Contents

1. Introduction

This document collects the Final Project titled "*Mining semantic relations from product features*". In this chapter, the main concepts of the project are described, such as the presentation of the problem found and the motivation to solve it.

1.1. Problem and motivation

At the beginning, the World Wide Web was developed to provide the most human knowledge and culture, which would let people everywhere to share their ideas and all the productions related with the project. Nowadays, the Web has become a vast resource of information where everyone collaborate providing contributions in many senses. One of these ways of collaboration is the fact of reviewing something. Websites, blogs, communities and forums let posting reviews or comments for many products or services.

As e-commerce is becoming more and more popular, the number of customer reviews that a product receives grows rapidly. Meanwhile normal products receive some reviews, most popular ones receive hundreds. "The more information, the better" has been the slogan of the information age [1], but sorting through huge amounts of information is tedious and infeasible manually. This makes it difficult for a potential customer to read all them and to make an informed decision on whether to purchase the product. Moreover, it obstructs to keep track and to manage customer opinions for the manufacturer of the product. There are additional difficulties because many merchant sites may sell the same product and the manufacturer normally produces many kinds of products [2].

Inside reviews, reviewers usually do not write only giving their objective vision of the product. Nevertheless, they manifest their feelings or impressions, praising or judging aspects of them. Writing about product features is often very different for both, reviewers and merchants, and between each others, too. Instead of merchants provide technical terms in the specifications, customers use them or not, and sometimes they introduce their own terms. Moreover, there are product features which are not found in the specifications. For example, the audio and video codecs that a smart-phone supports usually accompany them, such as *mp3*, *mp4*, *mpeg* or *avi*, but other product-features like *quality sound*, *resolution*, *volume*, *mode* or *track*, which are not included in specifications, are found in reviews.

Many researchers have been focused their studies on different tasks, such as opinion mining, product feature extraction and classification, reviews summarization, semantic

similarity measurement between features or semantic ontologies building. However, there is limited work focused on the field of grouping terms semantically related through semantic relations found between them, thus, it is the thesis followed in this document.

Considering facts such as a) technical specifications briefly group the main aspects of each specific smart-phone, separating these ones in categories, where single product features are located in their respective context; and b) reviews are usually accompanied by technical specifications, to concrete the model which customers are reviewing and to summarize at the same time the whole information about the product. It is thought that technical specifications suppose a semantic group formed by the terms which appear together, grouped by the same main characteristic. Due to that, they are also considered as a resource in this research, in the same way as reviews. Therefore, it is supposed that it will be possible to find new semantic relationships which relate more terms with the first ones, helped by these both resources, reviews and technical specifications, where it is assumed that terms are semantically related between each other, separated in categories. That is the point of this research, to increase these first groups of semantically related terms, technical specifications, looking for new semantic relations between terms from these groups and new terms located in reviews.

For example, in **Figure 1-1** there are three random customers who have reviewed three different smart phones. Taking these reviews as the corpus, it is illustrated how the goal of extracting and grouping semantically related product features is reached.

> [...] It doesn't seem to take much toll on the battery, and there is no Wi-Fi or big screen to drain it, so battery life is rated for the decent six hours plus of talk time, and the outstanding 25 days of stand-by. [...]

> [...] The manufacturer rates the 1320 mAh battery as being capable of providing juice for 4 hours of 3G talk time or 6 days of stand-by. [...]

> [...] The battery of the LG Mini GD880 can be quite robust in case you switch off 3G and Wi-Fi dependant functions. According to the manufacturer, it should be able to provide 7 hours of continuous talk time and keep the handset operational for 14 days in stand-by. [...]

Figure 1-1. Three fragments of reviews.

Figure 1-2. Specifications from a smart-phone.

As it is said before, technical specifications most of times accompany reviews, then, in this example it is had also some brief specifications only referred to the battery field which is analyzed here. **Figure 1-2** shows how specifications are stored on a website.

Nouns and compound nouns are identified as product features from both, reviews and technical specifications. From technical specifications it is taken terms like *battery*, *talk time*, *stand-by time*, *capacity*, *hour*, *minute*, *day* and *mAh*, which all are supposed to be semantically related between each other. From reviews it is processed the same, product features are underlined in **Figure 1-3**.

[...] It doesn't seem to take much toll on the battery, and there is no Wi-Fi or big screen to drain it, so battery life is rated for the decent six hours plus of talk time, and the outstanding 25 days of stand-by. [...]

[...] The manufacturer rates the 1320 mAh battery as being capable of providing juice for 4 hours of 3G talk time or 6 days of stand-by. [...]

[...] The battery of the LG Mini GD880 can be quite robust in case you switch off 3G and Wi-Fi dependant functions. According to the manufacturer, it should be able to provide 7 hours of continuous talk time and keep the handset operational for 14 days in stand-by. [...]

Figure 1-3 Extraction of product features from fragments of reviews.

[...] It doesn't seem to take much toll on the battery, and there is no Wi-Fi or big screen to drain it, so battery life is rated for the decent six hours plus of talk time, and the outstanding 25 days of stand-by. [...]

[...] The manufacturer rates the 1320 mAh battery as being capable of providing juice for 4 hours of 3G talk time or 6 days of stand-by. [...]

[...] The battery of the LG Mini GD880 can be quite robust in case you switch off 3G and Wi-Fi dependant functions. According to the manufacturer, it should be able to provide 7 hours of continuous talk time and keep the handset operational for 14 days in stand-by. [...]

Figure 1-4. Grouping product features semantically related with battery.

Regarding the meaning and the context of the product features obtained from technical specifications, the most semantically related terms with these first ones, are extracted. In **Figure 1-4**, green marked terms are supposed to be more semantically related with the battery context than the red marked ones, applying the knowledge as a human being. Terms such as *screen*, *juice*, *LG Mini GD880* and *function* have any relation in any way with the first battery

ones, meanwhile *battery*, *Wi-Fi*, *battery life*, *hour*, *talk time*, *day*, *stand-by*, *manufacturer*, *3G talk time*, *3G* and *handset*, are some of those first terms or have a strong relation with the battery terms named before.

Finally, these selected product features, in union with the first ones, form the semantic group looked for. Results like these may help and improve opinion and data mining tasks, knowledge representation, machine learning processes, information retrieval improvement through query expansion, etc. Unlike humans conceive knowledge helped by experiences, emotions and feelings, an automatic system has any of these contexts, only plain text, where many achievements have been done, but there is still a long way to do. This document explains how the same task can be carried out with tools such as information retrieval techniques, which have been used to achieve different goals for years, but which fit perfectly in this context, because their treatment of natural language in large amounts of data has been successful.

1.2. Distribution of chapters

The distribution of the chapters in this research is structured as follows:

Chapter two covers the state of the art of which studies have been done and how researchers have covered in some way the thesis presented before.

Chapter three describes the context where the analysis is run.

Chapter four contains the theoretical grounds and concepts of manual and automatic processes of new semantically related terms extraction.

Chapter five contains the complete analysis of performances of manual and information retrieval techniques analysis.

Chapter six contains the summarization and graphic visualization of the results obtained in the analysis.

Chapter seven concludes the research giving the conclusions companied with the vision of the future work that could continue this research, a summarization of the problems found and a personal reflexion of the whole process of making this project.

Chapter eight contains the bibliography references that have been used for the development of this research.

Next chapters contain the appendix used in the development of this research.

2. Related work

There are many studies focused on the different points of the whole process. Most of them are related with the product feature extraction step. However, there are also researches which cover the semantic similarity process, but they boarded it in some different ways, i.e., grouping and classifying product features, measuring semantic similarity between two words, or with ontologies. Moreover, an extended definition of semantic relations is covered in this chapter.

2.1. Product feature extraction

The product feature extraction process has been usually included in the opinion mining process, even though sometimes it has been treated separately. Studies like [2-12] have used the part of speech (POS) tagging as the first step before extracting the product features and only [2-4, 6, 7] have used in addition the stemming process, as it is done in this research. Moreover, in [6-8] it is argued that product features always tend to be nouns and compound nouns, and they have to satisfy a frequency threshold, as it is seen in chapter three. Looking at what techniques they have used to extract product features, this field of research is divided in different points of view: data mining, which is the most used, but also probability based algorithms, maximum entropy models, feature similarity methods and re-occurrence.

One of the earliest efforts was the system of Hu and Liu in [3], where they boarded the problem from the data mining sight, extracting language patterns and using association rule mining techniques. They improved the system to summarize features extracted by their polarity in [2], and Liu et al. in [4] developed an observer from the same system to visualize graphically the opinions about features of particular products. Association rule mining was also developed by Kim et al. in [11]. In the same line of data mining, Popescu and Etzioni developed OPINE, a system based on a feature assessor that using Point-wise Mutual Information (PMI) assessment to evaluate each candidate feature and incorporates Web PMI statistics to review data in its assessment which increased precision. Miao et al. [8] use the idea of getting the product features mined from reviews in format (2) (see section **3.1**) as a prior knowledge to extract the product features in format (3).

Apart from data mining, Scaffidi used in [6] a simplified version of the algorithm of the observer of Liu et al. [4] as a baseline for evaluating Red Opal's probability based feature extraction algorithm. He compared Liu et al.'s support-based algorithm [4] versus his probability-based algorithm [6], demonstrating improved utility of his method without

sacrificing scalability. Later, Scaffidi et al. in [7] employed these probability-based heuristics and baseline statistics of words in English to identify product features. Somprasertsri et al. [9] used a maximum entropy model with lexical and syntactic features, a framework for integrating information from many heterogeneous information sources. They redefined it as a classification problem, in which the task was to observe some syntactic dependency and context information, and predict the next. Lopez-Fernandez et al. [12] extracted a set of product features from the pros and cons and applying a feature similarity function they extract the new ones from the summaries. Then, they did the same but with another corpus of a different product. The main idea was to use the first corpus as a training corpus for the extraction on the second one. Finally, Guo et al. [10] described a method based on the re-occurrence of the product features in reviews. The first ones were taken from the pros and cons removing the opinion and stop words. Then, they verified the re-occurrence of the entire fragment or only a substring in the summaries, if they re-occurred they were valid to be product features.

However, the process is further simplified here, where only nouns and compound nouns, as it is done in [6-8], are taken as terms. They are selected applying a POS tagger, like in [2-12]. After that, a stemmer is run over them, like in [2-4, 6, 7], and those which only has 1 occurrence in the corpus, a threshold is set like in [6-8], are rejected, because they can be writing mistakes or reviewer's own words. Following this approach any term is discarded of being a product feature or to be related with the first ones. The spectrum is very large, but all possibilities are covered to increase semantic groups started from the technical specifications. Once product features are extracted, each appearance of each feature can be counted, getting the frequency of a feature in a review, the bag of words model, where the order of the terms does not worth, but the co-occurrence. This made possible the building of the term-document matrix. However, product feature extraction is not the most important step in this research, then, it is covered without losing much information.

2.2. Types of semantic relation

Concepts and relations are the foundation of knowledge and thought, is said in [13]. The variety of semantic relations and their properties play an important role in human comprehension and reasoning. Besides words, semantic relations can occur at higher levels of text - between phrases, clauses, sentences and larger text segments, as well as between documents and sets of documents. However, it is needed to be described what is meant by semantic relations, while it is talked about looking for them. Most researchers have identified a high variety classification of semantic relations, considering different aspects to be one of them. It is selected some different perspectives.

First consideration was described by Saussure [14], latterly covered by most researches, who recognized two categories: paradigmatic and syntagmatic relations. Paradigmatic relations are relations between pairs of words or phrases that can occur in the same position in the same sentence [15]. The words often have the same part-of-speech and belong to the same semantic class, and are to some extent grammatically substitutable. Examples include

ISA (is-a) (broader-narrower), part-whole and synonym relations [13]. Syntagmatic relations refer to relations between words that co-occur (often in close syntactic positions) in the same sentence or text [15]. It is a linear or sequence relation that is synthesized and expressed between two words or phrases when we construct a sentence. The relations are governed partly by syntactic and grammatical rules of a language.

Other researchers have a much more elaborate list of specific relations, often based on lexical-semantic relations and words found in a text (e.g. [16]). Lexical-oriented models often group relations into families of relations with the same core meaning or function. These relations are an important group of relations since they provide structure to lexicons, thesauri, taxonomies and ontologies. A word's relationship with other words is part of the meaning of the word. The vocabulary of a language is thus viewed as a web of nodes, each representing a sense of a word, and labeled links representing relations between the word senses. The main lexical-semantic relations are the paradigmatic relations of hyponymy (ISA or broader-narrower term), part-whole relation, synonymy and antonymy. However, frequently occurring syntagmatic relations between a pair of words can be part of our linguistic knowledge and considered lexical-semantic relations [13].

The final considered classification is where semantic relations are four of the five (selected in [13]) well-known paradigmatic relations often used in thesauri and ontologies, such as hyponym-hyperonym, meronym-holonym, synonymy and antonymy except troponymy, due to the omission of the verbs on the seeking of semantic relations, and the cause-effect relation, which is an important syntagmatic relation in human knowledge structures. These relations are often treated as unitary primitive relations. However, they are complex relations which can be subdivided into subtypes with different properties.

The hyponymy relation has been referred to in the literature under various names, including ISA (is-a), a-kind-of, taxonomic, superordinate-subordinate, genus-species and class-subclass relations. Hyponym refers to the narrower term/concept (e.g. cat), and hyperonym is the broader term/concept (animal). The relation implies class inclusion, i.e. all instances of cat are animals, the set of cat instances is a subset of animals, and the meaning of cat is included in the meaning of animal [17].

The meronymy relation is also referred to as part-whole relation and partonymy, and refers to the relation between a concept/entity and its constituent parts. The distinction between meronymy and hyponymy relations is clear for concrete concepts but fuzzy for abstract concepts [13]. Hyponymy relations can be said to exist within concepts, while meronymy relations are between concepts. Hyponyms inherit features from the hyperonyms but parts do not inherit features from the whole, though there is an upward inheritance for some attributes like color, material, and function [18].

Two words (or phrases) are synonyms when they have the same meaning and are absolutely synonymous if all their meanings are identical in all linguistic contexts, but this last is very rare [13]. Common types of synonyms are sense-synonyms (terms which share one or more senses), near-synonyms (which have no identical senses but are close in meaning), and partial synonyms (which share some senses but differ in some aspect, e.g. in the way they are used or in some dimension of meaning) [19, 20].

Antonymy, or opposites, is one of the most well-studied relations, and is the relation that people find easiest to learn and process [21]. Justeson and Katz in [22] concluded that "the patterns [of phrasal substitution] are so pervasive, that there is simply no chance for a genuine antonym pair to fail to show up in them, at a reasonable rate. So those that do not, cannot be antonymic". They suggested that the frequent co-occurrence of antonyms in text and discourse reinforces people's knowledge of antonymous pairs, which partly explains how antonymous pairs are learnt and why antonym relations are graded. Frequently co-occurring antonymous words are more likely to be judged as good antonyms than less frequently co-occurring antonyms [13].

The concept of causation is complex and surprisingly difficult to define, because many philosophers from Aristotle till the present have grappled with the concept [13]. It refers to the relation between two events. If an event E2 is caused by another event E1, then a causative relation exists between event E1 and event E2. E1 is the cause of E2 and E2 is the consequence of E1 [23]. However, these facts have been viewed from many perspectives, but it is only named to complete the selection.

After presenting which semantic relations can be found and in how many classifications they can be included, mining semantic relations without the distinctions described before is the goal in the process of this research. All the relations are treated as one, but it is important to know what it is understood by the concept of semantic relation. Psychologists consider semantic relations to be important in explaining the coherence and structure of concepts and categories. A category is not just a random set of entities - the entities in a category must belong together in some way. A category or concept is coherent - it must make meaningful sense [13]. Following these considerations, what it is searched here is a kind of category where all the terms included are semantically related between each other and make sense.

2.3. Semantic similarity

The most common way to determinate the semantic relation between two terms is defining a metric that measures the likeliness of their meanings. There are many methods that calculate the semantic similarity in many different ways. It could be possible to focus on semantic similarity between relations [24], but it starts from the assumption that the relations are already known and the point is to compare term pairs. However, finding semantic relations is the goal of this research, then, semantic similarity between attributes, terms or product features is what is looked for here. Sometimes this semantic similarity between terms is expressed in clustering features or building ontologies, not only in semantic similarity measure algorithms or functions. The most relevant studies are briefly described below.

2.3.1. Product feature clustering

Some researchers noted that customers may comment on different product features, but the words they use will converge. As it is thought here, one way of finding a semantic relation

between words may appear if there are such as latent groups or topics that can be identified in the corpus. Probabilistic Latent Semantic Indexing (PLSI) [25] and Latent Dirichlet Allocation (LDA) [26], which introduced topic modelling, have been the inspiration for many other studies, because of that they are deeply analyzed in next chapter five.

In one of them, Titov and McDonald [27] proposed a multi-grain unsupervised topic model. They objected that standard models tended to produce many topics that correspond to global properties of objects rather than the aspects of an object that tended to be rated by a user. The main difference was that they conceived the idea of two distinct types of topics: global topics and local topics. As in PLSI and LDA, the distribution of global topics is fixed for a document. However, the distribution of local topics is allowed to vary across the document. A word in the document is sampled either from the mixture of global topics or from the mixture of local topics specifically for the local context of the word.

Guo et al. [10] proposed an unsupervised product feature categorization method with multilevel latent semantic association. After extracting product features, as is said before, they constructed the first Latent Semantic Association (LaSA) model to group words into a set of concepts according to their virtual context documents. It generated the latent semantic structure for each product feature. The second LaSA model is constructed to categorize the product features according to their latent semantic structures and context snippets in reviews.

Zhai et al. [28] proposed a constrained semi-supervised learning method to solve the problem of producing a meaningful summary of terms, which are domain synonyms, under the same feature group. It is called semi-supervised, because they took as prior knowledge a small set of seeds labelled and the ideas of, feature expressions which shared words, e.g., *battery life* and *battery power*, and feature expressions which were synonyms in a dictionary, e.g., *movie* and *film*, were likely to belong to the same group. First, from the prior knowledge extracted what they considered as soft-labelled seeds and an initial naive Bayesian classifier learnt from them. Then, the current classifier using probabilistic label learnt from the unlabeled feature expressions. Finally, a new classifier was learnt using the labelled and the unlabelled features. They extended their research in [29] focusing on opinion mining, but following the same steps described before.

Finally, Pangos et al. [30] mixed concepts by grouping features with the idea of scoring them with semantic similarity measures. They computed two different semantic similarity measures, the first one was based on the same model that is covered here, the bag of words model. This measure computed wide-context similarity, because it considered the context given by a number of terms on the left and right of the term they were scoring. Helped by a weighting function to weight these terms around, they calculated the similarity between two terms. On the other hand, they developed a narrow-context similarity based on bigram language model, similar to the wide-context one, but taking only the closest terms from the left and right of the scored term, which was scored with the help of a probability to calculate the distances between terms. Then, through a weighted linear combination of these two similarity measures they grouped semantically related words.

As it is said, it is followed the bag of words model, as in [30], which explains the consideration of taking reviews as groups of terms, without considering any position or

sorting. Moreover, the goal is to form groups by finding semantic relations between terms in reviews, like in [10] and [27], discovering topics supported by groups of words. However, the difference is that here it has been noticed a resource of information/knowledge, the technical specifications, which can be used as prior knowledge, as in [28], but maintaining the unsupervised condition of considering the process completely automatic. Instead of developing new methods from zero, well-known techniques of information retrieval are used in a different way as usual, to identify semantic relations between terms which increase the members of the initials groups obtained in the technical specifications.

2.3.2. Semantic similarity metrics on corpus

Web search engines provide an efficient interface to this vast information. Page counts and snippets are two useful information sources provided by most Web search engines. They have been used to develop web-based semantic similarity metrics. Focusing on that, it is considered at first the work of Sahami and Heilman [31], which developed a web-based kernel function to measure the semantic similarity between snippets, regarding related terms, not the whole snippet, to suggest related queries to the user in web search engines. The kernel function is the product of the query expansion of the two terms they wanted to compare. This query expansion came from the truncated vector of the highest TF-IDF weighted terms of each term vector.

Bollegala et al. [32] took the same idea of the snippets, this time the returned by a web search engine, and combined it with the page counts returned. They developed a method that consisted of four page-count-based similarity scores (Jaccard, Overlap and Dice coefficients and PMI) and automatically extracted lexico-syntactic patterns helped by pairs of synonyms of WordNet synsets. Takale and Nandgaonkar [33] used the snippets returned by Wikipedia to demonstrate that they have a significant influence on the accuracy of semantic similarity measures such as the named before and the cosine similarity and simple matching coefficient.

Following the steps advanced by Pangos et al. [30] explained above, Iosif and Potamianos [34] proposed two web-based metrics. Both measures used a web search engine in order to exploit the retrieved information for the words of interest. The first type was fully text-based and included two approaches which were variations of the cosine similarity based on the usage of wide- and narrow-context, because they assumed that similarity of context implies similarity of meaning. They used the top ranked documents to apply these metrics on them. The second type was based on the page counts returned, they also used Jaccard and Dice coefficients and Mutual Information (MI), introduced by Church and Hanks [35], to compute the likelihood between words. In the next work, Iosif and Potaminos [34] focused their work on the wide-context metric, improving its performance with the different weighting schemes, from the binary to the Logaritmic TF-IDF (LTF-IDF). They compared the new proposal to the other page-count-based metrics and also included a new one, the Google similarity distance developed by Cilibrasi and Vitányi [36]. It is based on the concepts of the Kolmogorov complexity, from where it is extracted the Normalized Compression Distance; the Google Distribution, which is universal for all the individual web users distributions; and the Google Code, based on the

Google Semantics which said that Google events capture in particular sense all background knowledge about the search terms concerned available on the web.

Apart from web-based metrics, there are other metrics, for example, the probabilistic vision of Lin [37], whose information-theoretic definition of the similarity is applicable as long as there is a probabilistic model. It is considered that the similarity is not defined by a formula, but it is derived from a set of assumptions about similarity instead. He demonstrated the universality of the definition by its applications in different domains where he has employed different similarity measures.

Finally, introducing the idea of the semantic relation involving three actors of a sentence, subject, predicate and object, and their main representations, noun, verb and noun, Hindle [38] proposed a semantic similarity measure that calculated the similarity between nouns considering subject similarity and object similarity. For each noun and verb pair, he got two Mutual Information values, for subject and object. He defined the object similarity of two nouns with respect to a verb in terms of the minimum shared co-occurrence weights. The same occurred with the subject similarity. The overall similarity of two nouns was the sum across all verbs of the object and the subject similarity.

Some semantic similarity measures are used by the information retrieval methods analyzed here, like cosine similarity, but it is used to retrieve related documents with a determinate query, then, the semantic similarity is term-document oriented, and the goal looked for here is a term-to-term semantic similarity, which is automatically developed by these methods without applying any process more.

2.3.3. Semantic similarity metrics on taxonomies, dictionaries and ontologies

Ontology is an explicit specification of a conceptualization. In an ontology, definitions associate the names of entities in the universe of discourse with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms [39]. Then, it is presented here some different methods and techniques that used ontologies somehow to measures semantic similarity between terms.

WordNet is a thesaurus that contains a large lexical database and it can be interpreted as ontology, because of its hypernym and hyponym relationships. As a consecuence, Budanitsky and Hirst collected five different WordNet distance measures in their work [40].The idea behind measure of Hirst and St-Onge was that two lexicalized concepts were semantically close if their WordNet synsets were connected by a path that was not too long and that “did not change direction too often”. Leacock and Chodorow also relied on the length of the shortest path between two synsets for their measure of similarity. However, they limit their attention to IS-A links and scaled the path length by the overall depth of the taxonomy. Resnik’s approach said that the similarity between a pair of concepts may be judged by “the extent to which they shared information”. Resnik defined the similarity to be the information content, defined as negative the log likelihood of the probability of occurrence of the concept and its sub-concept in the corpus, of their lowest super-ordinate. Jiang and Conrath’s approach also used the notion of information content, but in the form of the conditional probability of

encountering an instance of a child-synset given an instance of a parent synset. Thus, the information content of the two nodes, as well as that of their most specific subsume, played a part. Notice that this formula measured semantic distance, the inverse of similarity. Finally, Lin's similarity measure followed from his theory of similarity between arbitrary objects. It used the same elements as Jiang and Conrath's, but in a different fashion.

In the same line, working inside WordNet, Li et al. [41] combined three sources of information to measure the semantic similarity: the information shared by the two words to be compared inside the corpus, based on the Resnik's information content; WordNet as the main semantic knowledge base for the calculation of semantic similarity; and Brown Corpus to assist for the calculation of the statistical information. They developed some strategies in which they combined, linearly and nonlinearly, concepts like information content, maximum depth of semantic hierarchy and (sometimes transferred) shortest path length, using both ISA and HASA relations.

Like it happens with WordNet, there has been significant progress toward extracting semantic information from Wikipedia, which is one of the largest sources of collaborative developed knowledge. In the work of King and Satuluri [42], the Wikipedia article abstracts were their corpus. They first tagged the corpus using WordNet. Then, they collected the dependency paths and using these as features, trained two classification algorithms to predict the existence of a semantic relation among noun pairs in the test corpus, which predictions they evaluated using human judgment. Moreover, Stoutenbourg et al. [43] used Wikipedia articles and their links to extract the meaning of the relationships. Their idea was to build a semantic network based on the extracted relationships, expressed in OWL in the Web, between each linked article pair that captured, in part, the knowledge contained in Wikipedia.

Apart from WordNet, there are other ontologies where the semantic similarity measure problem is addressed. Spagnola and Lagoze [44] used ConceptNet[1] as a semantic resource to semantic measure. Unlike Wikipedia and WordNet, edges in ConceptNet contained additional semantic information between two concepts. Each edge had a relation type and a score that correlated to how well ConceptNet users believed in the validity of the relation. There were also predefined edge types to calculate conditional edge type transition probability. Then, from the edge score and transition probability they built three vectors, scores, transitions and edges, to calculate the pathway score as the semantic similarity measure. Instead of the usage of stemming, they concluded that it inserted some noise in the recognition of the strong semantic relations.

Other techniques focused their work on the fact of the existence of several ontologies that has entities semantically related and a combination is looked for, i.e., to search more efficiently on the Semantic Web. The work of Han et al. [39] showed that linearly weighting an ontology higher on its deepest entities had more sense, because they were more meaningful and specific. After that, they applied a Tabu Search algorithm, which proceeded assuming that there was no point in accepting a new solution unless it was to avoid a path already investigated, to improve precision. In the same line of ontology matching, Tous and Delgado [45] based their work on the intuitive idea that similarity of two entities could be defined in

[1] csc.media.mit.edu

terms of how these entities related to the world they shared. They modelled this relationship with a vector space, adapting a graph matching algorithm to iteratively compute the similarities between two OWL[2] ontologies, modelled as RDF[3] labelled direct graphs.

Adapting the idea of Hindle [38], explained before, to the ontologies field, it is found studies that have focused their researches on the subject-predicate-object recognition. Akbik and Broß [46] developed Wanderlust, an algorithm that automatically extracted semantic relations from natural language texts, using deep linguistic and grammatical patterns. They tried to build a semantic Wiki for the English Wikipedia corpus, as Stoutenbourg et al. [43]. Nagano et al. [47] and Lin [48] followed also Hindle's idea. On one hand, Nagano et al. [47] focused their work on extracting a set of verb-nouns from the corpus, and then, considering syntactic patterns and co-occurrence of verb-noun pairs, they built an ontology from the hypernym-hyponym relation between two nouns. On the other hand, Lin [48] extracted directly these dependency triples from the text corpus considering grammar relations between terms to build a thesaurus.

It is seen that looking for semantic similarity between terms or words usually ends building an ontology, a thesaurus or a semantic network, but in this case, Kozima and Furugori [49] build a semantic network from a dictionary to measure semantic similarity from this new resource. Firstly, they took an English dictionary, which is possible to see it as a tangled network of words, and extracted a subgroup of words. Then, they built the semantic network from their new and classified group of words. Finally, they measured semantic similarity considering this semantic network and word significance in corpus.

2.3.4. The YAGO-NAGA project: A mixture of ontologies and a semantic search engine

The YAGO-NAGA project of Kanseci et al. [50] consisted in automatically building and maintaining a conveniently searchable, large and highly accurate knowledge base, by applying information-extraction (IE) methods to Wikipedia and other sources of latent knowledge. The project started in summer 2006 at the Max Planck Institute for Informatics, located in Saarbrücken (Germany), with continuous enhancements and extensions. YAGO knowledge base represents all facts in the form of unary and binary relations and it could be queried for knowledge discovery by the NAGA semantic search engine. Semantic similarity could be viewed as a particular goal in the Semantic Search.

On one hand, it was seen that there are many methods which interacted somehow with ontologies, some of these ontologies already developed, such as WordNet, and some other generated networks after applying techniques, but many differences existed between all of them. As Kozima and Furugori in [49] showed that it was possible to adequate the existing sources to a new structure that provided more exact information for measure semantic similarity, the same did Suchanek et al. [51]. They realized that there were many fields of

[2] www.w3.org/2004/OWL

[3] http://www.w3.org/TR/REC-rdf-syntax

research of information technology that used background knowledge, but the existing applications typically used only a single source of background knowledge (mostly WordNet and Wikipedia). Because of that, a huge ontology available with knowledge from several sources, whose quality, with accuracy close to one hundred percent, was comparable in quality to an encyclopaedia. It was called YAGO[4] (Yet Another Great Ontology) [51], a core semantic knowledge built from Wikipedia and WordNet, and it compromised not only concepts in the style of WordNet, but also named entities like people, organizations, geographic locations, books, songs, products, etc., and also relations among these entities. Going deeper into the YAGO structure, it used category pages from Wikipedia, list of articles that belong to a specific category, but Wikipedia category hierarchy was poor. WordNet, in contrast, provided a clean and carefully assembled hierarchy of thousands of concepts. YAGO is based on a data model of entities and binary relations, but it could also express relations between relations and general properties of relations. It also clearly defined the YAGO semantics. The basic structure was the *fact* which was composed of an entity, a relation and an entity, similar to the subject-predicate-object named before [38, 46-48] and it was organized in these facts. It was developed to be extendible to new knowledge and to be available in many formats, such as plain text, XML, RDFS and SQL databases.

On the other hand, there were also many semantic similarity measures which have been presented here, but there was no one that wanted to unify all the spread efforts to develop a homogeneous system to acquire information, in our case the semantic similarity between words. Many works have focused their efforts in information extraction from many sources on the Web, but they did not address the querying of the acquired knowledge. Graph querying, data mining, entity-oriented Web search and other forms of "semantic" information retrieval provided ranking but they were not so expressive. Due to this fact, Kanseci et al. developed NAGA[5] (Not Another Google Answer) [52], which was presented as a new semantic search engine. NAGA's data model is a directed, weighted and labelled multi-graph composed of a set of nodes, a multi-set of edges, a set of node labels and a set of edge labels and also it used the notion of facts, like YAGO. NAGA's knowledge consisted in facts extracted from semi-structured Web-based sources such as Wikipedia and IMDB as well as hand-crafted ontologies such as WordNet. In order to query the knowledge-graph, NAGA provided a graph-based query language, built on the concepts of the SPARQL[6] (the W3C[7] standard for querying RDF data). The query language allowed the formulation of queries with semantic information and also more complex graph queries with regular expressions over relationships as edge labels. It returns multiple answers for some queries. In order to rank these answers, they proposed a scoring mechanism based on the principles of generative language models for document-level information retrieval and applied these principles to the specific and unexplored setting of weighted, labelled graphs. Its scoring model was extensible and tuneable and took into consideration several intuitive notions such as compactness, informativeness and confidence of the results.

[4] http://www.mpii.mpg.de/~suchanek/yago
[5] http://www.mpii.mpg.de/~kasneci/naga
[6] www.w3.org/TR/rdf-sparql-query
[7] www.w3.org

One of the last efforts was the work of Hoffart et al. [53], YAGO2, an extension of the YAGO knowledge base with focus on temporal and spatial knowledge. It was also automatically built from Wikipedia, GeoNames[8], and WordNet. An enhanced data representation introduced time and location as first-class citizens. The wealth of spatiotemporal information in YAGO could be explored either graphically or through a special time- and space-aware query language. YAGO2 framework assigned time spans, location and contextual information to all entities, facts and events, if it was possible. To facilitate the querying, they added three more dimensions: time, location and context. This makes YAGO2 a valuable gazetteer, not just for geographical, but also for temporal and semantic data in general.

[8] www.geonames.org

3. Context of analysis

When approaching the field of automatic text classification, the first problem is that any mathematical method, that can be used for the task of classification, only operate on numbers and not on long, unstructured passages of text [54]. This fact brings the necessity of converting reviews in natural language to numbers. The next sections define the context that surrounds the thesis of this research and how the process is carried out in this research to transform plain texts into structured data.

3.1. Types of review

Before start analyzing texts and extracting data from them, it is important to know where the analysis is located. In the recent years, the number of freely available online reviews is increasing at a high speed [10]. However, not the whole current collection of reviews is analyzed here. This research is limited to the real life dataset crawled from Phonearena[9] by the Institute for Information Systems[10] at the TU Braunschweig[11], which focus the problem on the smart-phones.

Figure 3-1. A review from Cnet.com that uses only Pros and Cons, format (1).

First of all, what is understood as a review is a text that represents an evaluation written in natural language: the point of view that a customer has formed of a specific product or service. This concept has many different ways to be expressed. It usually depends on where the review is located and what offers the website, blog, forum or community to post it. Actually, many times it is found accompanied by videos, photos, drawings, captures, graphics, etc. However, paying attention to the written text, it can be defined three review formats [4]:

[9] www.phonearena.com
[10] www.ifis.cs.tu-bs.de
[11] www.tu-braunschweig.de

- Format (1) - Pros and Cons: The reviewer is asked to describe Pros and Cons separately. **Figure 3-1** shows a review from Cnet.com[12] that uses this format.

- Format (2) - Pros, Cons and detailed review: The reviewer is asked to describe Pros and Cons separately and also writes a detailed review. **Figure 3-2** shows a user review from Epinions.com[13] that uses this kind of expression. Although it presents the beginning of the detailed text, it details the main aspects of the smart phone, e.g., multimedia, performance or design.

- Format (3) - free format: The reviewer can write freely, i.e., no separation of Pros and Cons. Many sites use this kind of review. **Figure 3-3** shows a user review from Phonearena.com[14] which uses it. As it occurs in format (2), it is usually to separate the whole text in sections like multimedia, performance or design, as it is seen on the upper right corner of the figure.

Figure 3-2. A review from Epinions.com that includes Pros and Cons and a detailed review, format (2).

Figure 3-3. A review from Phonearena.com that uses the free review format, format (3).

[12] www.cnet.com
[13] www.epinions.com
[14] www.phonearena.com

For formats (1) and (2), opinion orientations of product features (positive or negative) are known, because Pros and Cons are separated and thus there is no need to identify them. However, in this case it is not necessary to have them separated, because the challenge is to identify the semantic relations between product features, not the opinion polarity, and there is no extra information in the Pros and Cons to help it. In other words, Pros and Cons are such a summarization of the entire review. This research is focused only in format (3), due to it implies format (1) and (2) keeping the whole information of the review.

It is necessary to clarify what is meant by *product feature*. In previous works [3], the term feature has been used to refer to those aspects of a product that customers have expressed opinions on, also called *opinion features* [12]. In this case, a *feature of a product* (or a *product feature*) is an attribute, component or other aspect of a product represented by a *noun* or a *compound noun* that it is also called *term*, e.g., *divx player*, *battery life* or even *event*.

3.2. Structured data

Websites, communities, blogs and forums usually allow posting reviews of products or services. Moreover, they include some specifications, characteristics or features about them, that give the users the security of reviewing and talking about the same article.

When it is talk about services, it is included a description about it or the context where it is got, but in the case of products, it may have a high level of precision, to distinguish between two or more products very similar. That is the point of this study, where technical products, such as smart-phones, have an enormus number of reviews due to each model may have few technical differences between the rest of them, but being a product completely different. Technical specifications allow users to know what are them reviewing. Then, it is assumed that each review refer to a concrete smart-phone, whose technical specifications are precise and clear.

These technical specifications are extracted manually, considering both resources of extraction: the previous product feature extraction done by the Information Systems[15] at the TU Braunschweig[16], extracted from Phonearena.com[17]; and the current extraction from the product features of the the technical specifications, which accompain the little group of random reviews used in the manual analysis. The issue is that most of times, product features are quite similar, such as *music player*, *calculator* or *capacity*, but in some other cases, like it occurs with the collection different video and audio codecs that a smart-phone supports, or the different versions of a program that are susceptible to be installed in the smart-phone. This fact brings the necessity of covering the widest range of product features of a smart-phone without analyzing the complete amount of technical specifications associated to reviews.

[15] www.ifis.cs.tu-bs.de
[16] www.tu-braunschweig.de
[17] www.phonearena.com

As an example, **Figure 3-4** shows an example of how product features are located in technical specifications on the mentioned website. There, it is possible to see how they are organized in categories, and one of those categories is the one shown here, *multiemedia*. Because of that, in this research it is referred to the *battery*, *organizer* and *multimedia* main product features, due to the naming of these categories on technical specifications.

Figure 3-4. *Multimedia* product features in technical specifications.

In this figure, terms such as multimedia, music player, mp3, aac, aac+, wma, video playback, mpeg4, streaming, audio, video and youtube player, are extracted as *multimedia* initial product features. In general, it is not considered cases where *music* or *player* occur, being both parts that form the compound noun *music player*, because there are many exceptions, and it is considered the whole term as an entity o expression.

Finally, for each main product feature, *battery*, *organizer* and *multimedia*, initial terms are extracted collecting in groups of 11, 18 and 47, repectively. In spite of having enough terms to perform a complete analysis, these terms was supposed to appear in reviews, but reviewers do not use the entire collection of terms extracted and these lists are reduced.

3.3. Natural Language Processing

Natural Language Processing is the field of study of computer science and linguistics which covers the interactions between computers and humans, in this case, reviews in natural language (English). Two of its well-known efforts are the part of speech (POS) tagging, applied to determine the class of a word, in a grammatical way, in a sentence, and the stemming process, which reduces a word to its lexical lemma or root. These two techniques are used to extract potential product features, extracting nouns and compound nouns, and calculate their number of occurrences, by counting how many times is repeated a lemma.

3.3.1. POS Tagging

POS tagging is the process of assigning a part of speech like noun, verb, pronoun, adverb, adjective or other lexical class marker to each word in a sentence [11]. Product features are usually nouns or compound nouns in review sentences. Thus the part of speech tagging is crucial [2].

In the studies presented in related work, it is said that they used some POS tagger to tag sentences from reviews or resources of data. However, there is not a unique way to develop this task; there are many researchers that have concentrated their efforts obtaining quite good results instead. In [2-4, 8] they have used the NLProcessor[18] linguistic parser, which outputs a XML tagged file from the input texts. In [6, 7] they used the MontyLingua[19] as their POS tagger and the TreeTager[20] is used in [12].

Apart from which taggers have been used by these researchers, some of the available and public POS taggers have been proved. The first of them is the Standford POS Tagger[21], which was already used in [11]. This tagger has three pre-trained different models, the last one with a bidirectional. However, trying it on the product features extracted from technical specifications, it is found that there are some mistakes with each model. Running the tagger with the first and second models, it tags the *mp3* term as an adjective (JJ), and divides the *aac+* and *eaac+* terms in *aac* and *eaac* as adjectives (JJ), and *+* as a noun (NN). It changes it results with the third pre-trained model, because rather than solving the found problems, it changes them. This time it treats the same terms as foreign words (FW), which conceives words under the same class. Moreover, it divides in the same manner the *aac+* and *eaac+* terms, and as a consequence, it does not distinguish between the classes of words tagged as FW. Another tagger that has been tested is the CRFTagger[22], whose results also have brought doubts about its best performance in the corpus. Some mistakes such as treating the *eaac* term as an adjective (JJ), the *3g2* as a cardinal number (CD) or the *3g* term as an interjection (UH), have been relevant not to choose it.

[18] http://www.infogistics.com/textanalysis.html
[19] http://web.media.mit.edu/~hugo/montylingua
[20] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger
[21] http://nlp.stanford.edu/software/tagger.shtml
[22] crftagger.sourceforge.net

Finally the OpenNLP[23] POS tagger is run and its performance is really adequated to the goal proposed in this research, because it tags most of the unknown or "difficult" words it find as nouns (NN). Thus, it brings many potential product features which may be new terms that have not been treated before, like *eaac* or *3g2*, and it tags them as nouns. Indeed, it is used accompanied by a Sentence Detector and a Tokenizer, from the OpenNLP Tools, which is the recommended way to run it. It is usually to find problems between a POS tagger and natural languages, because, in this case, reviewers do not may use the correct syntax or grammar, then, the POS tagger return wrong results.

3.3.2. Stemming / Lemmatisation

In information retrieval (IR), the relationship between a query and a document is determined primarily by the number and frequency they have in common. Unfortunately, words have many morphological variants which will not be recognized by term matching algorithms without some form of natural language processing. In most cases, these variants have similar semantic interpretations and can be treated as equivalent for information retrieval (as opposed to linguistic) applications. A number of stemming algorithms have been developed for information retrieval in order to reduce morphological variants to their root form [55].

Studies named before, like [2-4], besides removing stop-words, which here is implied on the nouns and compound nouns extraction process of the POS tagging, stemmed the rest of the words, and applied fuzzy matching, a technique used to deal with words variants and misspellings [2]. In [6, 7] a stemmer was also used, but included in the POS tagger, whose output summarizes both processes.

Two of the most popular algorithms in information retrieval, the Lovins stemmer [56] and the Porter stemmer [57] are based on suffix removal. Lovins find the longest match from a large list of endings while Porter uses an iterative algorithm with a smaller number of suffixes and a few sensitive recording rules [55].

After running some available open source Java implementations, such as the Paice/Husk Stemmer[24], an implementation[25] of the Lovins' algorithm, and the Porter's original implementation[26], which solve some of the many implementations of the Porter's algorithm, but there have been found some subtle flaws on all of their performances. However, once again Porter created Snowball[27], a framework for writing stem algorithms, which is also the name of the resultant stemmer. The Snowball Stemmer is selected, due to the almost perfect identification of the stem of words in the corpus. In addition with the POS tagger, reviews are processed and lemmas of nouns and compound nouns are extracted considering different scenarios described in the next section.

[23] incubator.apache.org/opennlp
[24] www.comp.lancs.ac.uk/computing/research/stemming/paice/article.htm
[25] www.cs.waikato.ac.nz/~eibe/stemmers
[26] tartarus.org/~martin/PorterStemmer
[27] snowball.tartarus.org

3.3.3. Getting term-document matrix

As it is said before, the bag of words model is used in this research as the pillar of the process, where the nouns and compound nouns have been extracted and stemmed by natural language processing techniques. At the same time those processes are applied, the term-document matrix is built, considering occurrences of terms on each review. Counting the number of times that a term appear in a document (actual review), its frequency is located on each element of the matrix. However, to avoid writing mistakes or possible wrong recognitions, every term whose total frequency in the entire corpus which does not overcomes a single occurrence, is deleted.

Besides frequencies of terms found in each review, it is proposed in this research another way of conceiving the same problem, not only applying techniques on the whole reviews, but considering also sections and paragraphs. What is followed with these new considerations is to see how the automatic techniques change their results in order to how organized is the scenario where they are run. The context for each term changes, because each one is surrounded by a higher number when they are run on documents, where the whole review is taken; lower when sections are considered where the information contained in a delimited section by a title like "*Camera and Multimedia:*", etc., as reviews are almost always reviewed following this pattern; and the lowest number of context terms when each paragraph is taken as a 'document'. As a consequence, three different scenarios are extracted from the corpus, one for each view, documents, sections and paragraphs.

Finally, to improve the performance of automatic methods, it is added a document, section or paragraph, depending on the circumstances, which gather all the product features found in the technical specifications, for each main product feature, *battery*, *organizer* and *multimedia*.

4. Theoretical grounds

4.1. The manual process

The manual process is done without any technology or automatic technique which could help to carry out the objective proposed in this research. This objective is the same like it is done automatically, extracting the most semantically related terms in reviews according to the initial terms related to each main product feature, such as *battery, organizer* or *multimedia*. Starting from their initial terms, each one of these has a list of terms that are supposed to be semantically related between each other, because they are extracted from their technical specifications, respectively, but the way of achieving the goal is completely different. Initial terms are the same for both, automatic techniques and manual method, but at this time it is taken a little random representation of the corpus, a group of reviews which contains only 16 of the 542 reviews of the entire corpus considered in this research. Considering the same corpus would be tedious and almost impossible to show better the differences between them.

The idea consists in calculate the percentage of the times that an initial term co-occur in the same context with another term regarding the times that the first term appear in the whole corpus. Then, the strongness of this potential semantic relation is based on overcoming a percentage threshold, which is set in both ways, from the first term to the term that co-occurs, and vice versa, to consider when a term pair is semantically related. This method finally evaluate whether the new semantic relation makes sense, and then, results are shown graphically. Those new semantic relationships between the initial term and the last terms increase the semantically related group of terms for each named main product feature.

4.2. The automatic process

The process is divided in two main tasks: the product feature extraction and the semantic similarity measurement of the relationships. Once it is seen how a natural language as English has been translated into numbers, which may be readable for automatic techniques, the whole process is summarized in a simple graphic **Figure 4-1**.

The first main task consists on the extraction of the product features contained on reviews, process that can be briefly described as the extraction of the nouns and compound nouns and their frequency of appearance inside each review. In order to recognize which words are nouns omitting the modifiers of them, the part of speech (POS) tagging is performed at first. It means that a program tags every word as noun, adjective, verb, etc., considering the role played by each word in its own sentence. As a result, the entire set of reviews is POS tagged and it easy to determine which terms can be extracted as nouns and compound nouns. The next step is performed by the stemmer, a program which reduces a word to its root or lemma. This brings the solution to count how many times a term occurs in each review. Finally, it is possible to build the term-document matrix crossing the terms and documents locating the frequency of the first in the second ones. Three term-document matrices are built, the first one taking as documents the whole review, the second one considering sections inside reviews delimited by titles such as "*Introduction and Design:*", and the third one dividing each review in paragraphs and taking these ones as contextual entities.

One of the most important differences between this and other studies is that it has been noticed that there is information given many times accompanying a review: the technical specifications. This information shows the different aspects of a smart phone grouped by their functionality. For example, *talk time* and *stand-by time* always summarize the time that the *battery* lasts, then, they are placed in the battery section. It is easy to see that here is the goal which is looked for, the semantic relation between product features, thus, the technical specifications are taken as the starting point of the process, because starting with this known relations it is easier to find more. For example, as it is said before, if *talk time* is related to *battery* and *stand-by* is related with *talk time*, it is possible to deduce that *battery* is also related with *stand-by.* Technical specifications are used in two ways: as starting point to look for terms which are semantically related with them; and as extra reviews, to include the assumed knowledge. The particularity comes when methods are run without the named *extra reviews*, because some of the initial terms, such as *talk time 3G*, are not found in the corpus, and its only occurrence is in these three *extra reviews*, then, performance may change.

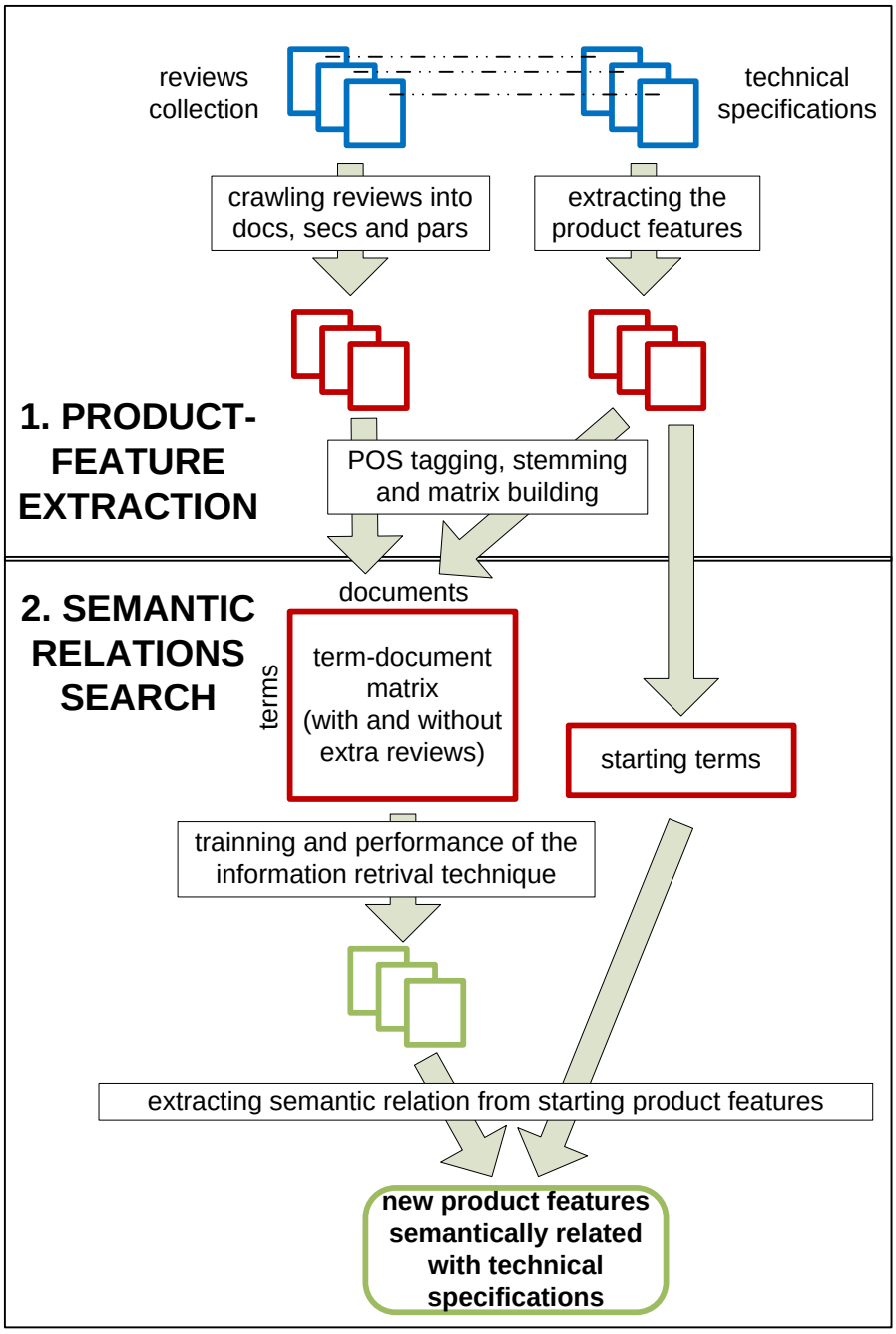


Figure 4-1 Graphic of the whole process of the semantic relations extraction.

Semantic relations between product features are wanted to be extracted here, and these semantic relations can be expressed as the semantic similarity between two terms. There are many methods and techniques that use semantic similarity measure between product features. However, this project is focused on the most common used and well-known techniques that have been and are the base of the information retrieval field of investigation for years. They are three methods: Latent Semantic Indexing (LSI, also called Latent Semantic Analysis, LSA) [58], Probabilistic Latent Semantic Indexing (PLSI, also called Probabilistic Latent Semantic Analysis, PLSA) [25] and Latent Dirichlet Allocation (LDA) [26]. Analysing how these techniques perform is deeply described in next chapters, because it is wanted to see the effectiveness of each method to extract semantic relations and the differences between them.

As a summarization, the whole process of extracting the semantic relations between product features from reviews is illustrated in **Figure 4-1**.

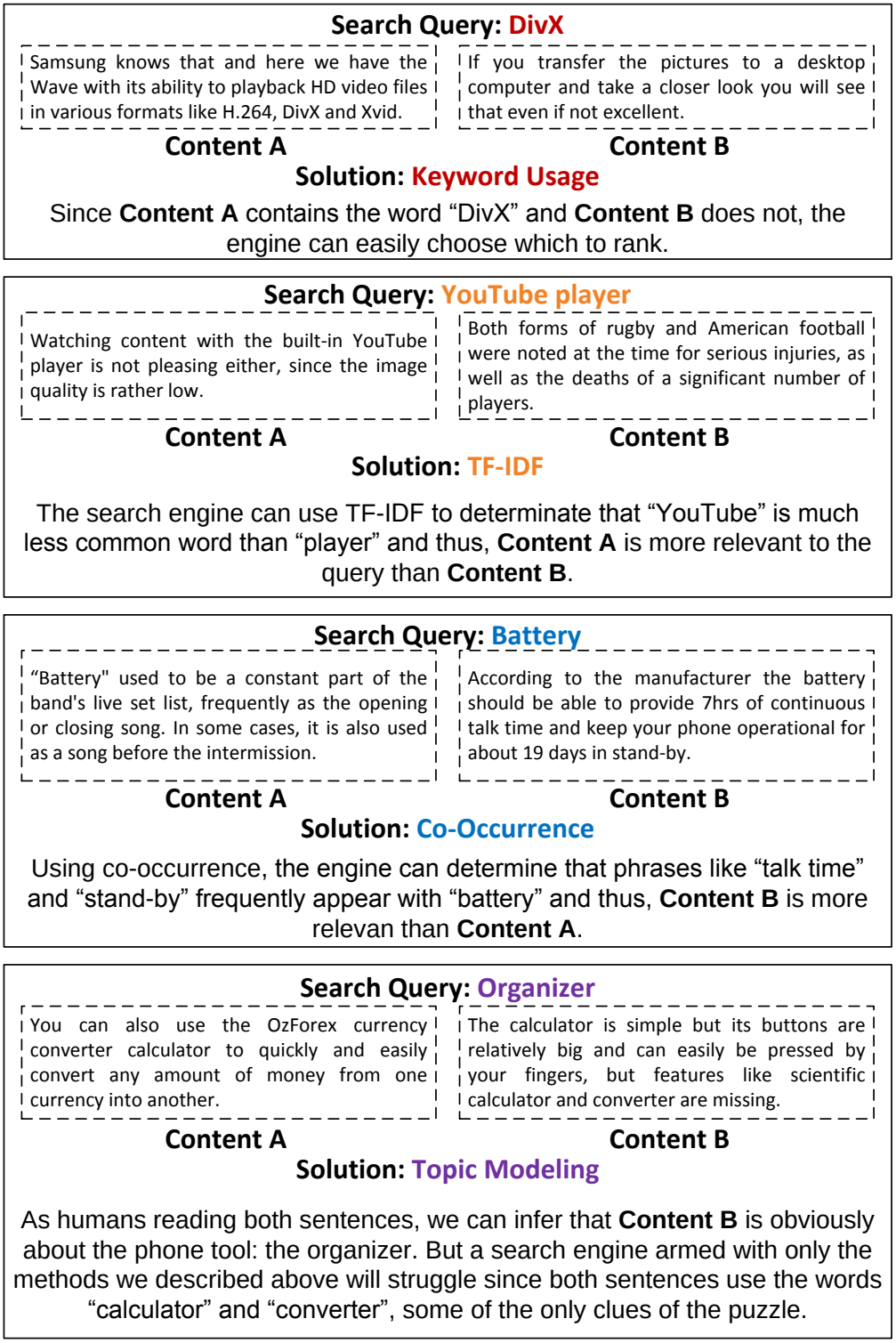


Figure 4-2. Evolution of the search engines techniques.

These three methods use the bag of words model, as it is wanted here, because the order of the words has not been considered and reviews have been treated as groups of words. It is a unigram language model that only calculates de probability of hitting an isolated word without considering the context that surrounds it. Bigram or trigram models are other

particular cases of the general concept of n-gram language models that calculate the probability of hitting sequences of two or three words considering the near context. Nevertheless, the main difference of the bag of words model compared to the other is that it can be said that it is language independent.

With these three techniques it is wanted to follow the evolution that has been experimented information retrieval by the different techniques during these years. At the beginning, appeared the keywords that were used by search engines to reach in an easy way the results that suited best the queries; afterwards came LSI, that used also co-occurrence and the Term Frequency-Inverse Document Frequency weight (TF-IDF), a statistic measure that weights how important is a word for a document; years later, PLSI, that has a solid statistical foundation, introduced the notion of classes represented by terms, but it still did not use a probabilistic model at the level of documents; and as a consequence, the topic modeling arrived with LDA, that described better the presence of latent topics in the data and uses a probabilistic model in every level. This seems to achieve what is looked for in this study, groups of words defining a topic. **Figure 4-2** shows the mentioned progression, locating the found problems and the proposed solutions for each one.

1. Run LSI on documents with all the possible values of its parameter, getting as output a matrix of scored terms organized in dimensions.
2. Choose which is the dimension or dimensions that have the initial product features of the *battery* better scored within of the sorted dimension, considering specially the first 30 terms.
3. Apply the discarding method to these potential new terms, which discard all the terms that have a number of better scores than in the analyzed dimension, higher than a threshold, e.g. 5 higher scores, in the rest of dimensions.
4. Judge as a human being the potential new terms classifying them in *true positive (tp),* a correct result*, false positive (fp),* an unexpected result*, true negative (tn),* a correct absence of result*,* and *false negative (fn)*, a missing result, omitting the initial product features, which are always correct results.
5. Calculate and show graphically the precision, accuracy and recall of these potential new terms.
6. Extract the new terms of the *battery* product feature which remain over the 80% of precision.

Table 4-1. Algorithm to extract product features semantically related.

As it is shown in **Figure 4-1**, the methods tested and the developed one are unsupervised. This means that they perform on their own, extracting knowledge with the only help of the training data, in this case the collection of reviews. Some other techniques are supported by WordNet[28], a thesaurus that contains a large lexical database of English nouns, adjectives and adverbs grouped into semantic and lexical-related sets; by Wikipedia[29], a free encyclopaedia where people have collaborated freely to build it; or by taxonomies, ontologies or other thesaurus built before to represent knowledge in some way. Although supervised techniques perform better than unsupervised, they depend on the maintenance of the knowledge and the challenge is to achieve the best results without extra knowledge.

[28] wordnet.princeton.edu
[29] www.wikipedia.org

Finally, what is looked for in the process is to extract automatically the best scored terms found in the dimensions (group of terms) where the main product features such as *battery*, *organizer* or *multimedia* are better defined by the high scored initial product features on the review assumptions of documents, sections and paragraphs. The steps included in **Table 4-1** are followed for each method, each main product feature and each assumption of review corpus, e.g. LSI, *battery* and documents.

4.3. Latent Semantic Indexing (LSI)

It is presented a short description [59] of the mathematical details of the method Latent Semantic Indexing (LSI), which is better described in [58]. The key of innovation of the LSI was to use Singular Value Decomposition (SVD) to decompose the original term-document matrix of the Vector Space Model (VSM). The term-document matrix, A_0 has *t* rows (one for each extracted term that appears in the selected set of documents) and *d* columns (one for each document in the set). The SVD, $A_0 = U_0 S_0 V_0^T$ results in a *t x m* matrix, U_0, the orthonormal columns (term dimensions) of which are called left singular vectors, an *m x m* diagonal matrix, S_0, of positive singular values sorted in decreasing order, and an *m x d* matrix, V_0, the orthonormal columns (document dimensions) of which are called right singular vectors. The value *m* is the rank of the matrix, A_0. **Figure 4-3** depicts the SVD of A_0. With the U_0, S_0 and V_0^T matrices, A_0 can be reconstructed precisely. The key innovation in LSI is to retain only the *k* largest singular values in the S_0 matrix and set the others to zero.

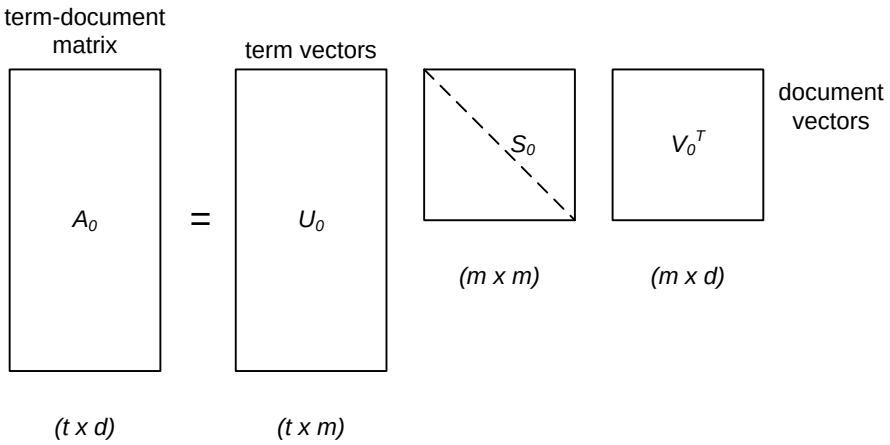


Figure 4-3. Singular Value Decomposition (SVD) of the term-document matrix A_0.

After the decomposition, the original matrix, A_0 is approximated by $A_k = U_k S_k V_k^T$, where U_k is a *t x k* matrix with orthonormal columns, S_k is a positive definite *k x k* diagonal matrix, and V_k^T is a *d x k* matrix with orthonormal columns as it is shown in **Figure 4-4**.

LSI performance can improve considerably after 10 or 20 dimensions, peaks between 70 and 100 dimensions, and then begins to diminish slowly. The number of dimensions k to keep in the reduced term-document matrix when *d* is very large is still open to study and debate, but experiments indicate that values of *k* between 100 and 300 typically give the best results [60]. Due to this fact, to see how it performs on each main product feature, with and without *extra reviews* of the initial product features, the *k* parameter is tested on the values of 50, 100, 150 and 200. Although it is talked about documents, it refers to the unity of consideration of reviews, which can be documents, sections and paragraphs.

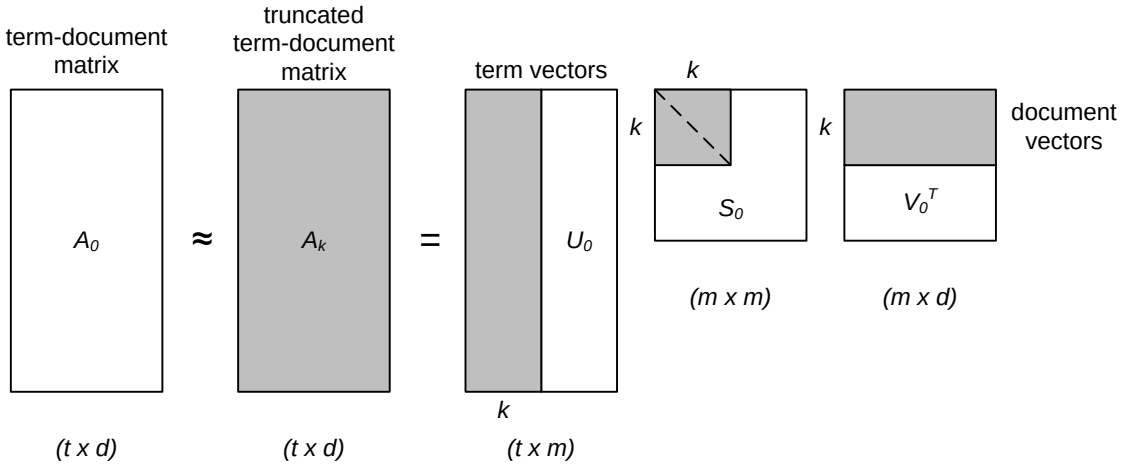


Figure 4-4. Singular Value Decomposition of the term-document matrix A_0 truncated to k singular values.

The LSI implementation used here is taken from the Institute for Information Systems[30] at the TU Braunschweig[31]. It also uses a weighting function to weight the term-document matrix entries. In **Table 4-2**, terms a_{ij} are calculated by the product of the local and global log entropy weighting function, where tf_{ij} is the term frequency of term *i* in document *j*, and gf_i is the total number of times that term *i* appears in the entire collection of *d* documents. It gives less weight to terms occurring frequently in a document collection, as well as taking into account the distribution of terms over documents. In [61], Dumais found that log-entropy gave the best retrieval results, 40% over raw term frequency.

$\boldsymbol{a_{ij} = g_i \cdot \log(tf_{ij} + 1)}$	$g_i = 1 + \sum_j \frac{p_{ij} \cdot \log(p_{ij})}{\log(n)}$	$p_{ij} = \frac{tf_{ij}}{gf_i}$

Table 4-2. Local and global log-entropy weighting function.

[30] www.ifis.cs.tu-bs.de
[31] www.tu-braunschweig.de

4.4. Probabilistic Latent Semantic Indexing (PLSI)

It is briefly presented the details of the method named Probabilistic Latent Semantic Indexing (PLSI) [25]. It is a statistical technique for analysis of two-mode and co-occurrence data, which has applications in information retrieval and filtering, natural language processing, machine learning from text and in related areas. Compared to standard LSI, PLSI is based on a mixture decomposition derived from a latent class model. This results in a more principled approach which has a solid foundation in statistics [62].

PLSI is based on a statistical model that is referred as an *aspect model*. An *aspect model* is a latent variable model for co-occurrence data, which associates unobserved class variables z_k, $k \in \{1, 2, \dots, K\}$ with each *observation*, where K is the number of latent classes. The number of latent classes is an important parameter that needs to be selected in the same way as LSI. The *observation* is an occurrence of a word w_j, $j \in \{1, 2, \dots, M\}$, in a particular document/context d_i, $i \in \{1, 2, \dots, N\}$, where M is the number of words and N is the number of documents in the collection. *Latent classes* can be understood as the *topics* that comprise the text. The probability distributions that associate the latent variables with words and documents describe how closely they are associated with each topic. The generative model for the observation is defined as follows [63]:

1. Obtain a document d_i in which a word occurrence will be observed with probability $P(d_i)$.
2. When the document d_i is known, select the topic z_k of the word with probability $P(z_k|d_i)$. This probability distribution is also a measure of the extent to which the document is relevant to each topic.
3. When the topic is known, select a word w_j whose occurrence is observed with probability $P(w_j|z_k)$.

In this way, the observation pair (d_i, w_j) can be formulated so that the latent class variable can be summed out. **Equation 5.1.1** shows the probability of observing a pair (d_i, w_j). Moreover, **Equation 5.1.2** is also perfectly symmetric with respect to both documents and words.

$$P(d_i|w_j) = P(d_i)P(w_j|d_i), \text{ where } P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \qquad \textbf{5.1.1}$$

$$P(d_i|w_j) = \sum_{k=1}^{K} P(z_k)P(d_i|z_k)P(w_j|z_k) \qquad \textbf{5.1.2}$$

Like virtually all statistical latent variable models the aspect model relies on a conditional independence assumption, i.e. d_i and w_j are independent conditioned on the state of the associated latent variable (the graphical model representing this is demonstrated in **Figure 4-5**). It is necessary to note that the aspect model can be equivalently parameterized by **Figure 4-6** [62].

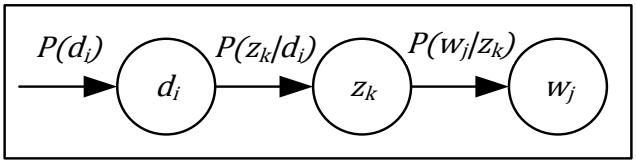


Figure 4-5. Graphical model representation of the aspect model in the asymmetric parameterization.

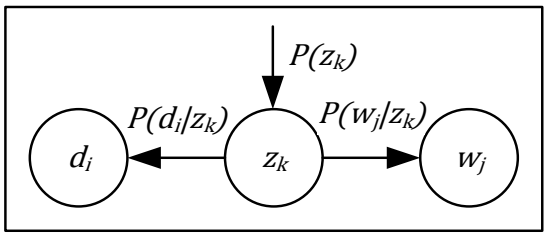


Figure 4-6. Graphical model representation of the aspect model in the symmetric parameterization.

Hoffman [25] proposed the use of the Expectation Maximization algorithm to identify the parameters for the probability mass function from the training data. The EM algorithm alternates between the following two steps: an *expectation* (*E*) step in which posterior probabilities are computed for the latent variables on the basis of the current estimates of the parameters (see **Equation 5.2**); and a *maximization* (*M*) step in which parameters are updated on the basis of the minimization criteria and in dependence on the posterior probabilities computed in the E-step (see **Equation 5.3** and **Equation 5.4**). In the equations, term $n(d_i|w_j)$ stands for the count of the word w_j in the document d_i.

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|d_i)} \qquad \textbf{5.2}$$

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i,w_j)P(z_k|d_i,w_m)} \qquad \textbf{5.3}$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{N} n(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{m=1}^{M} n(d_i,w_m)} \qquad \textbf{5.4}$$

The standard EM algorithm can, however, overfit the model to the training data and thus perform poorly with unseen data. Because this algorithm is iterative and converges relatively slowly, it can increase runtime dramatically, especially with large data sets. Because of this, Hoffman [25] proposed another approach that he calls *Tempered EM* (*TEM*), which is a derivation of standard EM algorithm. Essentially, it introduces a control parameter β (inverse computational temperature) and modifies the E-step as in **Equation 5.5**.

$$P(z_k|d_i, w_j) = \frac{P(z_k)[P(d_i|z_k)P(w_j|z_k)]^{\beta}}{\sum_{l=1}^{K} P(z_l)[P(d_i|z_l)P(w_j|z_l)]^{\beta}} \qquad \textbf{5.5}$$

Notice that $\beta = 1$ results in the standard E-step, while for $\beta < 1$ the likelihood part in Bayes' formula is discounted (additively on the log-scale) [25].

As it is said before, the most important parameter is the number of topics to run the method. In [64], Hoffman presented the model and tried it for values of the *k* parameter of 32, 48, 64, 80 and 128, because it did not perform better with higher values with four medium-size standard document collection with relevant assessment. From these performances, best results were obtained for 48 and 128 topics. Then, to be correlative with the selection done in the LSI analysis, the same values of the *k* parameter are selected. These values are 50, 100, 150 and 200. It is known that it makes not much sense to select almost half latent topics to appear in a collection of 542 documents (documents scenario), but it is also known that the accuracy can be increased of PLSI is increased by increasing the number of latent variables [63].

In order to clarify the relation with LSI, it is useful to reformulate the aspect model as parameterized by **Equation 5.1.2** in matrix notation. By defining $\widehat{U}_k = \left(P(w_j|z_k)\right)_{i,k}$, $\hat{V}_k = \left(P(d_i|z_k)\right)_{j,k}$ and $\hat{S}_k = diag\left(P(z_k)\right)_k$ matrices, it can be written the joint probability model P as a matrix product $P = \widehat{U}\hat{S}\hat{V}^T$. Comparing this with SVD, it can be drawn the following observations: outer products between rows of $\widehat{U}_k$ and $\hat{V}_k$ reflect conditional independence in PLSI; and the mixture proportions in PLSI substitute the singular values. Nevertheless, the main difference between PLSI and LSI lies on the objective function used to specify the optimal approximation [62]. Hence, the matrix, whose values are important in this research, is the one made by $P(w_j|z_k)$ results.

The implementation has been selected after several tries of different algorithms, because the number of paragraphs caused out of memory errors when they were run. Some MatLab and C implementations have been tested until a modified C++ version from Semantic Search Art[32] performed correct. The modification only differs from the original in the type of the data of the term-document matrix, changing from *float* to *short*.

[32] http://www.semanticsearchart.com/researchpLSA.html

4.5. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) [26] is a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.

While Hofmann's work [25] is a useful step toward probabilistic modeling of text, it is incomplete in that it provides no probabilistic model at the level of documents. In PLSI [25], each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers. This leads to several problems: (1) the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems with overfitting, and (2) it is not clear how to assign probability to a document outside of the training set [26].

Moreover, it is wished to consider exchangeable representations for documents and words, then, it is needed to consider mixture models that capture the exchangeability of both words and documents. This line of thinking has lead Blei to the Latent Dirichlet Allocation model [26].

The principles of LDA are similar to those of PLSI. The difference is that there is no need in LDA to estimate the probability of obtaining a document, and there is thus no need to perform the difficult estimation process when adding unseen documents to the model. Instead this is achieved by changing the generative model in such a way that it separates the process for each document and uses the word-latent class distribution to determine the document-latent class distribution. LDA assumes the following generative process for each document d_i in a corpus consisting of N documents that contains M distinct words and K distinct latent variables or topics [63]:

1. Choose the length of the document $L \sim Poisson(x)$. Note that it is (for most of the time) dealt with each sequence of words in a single document and not the distinct words in the corpus. It is therefore used indexing w_l for words in a single document $d_i = \{w_1, \dots, w_L\}$ and w_m for distinct words in a corpus.
2. Choose a parameter vector for the topic distribution $\theta \sim Dirichlet(\alpha)$, the parameter α is a $K - vector$ with components $a_k > 0$ and θ is a $K - vector$ so that $\theta_k \geq 0$ and $\sum_{k=1}^{K} \theta_k = 1$ and $P(\theta|\alpha)$ is the probability density function of the Dirichlet distribution.
3. For each of the L words w_l:
    a) Choose a topic $z_l \sim Multinomial(q)$. Note that z_l or z_L is used when it is discussed the topic for each word in a document, and that all topics in a document are referred as $z_d = \{z_1, \dots, z_l\}$.
    b) Choose a word w_l from $P(w_l|z_l, \beta)$, a multinomial probability conditioned on the topic z_l, where β is a KxM matrix so that

$\beta_{kj} = P(w_m|z_k)$ for all $1 \leq m \leq M$ and $1 \leq k \leq K$, where M is the number of distinct words in the corpus.

In order to build up the model for LDA, one should compute the posterior distribution of the latent variables for a given document in the way shown in **Equation 6.1**.

$$P(\theta, z_{d_i}|d_i, \alpha, \beta) = \frac{P(\theta, z_{d_i}, d_i|\alpha, \beta)}{P(d_i|\alpha, \beta)} \qquad \textbf{6.1}$$

But because **Equation 6.1** is intractable, it needs to be approximated. Blei [26] introduce an EM-based variational algorithm to approximate the equation and maximize the log likelihood of the model based on the α and β parameters. It is described the algorithm briefly at this point. Further details can be found at [26].

In the E-step, the density function in **Equation 6.1** needs to be approximated with a tractable model. The idea is to minimize the Kullback-Leibner Divergence between the tractable and intractable model by finding the minimal values for Dirichlet parameter γ and multinomial parameters $(\theta_1, \dots, \theta_L)$ in the tractable model. In order to find the optimal γ and ϕ for each document d_i, Blei in [26] obtained the updated equations (**Equation 6.2** and **Equation 6.3**) for these parameters,

$$\phi_{mk}(d_i) \propto \beta_{km} exp\{E_p[log(\theta_k)|\gamma(d_i)]\},$$

$$\text{where } E_p[log(\theta_k)|\gamma(d_i)] = \Psi(\gamma(d_i)) - \Psi\left(\sum_{j=1}^{K} \gamma_j(d_i)\right) \qquad \textbf{6.2}$$

$$\gamma_k(d_i) = \alpha_k + \sum_{i=1}^{N} \phi_{nk}(d_i) \qquad \textbf{6.3}$$

Because the γ parameter vector describes the distribution for each document, it can be used in the similar way to $P(z_k|d_i)$ in the PLSI model. These two equations are computed repeteatedly for all l, k and d_i until the lower bound achieved from Jensen's inequaly converges.

In the M-step, α and β parameters need to be estimated once the new values of ϕ and γ have been calculated. Blei [26] proposed the use of the Newton-Raphson optimization technique to find the stationary point of the α function by iterating **Equation 6.4**. The conditional multinomial parameters α and β are also updated as in **Equation 6.4** and **Equation 6.5**.

$$a_{new} = a_{old} - H(a_{old})^{-1} g(a_{old}) \qquad \textbf{6.4}$$

$$\beta_{km} \propto \sum_{i=1}^{N} \sum_{l=1}^{L_{d_i}} \phi_{lk}(d_i) eq(w_l, w_m), \qquad \textbf{6.5}$$

where $H(\alpha)$ and $g(\alpha)$ are the Hessian matrix and gradient respectively at point α and $eq(w_l, w_m)$ is 1 if a word w_l from the document d_i is the same word as mth distinct word w_m in the corpus otherwise 0. After each cycle in the EM algorithm the convergence of the model building is measured by means of the log-likelihood of the model.

In **Figure 4-7** on the top, it is possible to see graphically how PLSI is modelled, represented by the plate notation, where the plates or boxes are replicates. θ(called d_i before)

is the document variable, z is the topic drawn from the topic distribution for this document, $P(z|\theta)$ (called before $P(z_k|d_i)$), and w is a word drawn from the word distribution for this topic, $P(w|z)$ (called before the same). The w variables are the only observable variables, while the θ distribution and topic z are the latent variables.

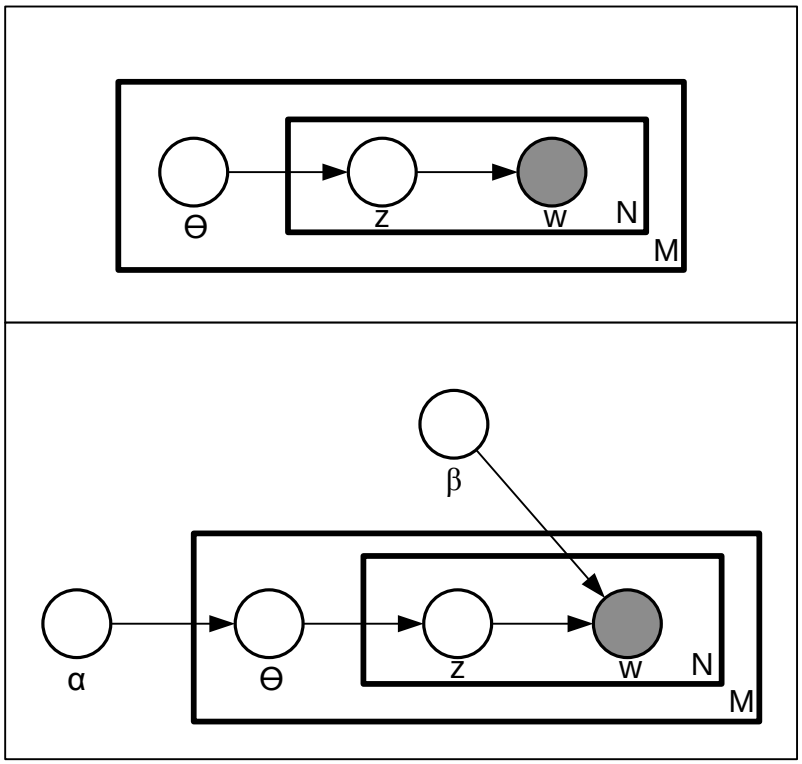


Figure 4-7. On top, PLSI model, on bottom, LDA model.

In the same **Figure 4-7** on the bottom, it is seen the same schema tha for PLSI, but for LDA, including the parameters of the Dirichlet distribution. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document, as it occurs with PLSI. However, α is the parameter of the Dirichlet prior on the per-document topic distribution, and β is the parameter of the Dirichlet prior on the per-topic word distribution. As it occurs in PLSI model, the only observable variable is w, while the paramenters α and β, and also θ and z are the latent variables.

The LDA implementation used in this section is the LDA package[33] of Daichi Mochihashi, which includes both versions, in C++ and MatLab. Although the Matlab one is more comfortable to treat with huge matrices, the C++ one is quite faster, then, performances are carried out by the executable, but results are treated in the same way like LSI and PLSI in MatLab.

[33] http://chasen.org/~daiti-m/dist/lda

4.6. Own method proposals

At first, the goal of this research was only to check whether information retrieval techniques satisfy the proposed objectives of increasing initial groups of semantically related terms. However, it was considered the idea of developing an own method that, designed for exclusively for this task, could improve results given by the presented methods. Meanwhile information was collected about state of the art and the first results were got from the first methods, this idea took many forms, and finally, after analysing all IR techniques, two ideas were implemented. Although both did not have such successful results as to be included on the analysis in next chapter, their theoretical concepts are explained here.

4.6.1. Co-occurence and context similarity

Based on the concepts presented to perform the manual analysis, the goal is to carry out the same process of new product feature extraction. It pretends to be the mixture of two concepts related to how terms are located int the corpus. The first concept refers to the co-occurrence of terms in a document (or context). It is assummed that those terms, whose percentage of co-occurrence with other terms in the same document is high regarding the number of times they appear in the corpus, are supposed to be strongly semantically related. Then, this percentage is calculated from the initial term-document matrix in a term-by-term matrix, where each terms is crossed with the entire collection of terms, and it is possible to see how high or low are their semantically relationships. Some efforts like [65], which has used and analysed the truncated term-by-term matrix resulting of the LSI performance after multiplying matrices, has shown that similarity between term pairs it is found also within the direct relation between terms. Because of that, it is thought that term vectors can be considered as term topics, supposing that each term vector represents its semantical group, where terms highly scored (except itself) are supposed to be strongly related with it.

The second part of the method consists in the assumption that semantically related terms usually appear in similar contexts. Each context of a term is the vector resultant of removing the current term from a document. Then, the goal consists in getting the highest value of the cosine distance between contexts of terms and building another term-by-term matrix, which will be combined with the co-occurrence one. However, the biggest problem found is that a single term has many different contexts where a term appears, and comparing all of these contexts with all of the rest term contexts suppose an enormous computing time when the number of terms considered is high, like it occurs with the used corpus. Although initial tests were run firstly on a mini corpus (like the used in [65]), the problems came when the time of calculating the cosine distance are too high due to the high number of different terms recognized in the corpus, more than 13.000 different terms.

The mixture of both concepts consists in the addition of both scores, the first from the co-occurrence and the second from the context similarity, but it was only tried on the mini corpus, where its performance is not concluyent to compare with the rest of the methods. The

fact that it is needed a dimensional reduction to get significant results or it is may needed a parallel implementation to calculate those values.

4.6.2. Cascade Latent Dirichlet Allocation

In this case, this proposal is based on the assumption that LDA has the best performance due to it has been developed to improve the other two. It does not mean that it performs better on the scenarios used in this research, but it is supposed that it overcomes the rest of performances. This issue satisfy the need of having a good dimension reduction, but it is thought that it could be increased its product feature extraction. If it is regarded that LSI, PLSI and LDA results are vectors of terms where terms are good or bad scored, following a probabilistic distribution or not, it is obtained the same structure that it is had initially, where the term-dimension or the term-topic matrix follow the similar structure of a term-document matrix: a matrix where a cross value between topics or dimensions with terms represent the interaction of these entities. In the initial case, input, they represent the frequency of appearing a term in a document (or context). In the second case, output, is quite similar, because it represents the probability of being related with a topic (PLSI and LDA) or the score given according to the SVD of the matrix (LSI). Then, it is proposed that these output values could be used as the intial ones, because they are structured in the same way. Because of that, this process has been named LDA on Cascade, or Cascade LDA (CLDA), where LDA is applied twice, one on the input data and the second on its results.

The problem arrived with the type of the data and the implementation of the method used. Fristly, as input of LDA it is had a term-document matrix of *integer* values, the frequency is represeneted by natural numbers equal or higher than 0, but its output consists in a group of vector which represent different probability distributions, which means that values of the crossed pair term-topic are *float* between 0 and 1. Although LDA's package[34] supports *float* type in both implementations, C++ and MatLab, it can not be run as well as it was thought. It almost always prints an internal error caused by the low values of the training data, then, it iterates until the end without calculating any value. However, it sometimes reaches a stable running, but it converges too fast (7-9 iterations) and looking at results, the first new topics usually are a mixture of popular terms (those which appear in most of documents like *phone*, *button*, etc.), and the rest of topics are different from it, but all with the same distribution. It is thought that it is caused due to the input data, because the variation between values in the term-topic matrix is almost nothing. It has been tried also multiplying those input values by a constant, but results are the same. Another improvement was to apply the same weighting function used in LSI implementation, where CLDA's performance seems to be more stable, obtaining results more times than before, but always in the same line, almost all new topics with the same probability distribution.

Finally, the objective of developing a new method is truncated and the following chapter covers the analysis proposed at first, manual, LSI, PLSI and LDA analysis on their respecitve corpus mentioned before.

[34] http://chasen.org/~daiti-m/dist/lda

5. Analysis

5.1. Manual analysis

It is assumed that, as a human being usually does, it is only consider the paragraph as the part of a review where a product feature can be related semantically with other ones. There, the analysis is performed as an automatic technique where each term, noun or compound noun, is analyzed by taking every paragraph where it appears, and there, it is counted which and how many times appear and how many of those times, they co-occurs with the rest of terms. There is no separation between with or without *extra reviews*, because it is done without *extra reviews* to check the actual performance of the pure data, and with a little corpus *extra reviews* would suppose almost the 16% of the reviews, which influence too much in results.

5.1.1. Battery

To start with this kind of procedure, the *battery* product feature initial terms are selected collected from the technical specifications, and each one has been counted. Another featured carried out is that when a not initial term, which has a part of one of them, such as *battery life* or *video playback hd*, if they reach the minimum of at least five times within the random corpus selected.

The co-occurrence between pairs of initial terms and any term is calculated by the percentages in both ways, how many times appear the first with the second, and vice versa, which with a high percentage in both way compose the relation between them. As it is done before, it is set a threshold of at least 2 occurrences for initial terms and 5 occurrences for new terms in the mini corpus to be considered as an analyzable term. **Table 5-1** shows the results obtained for the *battery* product feature, where it is seen how there are terms, e.g. *phone*, that occurs many times and due to that it has a poor relation with the other ones. It is selected the relations in both directions that satisfy at least 50% of co-occurrence, getting as new terms *hour, day* and *stand-by*, meanwhile *hd video playback, battery life* and *stand-by* have been analyzed because they appear more than 5 times in the selected reviews.

As a result, to complete the semantic network in **Figure 5-1** is presented below of the *battery* product feature, the new terms are added to it. In this case, new terms are extracted through at least one of the initial product features, although it is always said that they all correspond to the *battery* product feature, it is supposed that if it is looked for from the initial

product features which are extracted from the technical specifications of the *battery* product feature, it is deduced that they are like a extension of the *battery* product feature, without being the *battery* product itself.

Summarization battery	
Terms with 0 hits or less than 50% of relation from its side: • li-ion • stand-by time • video call time • capacity • stand-by 3g • li-polymer • talk time 3g • music playback • video playback	battery (24 hits) × battery ($^{12}/_{24} \cdot 100 = \mathbf{50}\% \leftrightarrow {}^{12}/_{39} \cdot 100 = \mathbf{30.76}\%$) manufacturer battery life(6 hits) × battery life($^{3}/_{6} \cdot 100 = \mathbf{50}\% \leftrightarrow {}^{8}/_{103} \cdot 100 = \mathbf{7.76}\%$) phone × battery life($^{3}/_{6} \cdot 100 = \mathbf{50}\% \leftrightarrow {}^{3}/_{42} \cdot 100 = \mathbf{7.14}\%$) smart-phone talk time (8 hits) ✓ talk time ($^{8}/_{8} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{12}/_{16} \cdot 100 = \mathbf{75}\%$) hour × talk time ($^{5}/_{8} \cdot 100 = \mathbf{62.5}\% \leftrightarrow {}^{7}/_{24} \cdot 100 = \mathbf{29.16}\%$) battery ✓ talk time ($^{5}/_{8} \cdot 100 = \mathbf{62.5}\% \leftrightarrow {}^{7}/_{10} \cdot 100 = \mathbf{70}\%$) stand-by × talk time ($^{5}/_{8} \cdot 100 = \mathbf{62.5}\% \leftrightarrow {}^{6}/_{103} \cdot 100 = \mathbf{5.82}\%$) phone × talk time ($^{5}/_{8} \cdot 100 = \mathbf{62.5}\% \leftrightarrow {}^{5}/_{39} \cdot 100 = \mathbf{12.82}\%$) manufacturer ✓ talk time ($^{6}/_{8} \cdot 100 = \mathbf{75}\% \leftrightarrow {}^{8}/_{15} \cdot 100 = \mathbf{53.33}\%$) day stand-by (10 hits) × stand-by ($^{7}/_{10} \cdot 100 = \mathbf{70}\% \leftrightarrow {}^{10}/_{103} \cdot 100 = \mathbf{9.71}\%$) phone ✓ stand-by ($^{10}/_{10} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{13}/_{16} \cdot 100 = \mathbf{81.25}\%$) hour ✓ stand-by ($^{8}/_{10} \cdot 100 = \mathbf{80}\% \leftrightarrow {}^{11}/_{15} \cdot 100 = \mathbf{73.33}\%$) day × stand-by ($^{8}/_{10} \cdot 100 = \mathbf{80}\% \leftrightarrow {}^{8}/_{24} \cdot 100 = \mathbf{33.33}\%$) battery ✓ stand-by ($^{7}/_{10} \cdot 100 = \mathbf{70}\% \leftrightarrow {}^{5}/_{8} \cdot 100 = \mathbf{6.2.5}\%$) talk time × stand-by ($^{6}/_{10} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{6}/_{39} \cdot 100 = \mathbf{15.38}\%$) manufacturer hd video playback (5 hits) × hd video playback ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{2}/_{120} \cdot 100 = \mathbf{1.66}\%$) handset × hd video playback ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{2}/_{11} \cdot 100 = \mathbf{18.18}\%$) test

Table 5-1. Summarization of the co-occurrence from the *battery* initial product features.

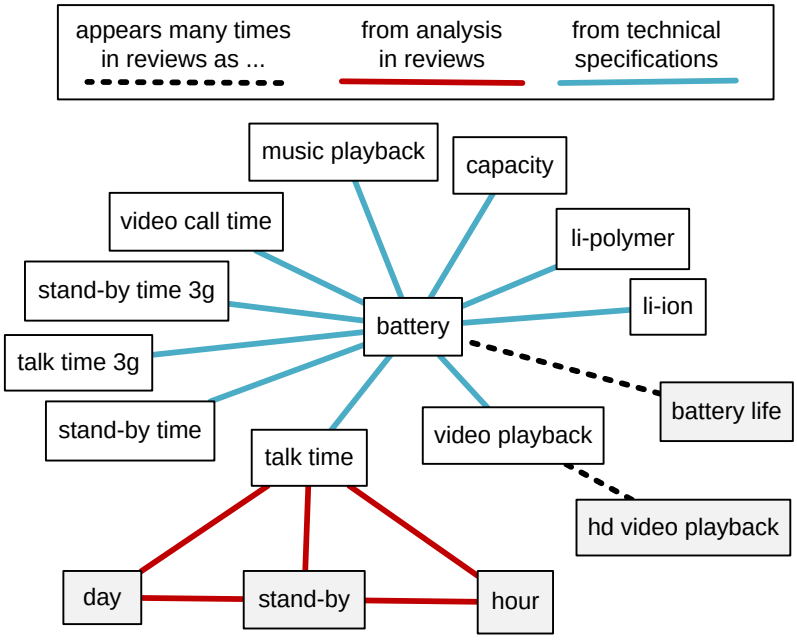


Figure 5-1. Semantic network of the battery product feature from the manual extraction.

This little analysis is presented here to demonstrate how can be perform in a small scenario, without scores, dimensions, precision and recall, obtaining a little result that introduces the main concept, which is to extract semantic relations from the product features. The named tools such as scores, dimensions, etc., are used by the automatic methods to save time and to cover such an extremely tedious task.

5.1.2. Organizer

The same process is followed with the organizer product feature, whose initial terms are analyzed and summarized in **Table 5-2**. In this section it is found that having more initial terms does not imply that there are more probabilities of finding strong semantic related terms, indeed, it is found a single term, *timer*, and a new relation between *alarm* and *world clock*, which does not increase the first semantically related group of *organizer* terms.

Summarization organizer	
Terms with 0 hits or less than 50% of relation from its side: • calendar • to-do • task • document viewer • office 2003	alarm (5 hits) ✓ alarm ($^{5}/_{5} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{4}/_{5} \cdot 100 = \mathbf{80}\%$) hour × alarm ($^{5}/_{5} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{5}/_{16} \cdot 100 = \mathbf{31.25}\%$) manufacturer
	office 2007(2 hits) × office 2007 ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{41} \cdot 100 = \mathbf{4.87}\%$) software × office 2007 ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{7} \cdot 100 = \mathbf{28.57}\%$) document × office 2007 ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{88} \cdot 100 = \mathbf{2.27}\%$) application
	pdf (2 hits) × pdf ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{9} \cdot 100 = \mathbf{22.22}\%$) office × pdf ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{7} \cdot 100 = \mathbf{28}.\%$) document
	calculator (4 hits) × calculator ($^{4}/_{4} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{4}/_{16} \cdot 100 = \mathbf{25}\%$) calendar × calculator ($^{2}/_{4} \cdot 100 = \mathbf{50}\% \leftrightarrow {}^{2}/_{12} \cdot 100 = \mathbf{16.66}\%$) task × calculator ($^{4}/_{4} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{5} \cdot 100 = \mathbf{40}\%$) world clock
	world clock (5 hits) × world clock ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{5}/_{88} \cdot 100 = \mathbf{5.68}\%$) application × world clock ($^{4}/_{5} \cdot 100 = \mathbf{80}\% \leftrightarrow {}^{6}/_{16} \cdot 100 = \mathbf{37.50}\%$) calendar × world clock ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{2}/_{12} \cdot 100 = \mathbf{16.66}\%$) task ✓ world clock ($^{4}/_{5} \cdot 100 = \mathbf{80}\% \leftrightarrow {}^{5}/_{5} \cdot 100 = \mathbf{100}\%$) alarm
	stopwatch (2 hits) × stopwatch ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{4}/_{88} \cdot 100 = \mathbf{4.54}\%$) application × stopwatch ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{16} \cdot 100 = \mathbf{12.50}\%$) calendar × stopwatch ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{3}/_{12} \cdot 100 = \mathbf{25}\%$) task × stopwatch ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{5} \cdot 100 = \mathbf{40}\%$) world clock × stopwatch ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{2} \cdot 100 = \mathbf{100}\%$) timer × stopwatch ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{22} \cdot 100 = \mathbf{9.09}\%$) facebook
	timer (2 hits) × timer ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{5} \cdot 100 = \mathbf{40}\%$) world clock ✓ timer ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{2} \cdot 100 = \mathbf{100}\%$) stopwatch

Table 5-2. Summarization of the co-occurrence from the *battery* initial product features.

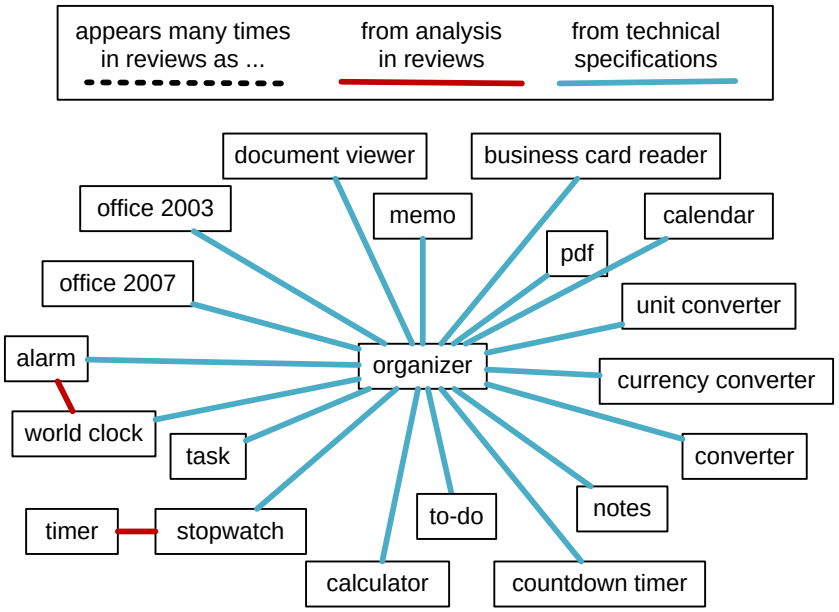


Figure 5-2. Semantic network of the *organizer* product feature from the manual extraction

Finally, in **Figure 5-2** the semantic network resulting from the process done is shown graphically. Although in this analysis *organizer* initial product features appear many times together, the rest of terms do not or their percentage of co-occurence is lower. Anyway, it is a single sample of how the same objective can be performed by hand and how the necessity of do it automatically is crucial after seeing how hard is to find something relevant following this kind of manual extraction.

5.1.3. Multimedia

Finally, the process is run on the *multimedia* initial product features, whose terms form a group which collects up to 42 terms. This is the product feature that more terms has, but results still remain in only several new terms semantically related with the first ones.

In **Table 5-3**, results of the analysis are summarized and there it is possible to see how many initial terms have less than 2 hits within the little corpues, then, they are removed from the analysis. On the other hand, there are more terms like with the other main product features that at first co-occur with many other terms, but at the end, the relation is not established in both ways. Thus, those terms are not considered to be strongly semantically related.

Like it occurs above, there are terms such as *hd video playback*, *radio fm*, *stereo headset* and *last.fm*, which are not initial product features, but they are considered as initial terms due to their lexical construction, although there are not relevant in the final extraction of new terms. Finally, *codec* and *movie* are the only new semantically related terms.

In **Figure 5-3**, it is shown graphically how the semantic network will be built, with the exception of the false initial terms, that are supposed to be related with the first ones, but they are not counted as new semantic terms.

Summarization multimedia	
Terms with 0 hits or less than 50% of relation from its side: • multimedia • music player • mp3 • aac • wma • wav • m4a • amr • midi • acc+ • eacc • eacc+ • ra • realaudio 7 • realaudio 8 • realaudio 10 • ogg • album art cover • background playback • music recognition • video player • video playback • wmv • avi • 3gp • 3g2 • 3gpp • realvideo 7 • realvideo 8 • realvideo 10 • flash video • fm transmitter • radio • fm radio • stereo headset • streaming audio • youtube player	mp4 (2 hits) × mp4 ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{120} \cdot 100 = \mathbf{1.66}\%$) handset × mp4 ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{18} \cdot 100 = \mathbf{11.11}\%$) file × mp4 ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{9} \cdot 100 = \mathbf{22.22}\%$) mpeg4 × mp4 ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{5} \cdot 100 = \mathbf{40}\%$) codec × mp4 ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{12} \cdot 100 = \mathbf{16.67}\%$) divx × mp4 ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{10} \cdot 100 = \mathbf{20}\%$) xvid
	mpeg4 (9 hits) × mpeg4 ($^{6}/_{9} \cdot 100 = \mathbf{66.66}\% \leftrightarrow {}^{11}/_{87} \cdot 100 = \mathbf{12.64}\%$) video × mpeg4 ($^{5}/_{9} \cdot 100 = \mathbf{55.55}\% \leftrightarrow {}^{6}/_{18} \cdot 100 = \mathbf{33.33}\%$) file × mpeg4 ($^{5}/_{9} \cdot 100 = \mathbf{55.55}\% \leftrightarrow {}^{7}/_{39} \cdot 100 = \mathbf{17.95}\%$) resolution ✓ mpeg4 ($^{5}/_{9} \cdot 100 = \mathbf{55.55}\% \leftrightarrow {}^{7}/_{12} \cdot 100 = \mathbf{58.33}\%$) divx × mpeg4 ($^{5}/_{9} \cdot 100 = \mathbf{55.55}\% \leftrightarrow {}^{4}/_{10} \cdot 100 = 40\%$) xvid
	hd video playback (5 hits) × hd video playback ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{2}/_{120} \cdot 100 = \mathbf{1.66}\%$) handset × hd video playback ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{2}/_{11} \cdot 100 = \mathbf{18.18}\%$) test
	h.263 (3 hits) × h.263 ($^{2}/_{3} \cdot 100 = \mathbf{66.66}\% \leftrightarrow {}^{2}/_{120} \cdot 100 = \mathbf{1.66}\%$) handset × h.263 ($^{3}/_{3} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{3}/_{9} \cdot 100 = \mathbf{33.33}\%$) mpeg4 × h.263 ($^{3}/_{3} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{4}/_{18} \cdot 100 = \mathbf{22.22}\%$) file × h.263 ($^{2}/_{3} \cdot 100 = \mathbf{66.66}\% \leftrightarrow {}^{2}/_{39} \cdot 100 = \mathbf{5.12}\%$) resolution × h.263 ($^{2}/_{3} \cdot 100 = \mathbf{66.66}\% \leftrightarrow {}^{2}/_{24} \cdot 100 = \mathbf{8.33}\%$) pixel × h.263 ($^{2}/_{3} \cdot 100 = \mathbf{66.66}\% \leftrightarrow {}^{3}/_{87} \cdot 100 = \mathbf{3.44}\%$) video × h.263 ($^{2}/_{3} \cdot 100 = \mathbf{66.66}\% \leftrightarrow {}^{4}/_{98} \cdot 100 = \mathbf{4.08}\%$) device × h.263 ($^{2}/_{3} \cdot 100 = \mathbf{66.66}\% \leftrightarrow {}^{4}/_{12} \cdot 100 = \mathbf{33.33}\%$) divx × h.263 ($^{2}/_{3} \cdot 100 = \mathbf{66.66}\% \leftrightarrow {}^{2}/_{10} \cdot 100 = \mathbf{20}\%$) xvid × h.263 ($^{2}/_{3} \cdot 100 = \mathbf{66.66}\% \leftrightarrow {}^{2}/_{5} \cdot 100 = \mathbf{40}\%$) h.264 ✓ h.263 ($^{2}/_{3} \cdot 100 = \mathbf{66.66}\% \leftrightarrow {}^{2}/_{2} \cdot 100 = \mathbf{100}\%$) movie × h.263 ($^{2}/_{3} \cdot 100 = \mathbf{66.66}\% \leftrightarrow {}^{2}/_{41} \cdot 100 = \mathbf{4.87}\%$) software
	h.264 (5 hits) × h.264 ($^{4}/_{5} \cdot 100 = \mathbf{80}\% \leftrightarrow {}^{7}/_{87} \cdot 100 = \mathbf{8.04}\%$) video × h.264 ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{5}/_{12} \cdot 100 = \mathbf{41.67}\%$) divx × h.264 ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{3}/_{10} \cdot 100 = \mathbf{30}\%$) xvid × h.264 ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{4}/_{9} \cdot 100 = \mathbf{44.44}\%$) mpeg4 × h.264 ($^{4}/_{5} \cdot 100 = \mathbf{80}\% \leftrightarrow {}^{4}/_{39} \cdot 100 = \mathbf{10.25}\%$) resolution × h.264 ($^{2}/_{5} \cdot 100 = \mathbf{20}\% \leftrightarrow {}^{3}/_{41} \cdot 100 = \mathbf{7.31}\%$) software
	xvid (10 hits) × xvid ($^{5}/_{10} \cdot 100 = \mathbf{80}\% \leftrightarrow {}^{8}/_{38} \cdot 100 = \mathbf{21.05}\%$) support × xvid ($^{7}/_{10} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{2}/_{12} \cdot 100 = \mathbf{16.67}\%$) divx × xvid ($^{5}/_{10} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{5}/_{5} \cdot 100 = \mathbf{100}\%$) codec × xvid ($^{5}/_{10} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{5}/_{120} \cdot 100 = \mathbf{4.17}\%$) handset × xvid ($^{5}/_{10} \cdot 100 = \mathbf{80}\% \leftrightarrow {}^{8}/_{18} \cdot 100 = \mathbf{44.44}\%$) file
	divx (12 hits) × divx ($^{9}/_{12} \cdot 100 = \mathbf{75}\% \leftrightarrow {}^{6}/_{38} \cdot 100 = \mathbf{21.05}\%$) support ✓ divx ($^{12}/_{12} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{7}/_{10} \cdot 100 = \mathbf{16.67}\%$) xvid ✓ divx ($^{7}/_{12} \cdot 100 = \mathbf{58.33}\% \leftrightarrow {}^{5}/_{9} \cdot 100 = \mathbf{100}\%$) mpeg4 × divx ($^{6}/_{12} \cdot 100 = \mathbf{50}\% \leftrightarrow {}^{4}/_{41} \cdot 100 = \mathbf{4.17}\%$) software × divx ($^{7}/_{12} \cdot 100 = \mathbf{58.33}\% \leftrightarrow {}^{5}/_{120} \cdot 100 = \mathbf{44.44}\%$) handset × divx ($^{9}/_{12} \cdot 100 = \mathbf{75}\% \leftrightarrow {}^{8}/_{18} \cdot 100 = \mathbf{44.44}\%$) file

	last.fm (5 hits) × last.fm ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{4}/_{114} \cdot 100 = \mathbf{3.51}\%$) feature × last.fm ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{2}/_{120} \cdot 100 = \mathbf{1.67}\%$) handset × last.fm ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{2}/_{17} \cdot 100 = \mathbf{11.76}\%$) social network × last.fm ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{2}/_{17} \cdot 100 = \mathbf{11.76}\%$) player
	rds (5 hits) × rds ($^{5}/_{5} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{5}/_{16} \cdot 100 = \mathbf{31.25}\%$) fm radio × rds ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{5}/_{114} \cdot 100 = \mathbf{4.39}\%$) feature
	stereo (2 hits) × stereo ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{14} \cdot 100 = \mathbf{14.29}\%$) loudspeaker × stereo ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{16} \cdot 100 = \mathbf{12.50}\%$) 3.5mm jack
	streaming video (2 hits) × streaming video ($^{2}/_{2} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{2}/_{17} \cdot 100 = \mathbf{11.76}\%$) handset
	codec (5 hits) × codec ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{2}/_{122} \cdot 100 = \mathbf{1.64}\%$) screen × codec ($^{4}/_{5} \cdot 100 = \mathbf{80}\% \leftrightarrow {}^{9}/_{87} \cdot 100 = \mathbf{10.34}\%$) video × codec ($^{5}/_{5} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{3}/_{12} \cdot 100 = \mathbf{25}\%$) divx ✓ codec ($^{5}/_{5} \cdot 100 = \mathbf{100}\% \leftrightarrow {}^{5}/_{10} \cdot 100 = \mathbf{50}\%$) xvid × codec ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{3}/_{9} \cdot 100 = \mathbf{33.33}\%$) mpeg4 × codec ($^{4}/_{5} \cdot 100 = \mathbf{80}\% \leftrightarrow {}^{4}/_{39} \cdot 100 = \mathbf{10.26}\%$) resolution × codec ($^{4}/_{5} \cdot 100 = \mathbf{80}\% \leftrightarrow {}^{3}/_{24} \cdot 100 = \mathbf{12.50}\%$) pixel × codec ($^{3}/_{5} \cdot 100 = \mathbf{60}\% \leftrightarrow {}^{2}/_{41} \cdot 100 = \mathbf{4.88}\%$) software

Table 5-3. Summarization of the co-occurrence of the *multimedia* initial product features.

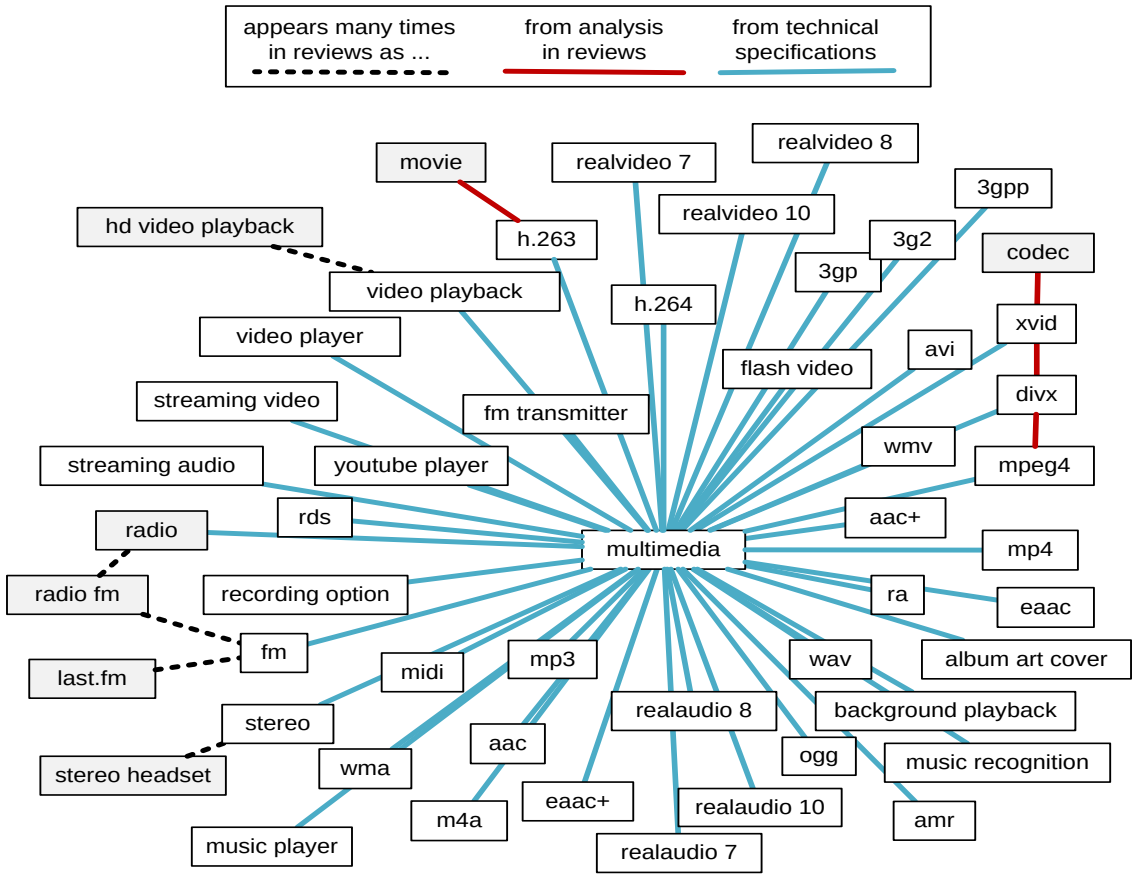


Figure 5-3. Semantic network of the *multimedia* product feature from the manual extraction.

5.2. LSI analysis

5.2.1. Battery

5.2.1.1. Documents

In order to follow the steps described in **Table 4-1**, LSI is run considering the classical environment, on documents, where each review is taken as an entity, obtaining the U_k, S_k and V_k^T matrices, where U_k is the only one which has information relative to terms. In **Figure 5-4**, it is possible to see how each column of the U_k matrix is called a dimension. It is supposed that each dimension has some terms better scores than others, creating a ranking. It is assumed that this ranking implies that the best scored terms define the dimension with some meaning. This meaning could be everything, but the goal is to find the dimension which better define a topic, the *battery* product feature. As it is said, the starting point is the terms extracted from the technical specifications, then, where these initial terms are found best scored and also best placed relatively to the top of a dimension, it will be possible to find more terms which extends the *battery* product feature meaning. These terms will have a semantic relation between them and the initial terms, and this is what is looked for here.

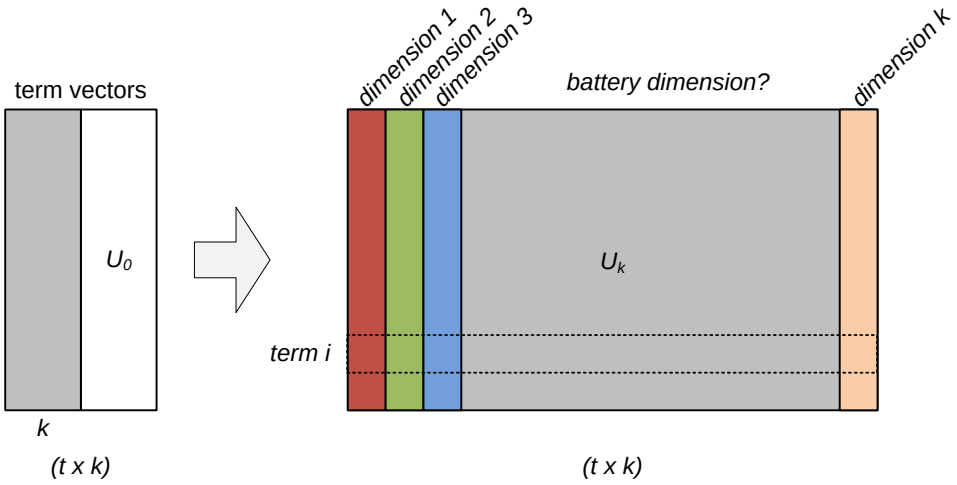


Figure 5-4. How are dimensions organized in the U_k matrix.

It is impossible to show all dimensions of each possibility here, hence, they are organized in tables which contain the most important dimensions found on each result of each performance, which means that they contain the dimensions which have better scored the initial terms of *battery* within the 30 first terms of the dimension. Although only **Table 5-4** is included here, the rest of the tables related to the analysis of LSI performances are included in **Annex A: LSI's running tables**.

To define better how results are represented by this table, it is analyzed the first cell, with *k=50* (row) and without *extra reviews* (column). It shows the best scored dimensions, including the places of initial terms of *battery* in the sorted dimension accompanied by their scores.

There, it is possible to see two dimensions, *dimension 1* and *dimension 5*, accompanied by the value of the top of the dimension, to see how far or close are the initial terms relatively.

Far from placing some initial term in the top 30, it is seen that in *dimension 1*, there are 190 terms better scored than the first initial term found, *battery*, in this case, and more terms until the rest. Then, although initial terms are best scored in this dimension than in others, it means that *dimension 1* does not define the *battery* product feature. The same happens with *dimension 5*, but here there are less initial terms, because the first one already is further than in *dimension 1* from the top of the dimension, and also there are 3 negative scored values of the whole list of 11 initial terms for the *battery* product feature. Finally, it is concluded that any dimension can be extracted as a potential *battery dimension* with *k=50* and without *extra reviews*.

	Without	With
K=50	× **Dim 1** · ____ 0.6014 – top dimension · 191st with 0.2470 – battery · 379th with 0.1855 – capacity · 444th with 0.1738 – video playback · 600th with 0.1439 – talk time · 1220th ... → No negative values × **Dim 5** · ____ 0.7153 – top dimension · 142nd 0.1043 – battery · 205th ... → 3 negative values	× **Dim 1** · ____ 0.6026 – top dimension · 191st with 0.2477 – battery · 385th with 0.1874 – capacity · 455th with 0.1723 – video playback · 597th with 0.1447 – talk time · 1225th ... → No negative values × **Dim 5** · ____ 0.7150 – top dimension · 142nd 0.1049 – battery · 203rd ... → 3 negative values × **Dim 10** · ____ 0.3931 – top dimension · 71th with 0.1115 – capacity · 189th ... → 3 negative values
K=100	× **Dim 1** (same as k=50) × **Dim 10** · ____ 0.3933 – top dimension · 68th with 0.1120 – capacity · 175th ... → 3 negative values × **Dim 5** (negative values)	× **Dim 1** (same as k=50) × **Dim 10** (same as k=50) × **Dim 46** · ____ 0.9598 – top dimension · 22nd with 0.1724 – li-ion · 221st ... → 3 negative values × **Dim 5** (negative values)
K=150	× **Dim 1** (same as k=50, k=100) × **Dim 5** (same as k=50) × **Dim 10** (same as k=100)	× **Dim 1** (same as k=50, k=100) × **Dim 5** (same as k=50) × **Dim 10** (same as k=100) × **Dim 46** (negative values)
K=200	× **Dim 1** (same as k=50, k=100, k=150) × **Dim 10** (same as k=100, k=150) × **Dim 5** (negative values)	× **Dim 1** (same as k=50, k=100, k=150) × **Dim 5** (same as k=50, k=100) × **Dim 10** (same as k=100) × **Dim 46** (same as k=100)

Table 5-4. Best scored dimensions of LSI considering the *battery* product feature on documents.

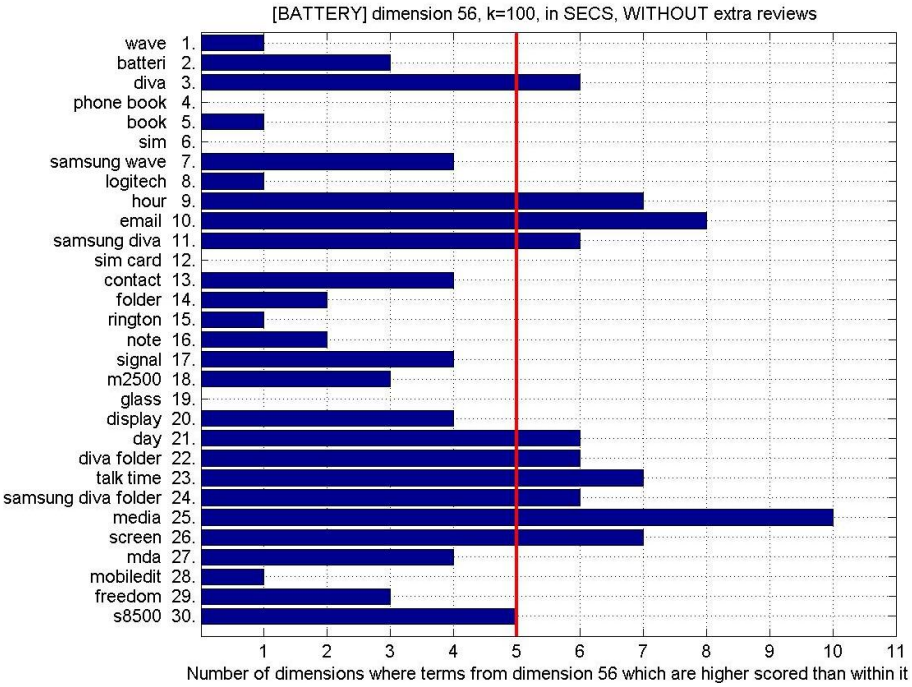
Taking a look at the next performances where the *k* parameter is increased (cell down), it is seen that *capacity* term is well scored relatively to the place that it occupies in *dimension 10* with k=100, besides not being in the top 30. It is also seen that *dimension 1* has the same score that with *k=50*, but surprisingly, terms that where relevant for *dimension 5* with *k=50*, now with *k=100* some of the initial terms have the same value, but negative, and as a consequence, *dimension 5* is worse scored than above. With *k=150*, all the dimensions (1, 5 and 10) remain as they were noticed, and with *k=200* it is had the same situation like with *k=100*, with

dimension 5, wich has higher values negative scored. These negative values may appear, because terms occupy a syntactic place in sentences that can be interpreted with a meaning, with the opposite, or with each one that fits that sentence, and if the method has difficulties scoring these terms, it occurs that sometimes they are positive and sometimes negative.

Observing what occurs when *extra reviews* are included, it is seen that there are not many differences between this column and the first one. It can be appreciated that scores are a bit better in general, but places differ on each dimension. Here, it is found that only *dimension 46*, with *extra reviews* and *k=100*, has 1 of the initial terms in a place higher than the 30^{th}. Accompanied by the low score regarding the top of the dimension and the place of the next initial term, on the 221^{st} place, it means that this dimension, in spite of being the closer to be a *battery dimension*, it cannot be consider as a *battery dimension*. The best scored terms, that usually are *battery* and *talk time*, are the result of a bad performance that bring a poor connexion between initial terms that not define a *battery dimension*. The next step would consist in applying the discarding method on the 30 first potential new terms, analyzing which one has higher scores in other dimensions apart from the analyzed one, but the extraction is stopped here.

5.2.1.2. Sections

The next performances are run on sections, where each review is divided in sections, also the *extra reviews*, keeping the initial terms all in a single section. In **Table 9-2** (located in **Annex A: LSI's running tables**), it is seen that scores and places of the initial terms in dimensions are better than in documents. Although it happens the same with negative and positive values, because sometimes the best values turn into negative sign, it is extracted the best performances of dimensions, paying attention to the values reached in the experiments.



Graphic 5-1. Discarding method applied to top 30 terms from *dimension 56* (without).

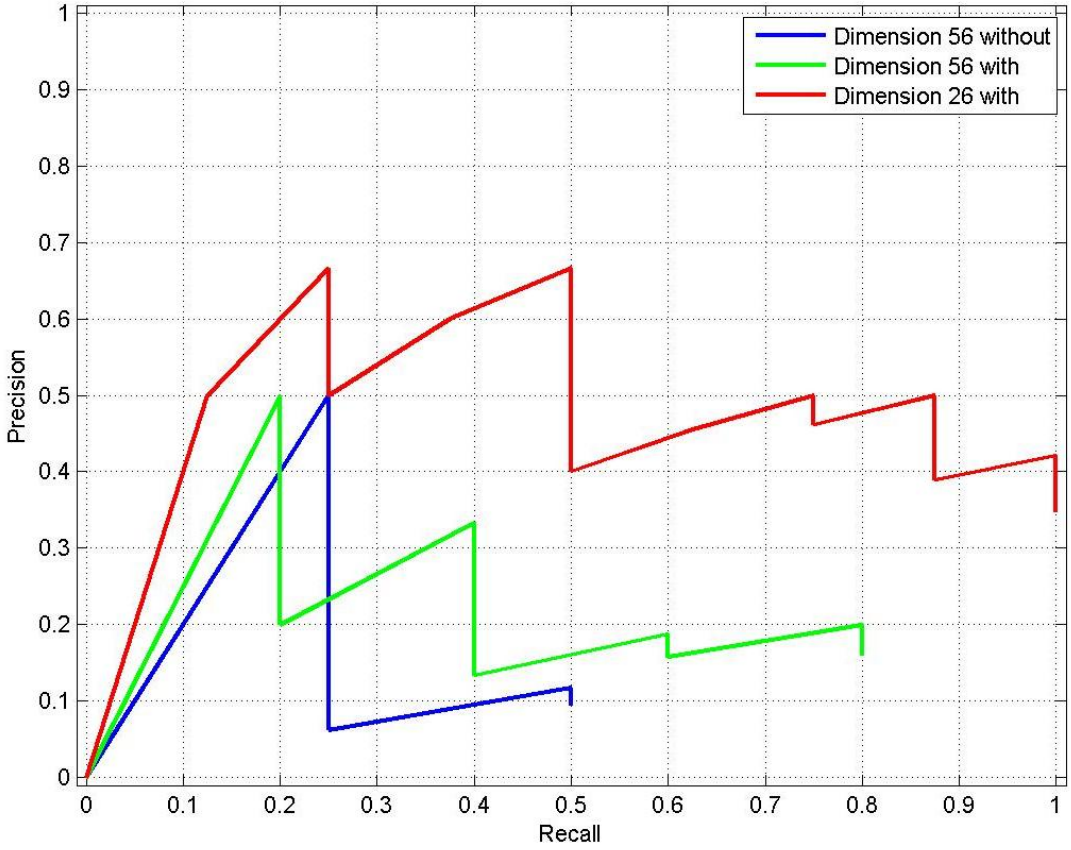
Dimension 12, *24* and *56* without *extra reviews* have initial terms very close to the top, being *dimension 56* which have best placed the initial terms and because of having 2 terms in the top 30 and the rest are all the 11 initial terms scored with positive values, it is selected to extract potential new terms. Although *dimension 12* has a little better scored the two first initial terms, they are further from the top, within top 30, but also it has 3 negative values of the 11 initial terms. The same happens with *dimension 24* and *dimension 25*, where the second term is placed on the 44th and on the 89th, respectively, although *dimension 24* has no negative values and *dimension 25* has only 2 of them. Apart from them, initial terms appear well scored within top 30 in *dimension 26*, but only with *extra reviews*, having only 1 negative value of the 11 initial terms. As a result, *dimension 56* from performances without (*k=100*) and with (*k=100, k=150, k=200*) *extra reviews* and *dimension 26* only without (*k=50, k=200*) *extra reviews* are selected as potential *battery dimensions.*

Place	*Dimension 56* without *k=100*	*Dimension 56* with *k=100, 150, 200*	*Dimension 26* with *k=50, 200*
1	0.3375 wave (*fp*)	0.3255 wave (*fp*)	0.4802 lg (*fp*)
2	**0.2914 battery** (***tp***)	**0.2936 battery** (***tp***)	0.3405 hour (***tp***)
3	~~0.2522 diva~~ (***tn***)	0.2512 diva (*fp*)	0.3399 message (***tp***)
4	0.2512 phone book (*fp*)	0.2508 phone book (*fp*)	0.3286 cookie (*fp*)
5	0.2505 book (*fp*)	0.2507 book (*fp*)	0.3022 performance (***tp***)
6	0.2407 sim (*fp*)	0.2408 hour (***tp***)	**0.2895 battery** (***tp***)
7	0.2383 samsung wave (*fp*)	0.2387 sim (*fp*)	~~0.2792 window~~ (***tn***)
8	0.2365 logitech (*fp*)	0.2374 logitech (*fp*)	0.2778 text (*fp*)
9	~~0.2357 hour~~ (*fn*)	~~0.2299 email~~ (***tn***)	0.2767 account (*fp*)
10	~~0.2328 email~~ (***tn***)	0.2277 samsung diva (*fp*)	0.2692 lg cookie (*fp*)
11	~~0.2286 samsung diva~~ (***tn***)	0.2273 samsung wave (*fp*)	~~0.2618 htc~~ (***tn***)
12	0.2234 sim card (*fp*)	0.2208 sim card (*fp*)	0.2518 k850 (*fp*)
13	0.2141 contact (*fp*)	~~0.2163 contact~~ (***tn***)	0.2500 talk (***tp***)
14	0.2078 folder (*fp*)	0.2097 ringtone (*fp*)	~~0.2342 email~~ (***tn***)
15	0.2036 ringtone (*fp*)	0.2066 folder (*fp*)	~~0.2199 keyboard~~ (***tn***)
16	0.1954 note (*fp*)	0.1987 note (*fp*)	**0.2140 talk time** (***tp***)
17	0.1810 signal (*fp*)	0.1799 signal (*fp*)	0.2045 sound (*fp*)
18	0.1730 m2500 (*fp*)	0.1713 day (***tp***)	0.2040 hardware (***tp***)
19	0.1706 glass (*fp*)	0.1709 glass (*fp*)	0.2028 corby (*fp*)
20	0.1703 display (*fp*)	0.1698 display (*fp*)	~~0.2009 quality~~ (***tn***)
21	~~0.1677 day~~ (*fn*)	0.1662 m2500 (*fp*)	~~0.1985 window mobile~~ (***tn***)
22	~~0.1595 diva folder~~ (***tn***)	**0.1608 talk time** (***tp***)	~~0.1931 explorer~~ (***tn***)
23	**0.1549 talk time** (*tp*)	0.1590 diva folder (*fp*)	0.1871 cookie fresh (*fp*)
24	~~0.1528 samsung diva folder~~ (***tn***)	~~0.1533 screen~~ (***tn***)	0.1870 client (*fp*)
25	~~0.1518 media~~ (***tn***)	~~0.1530 media~~ (***tn***)	0.1854 phone (***tp***)
26	~~0.1517 screen~~ (***tn***)	0.1524 samsung diva folder (*fp*)	0.1852 time (***tp***)
27	0.1490 mda (*fp*)	0.1502 mda (*fp*)	0.1809 fresh (*fp*)
28	0.1465 mobiledit (*fp*)	0.1452 freedom (*fp*)	0.1798 input (*fp*)
29	0.1452 freedom (*fp*)	0.1444 mobiledit (*fp*)	0.1787 messenger (*fp*)
30	0.1437 s8500 (*fp*)	~~0.1429 talk~~ (*fn*)	0.1755 result (*fp*)

Table 5-5. Judgement applied over the top 30 terms of *dimensions 56* (without)*, 56* (with) and *26* (without) after discarding method.

Once it is selected the dimensions, as it is said in **Table 4-1**, it is time to apply the discarding method, which analyze the behaviour of the top 30 terms in other dimensions. If it is found a term whose scores in more than 5 different dimensions than the analyzed are higher, it satisfies the condition to be discarded. In **Graphic 5-1** it is shown graphically how some terms of the top 30, in *dimension 56* without *extra reviews* and *k=100 ,* have scored better than *dimension 56* more number of dimensions than this threshold. These terms are

diva, hour, email, samsung diva, day, diva folder, talk time (this last one is not discarded, because it is an initial term)*, samsung diva folder, media* and *screen*. Except *talk time*, it is understood that these 9 terms are discarded. The same process is done with *dimension 56,* with *extra reviews* and *k=200*, and *dimension 26*, with *extra reviews* and *k=200*, whose graphics are located in **Annex D: Discarding method applied to LSI dimensions**. The highest value of the *k* parameter is taken to include more possible dimensions with higher scores in the discarding method.



Graphic 5-2. Precision and recall graph of selected dimensions of *battery* on sections.

After choosing dimensions and discarding terms, it is given an interpretation as a human being of the results obtained by LSI. This judgement is presented in **Table 5-5**, represented by: *true positive (tp),* when a top 30 term is correct retrieved, it means that there is such a latent semantic relation with one or more initial terms in the context of the *battery* product feature in the smart-phones field; *false positive (fp),* if a top 30 term is an unexpected result, a term that is thought to be semantically related due to it is within the top 30 terms of the dimensions; *true negative (tn),* a correct discarded term from the discarding method and also, it should be discarded semantically*;* and *false negative (fn)*, a missing term that has been discarded, but it is semantically related, as the *true positive* ones. Initial terms are always *true positive*, because they have been returned within the top 30, and because of that the selected dimensions have been selected. In *dimension 56* without *extra reviews*, almost the *positives* are *false*, except the named initial reviews. However, terms such as *hour* or *day* are bad discarded by the method; the other ones are good discarded. Meanwhile, when *extra reviews* are included in *dimension 56*, *hour* and *day* appear as *true positive*, but *talk* appears as *false*

negative. The rest are good discarding or *false positives*. The best precision is found in *dimension 26* with *extra reviews*, where *hour, message, performance, talk* and *hardware* are found as true positive, and there is no *false negative*, it means that discarding method has worked right, except that there are a lot of *false positive* yet.

Once it is had the table with the evaluation of terms, it can be calculated measures such as precision, recall and accuracy of these dimensions of the top 30 terms. It is assumed that LSI has scored these terms in these dimensions to define the *battery* product feature, then, the goal is to see which dimension defines better *battery,* regarding how far or close are dimensions from the expected results (evaluation). In **Graphic 5-2**, precision vs. recall graph are seen in the different selected scenarios, such as *dimension 56*, with and without *extra reviews*, and *dimension 26*, only with *extra reviews*. In general, it confirms what it was supposed in **Table 5-5**, with the evaluation of terms, that *dimension 26* performs better than the others, with a considerably improvement in all the calculated measures, e.g. being the only one whose recall reaches 1 (in this top 30 terms). The rest of the graphs are included in **Annex G: Precision, recall and accuracy graphs of LSI** where precision, recall and accuracy are presented in single graphics. However, any conclusion can be extracted from these graphics, because in each performance, precision and accuracy starts from the bottom, and they increase their values, but they do not remain with the highest values until any term, then, any term neither percentage is extracted from this performance of LSI.

5.2.1.3. Paragraphs

After running LSI on paragraphs, **Table 9-3** (located in **Annex A: LSI's running tables**) shows the scores given by the method to the best scored dimensions.

Firstly, it is selected the dimensions which collect the best places and scores at the same time, regarding all the scores of the initial terms relative to the *battery* product feature. Dimensions such as *dimension 17,* in the unique performance with *k=150* without *extra reviews*, having 1 negative initial term, *capacity* scored with -0.0670, and *dimension 18,* in performances with *k=100,150,200* with and without *extra reviews*, which have all the initial terms scored positive, are the dimensions selected as the closest ones to define themselves as *battery dimensions.* The rest of dimensions have lower scores for the initial terms to be considered here or have more or equal than 2 negative values of the list of 11 initial values, which suppose more than 18% of the terms which define the dimension, do not define it, thus they are discarded. The particular case of discarding *dimension 7* is due to its 3rd scored initial term is on the 462nd place, then, having only 2 of the 11 initial terms on the top 30 terms.

Once dimensions are selected, it is possible to apply the discarding method as it is done on sections before. In **Annex D: Discarding method applied to LSI dimensions** is presented the behaviours of top 30 terms of these selected dimensions, where it is seen how in this case, any initial term is discarded by the discarding method. In *dimension 17*, the percentage of discarded terns is up to 30%, but not all of them are good discarded, as it is shown in the **Table 5-6**, where the evaluation as a human being is given. In *dimension 18*, there are many differences when the method is applied, because it goes from the 23.3% of discarded terms

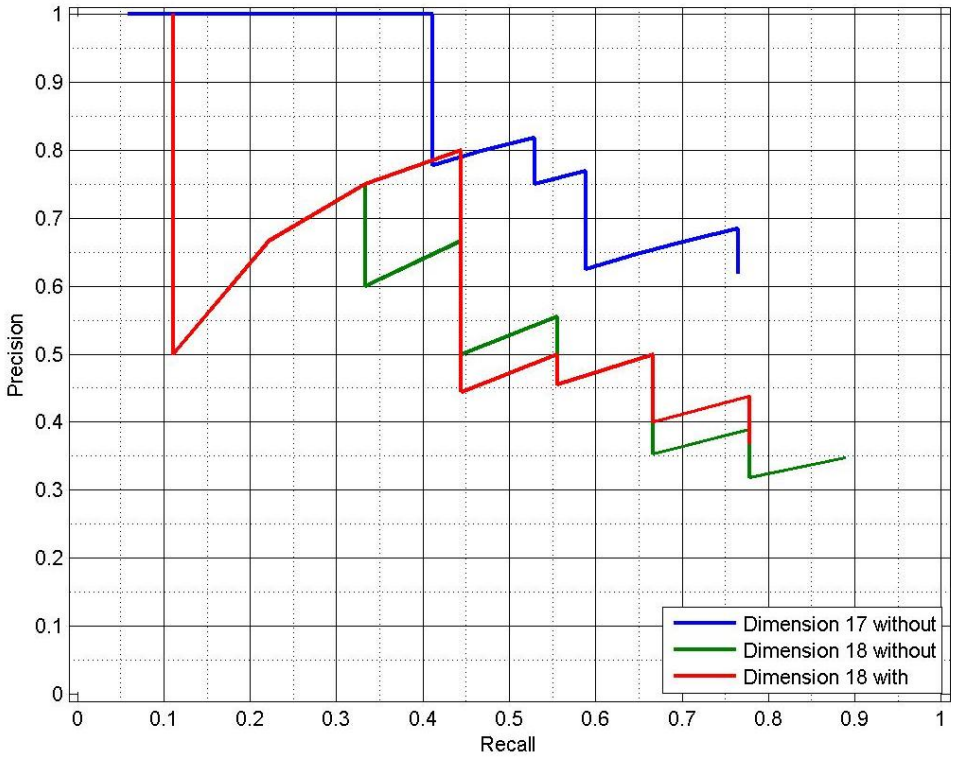
without *extra reviews*, up to the 40% with them. As it happens with *dimension 17*, in *dimension 18*, not all the discarded terms are well discarded and it is found some false negative terms.

Place	*Dimension 17* without *k=100*	*Dimension 18* without *k=100, 150, 200*	*Dimension 18* with *k=100, 150, 200*
1	0.9383 hour (***tp***)	**0.6716 battery** (***tp***)	**0.6550 battery** (***tp***)
2	**0.7465 battery** (***tp***)	0.5249 card (*fp*)	0.5215 card (*fp*)
3	0.5164 talk (***tp***)	0.5030 hour (***tp***)	0.4868 hour (***tp***)
4	**0.4850 talk time** (***tp***)	0.4581 message (***tp***)	0.4675 message (***tp***)
5	0.4649 gsm (***tp***)	0.4034 samsung (*fp*)	~~0.4108 samsung~~ (***tn***)
6	0.3599 network (***tp***)	0.3958 talk (***tp***)	0.3941 talk (***tp***)
7	0.3472 day (***tp***)	0.3698 application (*fp*)	0.3745 application (*fp*)
8	0.3199 tmobile (*fp*)	0.3582 menu (*fp*)	0.3577 menu (*fp*)
9	~~0.2948 time~~ (*fn*)	0.3454 slot (***tp***)	0.3434 slot (***tp***)
10	0.2926 standby (***tp***)	0.3364 home screen (*fp*)	~~0.3301 memory~~ (***tn***)
11	~~0.2892 icon~~ (*fn*)	0.3310 memory (*fp*)	0.3282 home screen (*fp*)
12	0.2864 edge (***tp***)	**0.3148 talk time** (***tp***)	**0.3143 talk time** (***tp***)
13	0.2676 umts (***tp***)	~~0.3056 icon~~ (*fn*)	~~0.2990 icon~~ (*fn*)
14	0.2589 data (*fp*)	0.3009 home (*fp*)	0.2921 home (*fp*)
15	~~0.2443 page~~ (***tn***)	0.2890 microsd (*fp*)	~~0.2888 game~~ (***tn***)
16	~~0.2397 contact~~ (***tn***)	~~0.2874 game~~ (***tn***)	0.2879 microsd (*fp*)
17	0.2386 battery life (***tp***)	~~0.2666 interface~~ (***tn***)	~~0.2664 interface~~ (***tn***)
18	~~0.2375 music~~ (***tn***)	0.2626 theme (*fp*)	0.2593 theme (*fp*)
19	0.2371 at&t (*fp*)	0.2551 life (*fp*)	~~0.2542 nokia~~ (***tn***)
20	0.2359 life (*fp*)	0.2491 program (*fp*)	0.2517 program (*fp*)
21	0.2248 mhz (*fp*)	0.2489 standby (***tp***)	0.2514 standby (***tp***)
22	0.2248 gsm phone (***tp***)	~~0.2469 nokia~~ (***tn***)	0.2480 life (*fp*)
23	~~0.2204 music player~~ (*fn*)	0.2331 sim (*fp*)	~~0.2366 document~~ (***tn***)
24	~~0.2159 player~~ (*fn*)	~~0.2323 document~~ (***tn***)	~~0.2314 multimedia~~ (***tn***)
25	0.2096 wifi (***tp***)	~~0.2294 multimedia~~ (***tn***)	0.2299 sim (*fp*)
26	0.2072 connection (***tp***)	0.2279 sim card (*fp*)	0.2247 sim card (*fp*)
27	0.2070 usa (*fp*)	0.2276 signal (*fp*)	~~0.2247 signal~~ (***tn***)
28	0.2028 album (*fp*)	0.2168 excel (*fp*)	~~0.2207 phone~~ (*fn*)
29	~~0.1969 windows~~ (***tn***)	~~0.2153 system~~ (***tn***)	~~0.2205 excel~~ (***tn***)
30	~~0.1957 computer~~ (***tn***)	0.2136 battery life (***tp***)	~~0.2176 system~~ (***tn***)

Table 5-6. Judgement applied over the top 30 terms of *dimensions 17* (without)*, 18* (without) and *18* (with) after discarding method.

As it was thought after showing **Table 5-6**, in **Graphic 5-3** it is shown graphically how *dimension 17* performs much better than *dimension 18* in both performances, with and without *extra reviews*. This last dimension improves a little its performance when *extra reviews* are included, but not enough to overcome *dimension 17*, then, *dimension 17* is the dimension selected to extract from it semantic relations. The rest of the graphs are included in **Annex G: Precision, recall and accuracy graphs of LSI** where precision, recall and accuracy are presented in single graphics.

According to the threshold of precision over 80% of precision, in ***dimension 17* without** *extra reviews*, terms such as *hour*, *talk*, *gsm*, *network* and *day*, are extracted with a precision and accuracy of 100%, but a recall only over 41.18%. LSI has been performed on each scenario, on documents, sections and paragraphs, but this last one is the best performance of the three of them, being the only one where significant results have been reached. Recall does not reach the total of its value, this occurs due to potential semantically related terms have been discarded by the discarding method, and then it never takes the complete 100% in graphics.



Graphic 5-3. Precision and recall graph of selected dimensions of *battery* on paragraphs.

5.2.2. Organizer

5.2.2.1. Documents

In this part, it is talked about the *organizer* product feature, taking the same steps described in **Table 4-1** to achieve the goal of extracting the major number of well semantically related terms with the initial terms from the *organizer* product feature. The first analysis, as it is done with the *battery* feature, is done on documents, considering each review as an entity to input of LSI.

Looking inside **Table 9-4**, the best scored dimension is *dimension 1*, like it occurs with *battery* on documents. There it is seen how are scored the initial terms of *organizer*, getting values up to two times better the scores given to the *battery* ones. This does not mean that *dimension 1* defines the *organizer* product feature, because although they are much better scored, they are placed over the 39th place, if it is seen as an ordered dimension, and this fact means that it is very difficult to assume that this dimension refers to the *organizer* topic. The same happens with dimensions such as *dimension 3* and *dimension 10*, but they have not scored the initial terms as well as *dimension 1*. However, initial terms like *memo* and *task* achieve the 24th and the 25th respective places in the sorted dimensions, *dimension 6* and *dimension 87*, respectively. Differences between executions with and without *extra reviews* are no appreciable. Therefore, it can be said that, like it occurs with *battery*, it exists any dimension that defines the *organizer* product feature in performances of LSI on documents, at any proposed value of the *k* parameter.

5.2.2.2. Sections

In this part, sections are given as input to the LSI method. Here, it is explained and shown the results of LSI performances. Taking a look at **Table 9-5**, it is found that, in general, initial terms have been scored too much better than in the rest of performances analyzed above. There, *dimension 4* is the first dimension found in performances with (*k=50, 150*) and without (*k=50*) *extra reviews*, having up to 7 initial term within the top 30 without any negative value of the list of 16 initial terms. It is followed by *dimension 8*, which is only found in performances without (*k=200*) *extra reviews* and with (*k=150*) them, having 6 initial terms within the top 30 and 3 negative values of the list. One singular case occurs with *dimension 34*, which has one of the best scores found with *extra reviews*, but without them it has very different scores and initial terms are also not sorted by the same pattern. Another particular case occur with *dimension 35* in all the performances except with values of *k=50, 200* with *extra reviews*, where it is seen that there is any improvement when *extra reviews* are included, but also scores of initial terms are poorer. They are all the dimensions which are supposed to be the *organizer dimensions.*

Running the discarding method, following the steps introduced in **Table 4-1**, it is found that there is any discarded term of the top 30 in *dimension 4* (with and without *extra reviews*), and only 5 in *dimension 8*, some of them are related with smart-phone models and

manufacturers, which are well discarded, but *picture* and *contact* are wrong discarded. In *dimension 34*, there are more discarded without *extra reviews*, what decreases its performance, meanwhile in *dimension 35* occurs the inverse, the number of discarded terms increases when *extra reviews* are included. The resultant graphics are located in **Annex D: Discarding method applied to LSI dimensions**, in section **12.2.1.**

Place	*Dimension 4* without *k=50*	*Dimension 4* with *k=50,150*	*Dimension 8* without *k=200*	*Dimension 8* with *k=150*
1	0.7573 contact (***tp***)	0.7558 contact (***tp***)	0.4792 week (***tp***)	**0.4787 alarm** (***tp***)
2	**0.6480 calendar** (***tp***)	**0.6490 calendar** (***tp***)	**0.4788 alarm** (***tp***)	0.4770 week (***tp***)
3	**0.5613 alarm** (***tp***)	**0.5624 alarm** (***tp***)	**0.4325 note** (***tp***)	**0.4325 note** (***tp***)
4	**0.5102 note** (***tp***)	**0.5111 note** (***tp***)	0.3875 day (***tp***)	0.3865 day (***tp***)
5	0.4771 menu (***tp***)	0.4765 menu (***tp***)	0.3862 menu (***tp***)	0.3840 menu (***tp***)
6	**0.4634 task** (***tp***)	**0.4649 task** (***tp***)	0.3830 field (***tp***)	0.3810 field (***tp***)
7	0.4591 clock (***tp***)	0.4604 clock (***tp***)	0.3815 clock (***tp***)	0.3807 clock (***tp***)
8	0.4437 application (***tp***)	0.4430 application (***tp***)	0.3524 number (***tp***)	0.3500 number (***tp***)
9	0.4391 email (***tp***)	0.4379 email (***tp***)	0.3333 date (***tp***)	**0.3320 calendar** (***tp***)
10	0.4377 week (***tp***)	0.4376 week (***tp***)	**0.3328 calendar** (***tp***)	0.3320 date (***tp***)
11	0.4325 list (***tp***)	0.4318 list (***tp***)	0.3054 month (***tp***)	0.3056 camera (*fp*)
12	0.4188 message (***tp***)	0.4176 message (***tp***)	~~0.3036 ericsson~~ (***tn***)	~~0.3053 ericsson~~ (***tn***)
13	0.4107 option (***tp***)	0.4098 option (***tp***)	0.3033 appointment (***tp***)	0.3039 month (***tp***)
14	0.4046 field (***tp***)	0.4041 field (***tp***)	~~0.3009 camera~~ (***tn***)	0.3015 appointment (***tp***)
15	0.4046 number (***tp***)	0.4038 number (***tp***)	0.2983 duration (***tp***)	~~0.2993 sony ericsson~~ (***tn***)
16	0.3989 file (***tp***)	0.3983 file (***tp***)	~~0.2977 sony ericsson~~ (***tn***)	0.2977 duration (***tp***)
17	0.3948 day (***tp***)	**0.3954 organizer** (***tp***)	~~0.2960 sony~~ (***tn***)	~~0.2975 sony~~ (***tn***)
18	0.3941 name (***tp***)	0.3949 day (***tp***)	0.2912 reminder (***tp***)	0.2901 reminder (***tp***)
19	**0.3941 organizer** (***tp***)	0.3934 name (***tp***)	0.2895 contact (***tp***)	~~0.2855 contact~~ (*fn*)
20	**0.3768 calculator** (***tp***)	**0.3784 calculator** (***tp***)	**0.2848 task** (***tp***)	**0.2844 task** (***tp***)
21	0.3639 month (***tp***)	0.3641 month (***tp***)	0.2739 priority (***tp***)	0.2725 priority (***tp***)
22	0.3588 phone book (***tp***)	0.3577 phone book (***tp***)	0.2692 name (***tp***)	~~0.2670 picture~~ (*fn*)
23	0.3368 date (***tp***)	0.3365 date (***tp***)	0.2686 day week (*fp*)	0.2670 day week (*fp*)
24	0.3312 appointment (***tp***)	0.3313 appointment (***tp***)	~~0.2634 picture~~ (*fn*)	0.2668 name (***tp***)
25	0.3259 home (*fp*)	0.3251 home (*fp*)	**0.2632 organizer** (***tp***)	**0.2634 organizer** (***tp***)
26	0.3138 search (***tp***)	0.3134 search (***tp***)	0.2540 image (***tp***)	0.2576 image (***tp***)
27	0.2939 page (***tp***)	0.2931 page (***tp***)	**0.2503 calculator** (***tp***)	**0.2508 calculator** (***tp***)
28	0.2922 home screen (***tp***)	0.2917 home screen (***tp***)	0.2472 light (*fp*)	0.2503 light (*fp*)
29	**0.2853 office** (***tp***)	**0.2888 office** (***tp***)	0.2346 color (***tp***)	0.2362 color (***tp***)
30	0.2822 windows (*fp*)	**0.2827 world clock** (***tp***)	0.2223 command (*fp*)	0.2210 command (*fp*)

Table 5-7. Judgement applied over the top 30 terms of *dimensions 4* (with and without) and *8* (with and without) after discarding method.

After running the discarding method, the judgement as a human being is applied and their results are shown in tables **Table 5-7** and **Table 5-8**. At first sight, it is seen that there are not many terms discarded, but it does not mean that they are correctly related with the *organizer* product feature. The best case occurs in *dimension 4*, where although any term has been discarded, it is almost perfectly done. Without *extra reviews*, it only has 4 unexpected results (*fp*), but with them it even increases its performance, getting only 3. *Dimension 8* also seems to have strongly related terms, even though discarding some terms wrongly, it includes only these *picture* and *contact* as missing results (*fn*), and some more unexpected results (*fp*) than *dimension 4*. The problem comes with *dimension 34* and *dimension 35*, where the appearance of terms related with smart-phones and manufactures is really high, also in the best performance of these dimensions (*dimension 34* with and *dimension 35* without *extra reviews*).

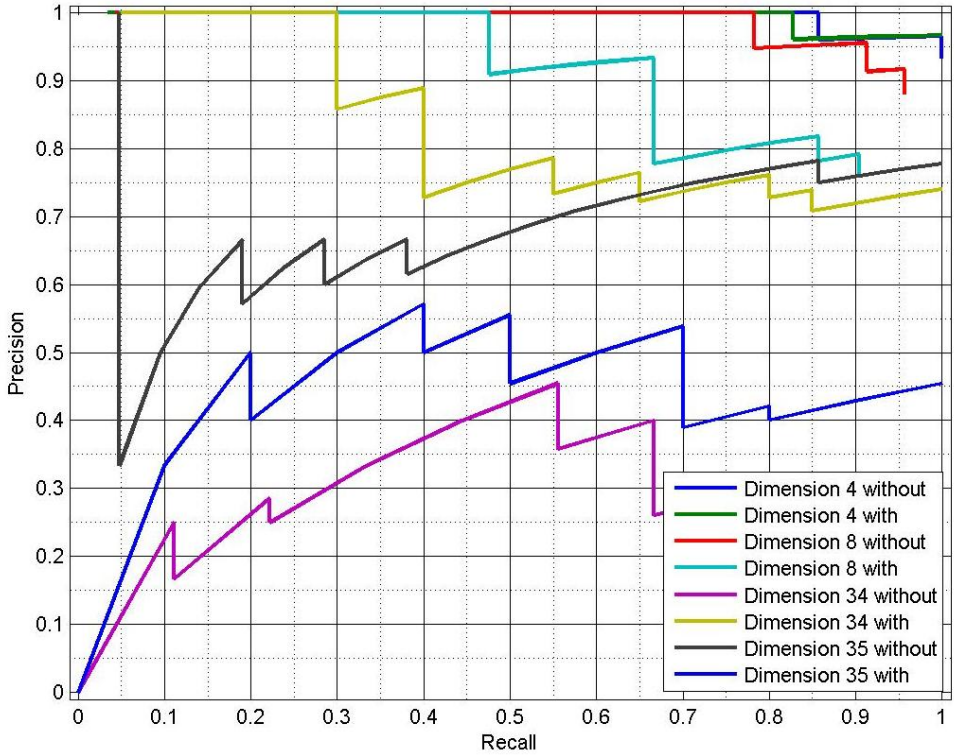
Place	*Dimension 34* without *k=100*	*Dimension 34* with *k=100,150*	*Dimension 35* without *k=50,100,150,200*	*Dimension 35* with *k=100,150*
1	0.3754 new chocolate (*fp*)	**0.4511 alarm** (***tp***)	**0.4625 alarm** (***tp***)	0.5405 samsung (*fp*)
2	0.3437 lg new (*fp*)	**0.3466 organizer** (***tp***)	0.4350 samsung (*fp*)	0.4028 galaxy (*fp*)
3	0.3437 lg new chocolate (*fp*)	**0.3180 calendar** (***tp***)	0.3875 galaxy (*fp*)	**0.3439 alarm** (***tp***)
4	**0.3234 alarm** (***tp***)	0.2935 voice (***tp***)	**0.3482 calculator** (***tp***)	0.3401 android (***tp***)
5	0.3146 chocolate (*fp*)	0.2814 timer (***tp***)	**0.3456 organizer** (***tp***)	0.3346 samsung galaxy (*fp*)
6	0.2818 bl40 (*fp*)	0.2637 page (***tp***)	0.3401 android (***tp***)	**0.2975 calculator** (***tp***)
7	0.2695 voice (***tp***)	0.2557 new chocolate (*fp*)	~~0.3235 htc~~ (***tn***)	~~0.2770 htc~~ (***tn***)
8	0.2604 new (*fp*)	0.2385 browser (***tp***)	0.3221 samsung galaxy (*fp*)	**0.2632 organizer** (***tp***)
9	0.2579 page (***tp***)	**0.2382 calculator** (***tp***)	0.2645 timer (***tp***)	0.2606 hd (*fp*)
10	**0.2577 calendar** (***tp***)	0.2357 lg new chocolate (*fp*)	**0.2573 memo** (***tp***)	~~0.2359 mini~~ (***tn***)
11	~~0.2493 sony~~ (***tn***)	0.2357 lg new (*fp*)	0.2250 world (*fp*)	~~0.2205 omnia~~ (***tn***)
12	~~0.2451 ericsson~~ (***tn***)	**0.2156 note** (***tp***)	**0.2227 calendar** (***tp***)	**0.2176 memo** (***tp***)
13	~~0.2450 sony ericsson~~ (***tn***)	~~0.2139 blackberry~~ (***tn***)	0.2055 day (***tp***)	0.1976 cover (*fp*)
14	**0.2444 organizer** (***tp***)	~~0.2092 htc~~ (***tn***)	0.2016 hero (*fp*)	0.1970 motorola (*fp*)
15	~~0.2170 blackberry~~ (***tn***)	**0.2078 stopwatch** (***tp***)	**0.1984 stopwatch** (***tp***)	0.1949 support (***tp***)
16	0.2121 bl20 (*fp*)	**0.1985 memo** (***tp***)	0.1974 file manager (***tp***)	0.1948 timer (***tp***)
17	~~0.2096 browser~~ (*fn*)	0.1969 chocolate (*fp*)	0.1973 event (***tp***)	0.1933 world (*fp*)
18	0.2051 new ... (*fp*)	0.1946 event (***tp***)	**0.1963 world clock** (***tp***)	0.1929 wave (*fp*)
19	0.2051 chocolate bl40 (*fp*)	**0.1938 world clock** (***tp***)	**0.1947 note** (***tp***)	~~0.1846 x10~~ (***tn***)
20	0.2040 timer (***tp***)	0.1909 bl40 (*fp*)	**0.1916 converter** (***tp***)	0.1815 divx (*fp*)
21	~~0.1911 hour~~ (*fn*)	0.1892 countdown (***tp***)	0.1892 quickoffice (***tp***)	0.1806 hero (*fp*)
22	0.1862 elm (*fp*)	0.1825 day (***tp***)	0.1864 option (***tp***)	0.1801 viewty (*fp*)
23	0.1788 lg new ... (*fp*)	0.1790 anniversary (***tp***)	0.1863 anniversary (***tp***)	0.1765 file manager (***tp***)
24	0.1725 chocolate bl20 (*fp*)	0.1766 wifi (*fp*)	~~0.1857 hd~~ (***tn***)	~~0.1753 text~~ (*fn*)
25	0.1725 new chocol bl20 (*fp*)	**0.1758 converter** (***tp***)	0.1765 support (***tp***)	~~0.1738 option~~ (*fn*)
26	0.1710 rim (*fp*)	~~0.1757 hour~~ (*fn*)	0.1756 cover (*fp*)	0.1666 nexus (*fp*)
27	0.1668 sony ericsson elm (*fp*)	**0.1755 countdown timer** (***tp***	~~0.1718 mini~~ (***tn***)	0.1617 month (***tp***)
28	0.1668 ericsson elm (*fp*)	~~0.1753 rim~~ (***tn***)	**0.1695 countdown timer** (***tp***)	**0.1616 converter** (***tp***)
29	0.1638 lg new ... (*fp*)	0.1706 bold (***tp***)	0.1685 business card (***tp***)	~~0.1606 omnia hd~~ (***tn***)
30	0.1629 volume (***tp***)	0.1637 volume (***tp***)	0.1679 month (***tp***)	~~0.1579 xperia x10~~ (***tn***)

Table 5-8. Judgement applied over the top 30 terms of *dimensions 34* (with and without) and *35* (with and without) after discarding method.

As a result, precision, recall and accuracy graphs as single graphics are included in **Annex G: Precision, recall and accuracy graphs of LSI**. However, here it is found the **Graphic 5-4** of precision and recall which show each performance of each dimension with and without *extra reviews*. It is possible to appreciate how *dimension 4* overcomes the rest of dimension's performances, keeping precision over 80% until the 30th, having better precision with *extra reviews* at the end. Moreover, this dimension has the same accuracy and precision graphs. *Dimension 8* performs also very good, keeping its precision over 80% during the first 30th terms without *extra reviews*, but only until term 21st with them included. *Dimension 34* and *dimension 35* have very different performance from the named before, except *dimension 34* with *extra reviews*, whose precision remains over 80% until the 7th term. The rest are poorer performances regarding that the first ones are the most representatives. Recall graphs are very similar to each other, with the exception of *dimension 35* without *extra reviews*, which increases a bit slower.

To finalize the process, dimensions such as *dimension 4* and *dimension 8* are selected in both performances, with and without *extra reviews*, and *dimension 34* only in performances with them. From these dimensions it is extracted the new semantically related terms with the initial product features relate to the *organizer* product feature in the context of smart-phones. From ***dimension 4*** in **both** performances, terms such as *contact*, *menu*, *clock*, *application*,

email, *week*, *list*, *message*, *option*, *field*, *number*, *file*, *day*, *name*, *month*, *phone book*, *date*, *appointment*, *search*, *page* and *home screen* are extracted with a precision and accuracy of 96.55%, and a recall of 100% **without** *extra reviews*, but a precision and accuracy of 96.67% and a recall also of 100% **with** them. From ***dimension 8***, first in performance **with** *extra reviews,* terms such as *week*, *day*, *menu*, *field*, *clock*, *number*, *date*, *month* and *appointment*, are extracted with precision of 82.35%, accuracy of 80.95% and recall of 66.67%. Then, **without** *extra reviews* are extracted the same terms as with them, but also terms such as *duration*, *reminder*, *contact*, *priority*, *name*, *image* and *color*, are extracted with precision of 91.67%, accuracy over 89.66% and recall of 95.65%. Finally, from ***dimension 34*** **with** *extra reviews*, only terms such as *voice*, *timer*, *page* and *browser* are extracted with precision and accuracy of 88.89%, but recall only of 40%.



Graphic 5-4. Precision and recall graph of selected dimensions of *organizer* on sections.

5.2.2.3. Paragraphs

To finalize the analysis over the *organizer* product feature on LSI performances, it is check its performance on paragraphs, taking the collection of reviews as a collection of paragraphs to be the input of LSI. The most important dimensions extracted from the U_k matrix are located in **Table 9-6**. In this table it is seen how dimensions where the initial terms relative to the *organizer* product feature are scored better are: *dimension 2* in performances without *extra reviews* (*k=50,100*) and with them (*k=50*), having all the initial terms positive scored; *dimension 9* in performances with and without *extra reviews* (*k=150*), having also all positive scored; *dimension 17* in all performances with and without *extra reviews*, except with *k=150* without them, having also all positive scored; and finally, *dimension 37* in all performances

without *extra reviews*, but only with *k=100* when they are included, having 3 initial terms negative scored of the list of 15.

Place	*Dimension 2* without *k=50,100*	*Dimension 2* with *k=50*	*Dimension 9* without *k=150*	*Dimension 9* with *k=150*
1	1.1088 contact (***tp***)	1.1045 contact (***tp***)	0.7830 day (***tp***)	0.7613 day (***tp***)
2	0.7872 number (***tp***)	0.7856 number (***tp***)	**0.7422 alarm** (***tp***)	**0.7443 alarm** (***tp***)
3	0.6855 email (***tp***)	0.6824 email (***tp***)	**0.7283 calendar** (***tp***)	**0.7267 calendar** (***tp***)
4	0.6514 message (***tp***)	0.6487 message (***tp***)	0.5932 week (***tp***)	0.5825 week (***tp***)
5	0.5805 name (***tp***)	0.5774 name (***tp***)	**0.5353 task** (***tp***)	**0.5399 task** (***tp***)
6	**0.5376 calendar** (***tp***)	**0.5404 calendar** (***tp***)	0.4643 month (***tp***)	0.4560 month (***tp***)
7	0.5107 phone book (***tp***)	**0.5124 alarm** (***tp***)	0.4375 clock (***tp***)	**0.4428 note** (***tp***)
8	**0.5083 alarm** (***tp***)	0.5090 phone book (***tp***)	**0.4337 note** (***tp***)	0.4424 clock (***tp***)
9	0.5065 menu (***tp***)	0.5047 menu (***tp***)	0.4158 menu (***tp***)	**0.4200 calculator** (***tp***)
10	**0.4843 note** (***tp***)	**0.4870 note** (***tp***)	**0.4056 calculator** (***tp***)	0.3947 menu (***tp***)
11	0.4746 list (***tp***)	0.4723 list (***tp***)	0.3970 appointment (***tp***)	0.3895 appointment (***tp***)
12	0.4478 field (***tp***)	0.4460 field (***tp***)	0.3711 option (***tp***)	0.3805 camera (*fp*)
13	0.4407 text (***tp***)	0.4396 text (***tp***)	~~0.3680 camera~~ (***tn***)	0.3626 option (***tp***)
14	0.3700 address (*fp*)	0.3685 address (*fp*)	0.3457 event (***tp***)	0.3380 event (***tp***)
15	~~0.3460 application~~ (*fn*)	0.3437 application (***tp***)	**0.3247 organizer** (***tp***)	**0.3327 organizer** (***tp***)
16	0.3343 search (***tp***)	0.3332 search (***tp***)	0.3221 reminder (***tp***)	0.3189 reminder (***tp***)
17	0.3067 date (***tp***)	0.3065 date (***tp***)	0.3145 date (***tp***)	0.3071 date (***tp***)
18	0.2997 clock (***tp***)	0.3037 clock (***tp***)	~~0.3068 time~~ (*fn*)	~~0.2987 card~~ (***tn***)
19	0.2975 information (***tp***)	**0.2981 task** (***tp***)	~~0.2978 mode~~ (***tn***)	0.2966 timer (***tp***)
20	**0.2943 task** (***tp***)	0.2948 information (***tp***)	0.2839 timer (***tp***)	~~0.2947 mode~~ (***tn***)
21	0.2794 week (***tp***)	0.2804 week (***tp***)	~~0.2729 card~~ (***tn***)	~~0.2884 time~~ (*fn*)
22	0.2741 dial (*fp*)	0.2747 dial (*fp*)	~~0.2690 soni ericsson~~ (***tn***)	~~0.2805 soni ericsson~~ (***tn***)
23	~~0.2718 option~~ (*fn*)	**0.2723 calculator** (***tp***)	~~0.2678 ericsson~~ (***tn***)	~~0.2794 ericsson~~ (***tn***)
24	**0.2677 calculator** (***tp***)	0.2662 account (*fp*)	~~0.2662 sony~~ (***tn***)	~~0.2777 sony~~ (***tn***)
25	0.2675 account (*fp*)	0.2650 option (***tp***)	**0.2517 world clock** (***tp***)	**0.2689 converter** (***tp***)
26	0.2639 sms (*fp*)	**0.2642 organizer** (***tp***)	~~0.2510 hour~~ (*fn*)	**0.2630 world clock** (***tp***)
27	0.2623 type (***tp***)	0.2632 sms (*fp*)	**0.2448 converter** (***tp***)	~~0.2489 slot~~ (***tn***)
28	**0.2609 organizer** (***tp***)	0.2615 type (***tp***)	~~0.2400 view~~ (*fn*)	~~0.2436 flash~~ (***tn***)
29	0.2595 letter (***tp***)	0.2588 letter (***tp***)	~~0.2394 flash~~ (***tn***)	**0.2425 memo** (***tp***)
30	0.2582 entry (***tp***)	0.2576 entry (***tp***)	~~0.2382 battery~~ (***tn***)	~~0.2358 view~~ (*fn*)

Table 5-9. Judgement applied over the top 30 terms of *dimensions 2* (with and without) and *9* (with and without) after discarding method.

Running the discarding method gives the resultant graphics located in **Annex D: Discarding method applied to LSI dimensions**, in section **12.2.2**, where it is seen that in *dimension 2* without *extra reviews* there are only 2 terms discarded, meanwhile when they are included any term is discarded. The most particular case occurs in *dimension 37*, where it is found that in both performances, with and without *extra reviews*, there are many terms which are better scored in other dimensions. Due to this fact, it could be assumed that *dimension 37* may not be the *organizer dimension* what it is looked for, but it is also analyzed to check the differences between the others. In general, any initial term is discarded by the discarding method, which shows that although some discarded terms are wrong discarded, at least it has not dropped the initial ones.

After the application of the discarding method, it is applied the judgement as a human being, deciding which term is correctly retrieved and discarded considering the semantic relation with the initial product features related with the *organizer* product feature. In tables **Table 5-9** and **Table 5-10** are shown the judged top 30 terms of the selected dimensions considered to be the *organizer dimensions*.

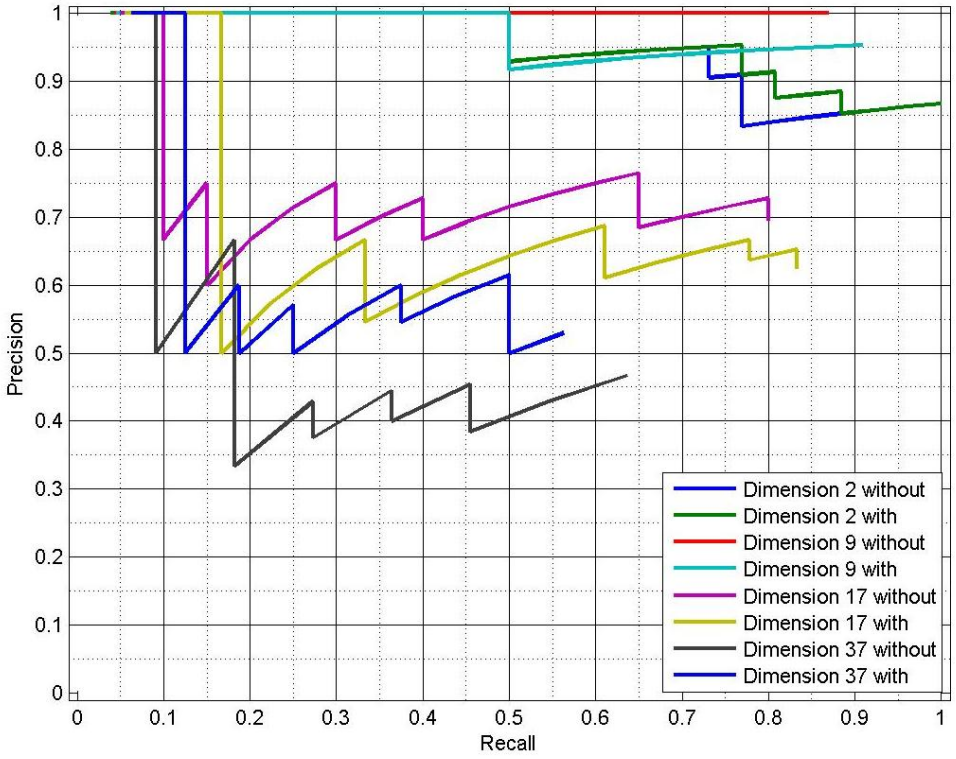
Place	*Dimension 17* without *k=50,100,200*	*Dimension 17* with *k=50,100,150,200*	*Dimension 37* without *k=50,100,150,200*	*Dimension 37* with *k=100*
1	0.7416 voice (***tp***)	0.7348 voice (***tp***)	**0.5605 alarm** (***tp***)	0.5454 clock (***tp***)
2	0.4414 application (***tp***)	0.4244 application (***tp***)	0.5159 radio (*fp*)	**0.5331 alarm** (***tp***)
3	0.4109 nokia (*fp*)	0.3800 sound (***tp***)	0.4767 clock (***tp***)	0.4609 htc (*fp*)
4	0.3775 sound (***tp***)	0.3713 nokia (*fp*)	0.4100 fm (*fp*)	~~0.3891 voice~~ (*fn*)
5	0.3247 quality (*fp*)	0.3303 quality (*fp*)	0.3432 fm radio (*fp*)	0.3805 world (*fp*)
6	0.3218 feature (***tp***)	0.3228 k750i (*fp*)	~~0.3226 htc~~ (***tn***)	**0.3589 calculator** (***tp***)
7	0.3071 software (***tp***)	~~0.3196 feature~~ (***tn***)	0.3010 blackberry (*fp*)	0.3378 blackberry (*fp*)
8	0.3047 word (***tp***)	0.3114 word (***tp***)	**0.2883 calculator** (***tp***)	**0.3248 world clock** (***tp***)
9	~~0.3028 game~~ (***tn***)	**0.3060 office** (***tp***)	~~0.2575 motorola~~ (***tn***)	0.2979 radio (*fp*)
10	0.2925 car (*fp*)	~~0.3033 game~~ (***tn***)	0.2481 world (*fp*)	~~0.2755 motorola~~ (***tn***)
11	**0.2920 office** (***tp***)	~~0.3022 handset~~ (***tn***)	**0.2441 world clock** (***tp***)	0.2645 recorder (***tp***)
12	0.2899 excel (***tp***)	0.3008 document (***tp***)	0.2435 station (*fp*)	~~0.2618 home~~ (***tn***)
13	0.2885 k750i (*fp*)	0.2953 car (*fp*)	~~0.2350 video~~ (***tn***)	~~0.2530 flash~~ (***tn***)
14	0.2866 document (***tp***)	0.2943 iphone (*fp*)	0.2331 recorder (***tp***)	**0.2406 memo** (***tp***)
15	~~0.2866 handset~~ (***tn***)	0.2938 excel (***tp***)	~~0.2244 lg~~ (***tn***)	0.2373 fm (*fp*)
16	**0.2757 note** (***tp***)	~~0.2913 software~~ (*fn*)	~~0.2206 battery~~ (***tn***)	~~0.2341 screen~~ (***tn***)
17	**0.2633 alarm** (***tp***)	**0.2695 note** (***tp***)	~~0.2106 mobile~~ (***tn***)	~~0.2335 home screen~~ (*fn*)
18	0.2610 program (***tp***)	**0.2559 calculator** (***tp***)	0.1923 rim (*fp*)	**0.2325 converter** (***tp***)
19	**0.2567 calculator** (***tp***)	0.2556 program (***tp***)	~~0.1903 flash~~ (***tn***)	~~0.2208 function~~ (*fn*)
20	~~0.2543 iphone~~ (***tn***)	**0.2550 alarm** (***tp***)	~~0.1815 feature~~ (***tn***)	0.2189 widget (***tp***)
21	0.2323 kit (*fp*)	0.2420 kit (*fp*)	0.1796 addition (*fp*)	0.2187 rim (*fp*)
22	~~0.2308 system~~ (*fn*)	0.2216 party (*fp*)	**0.1781 converter** (***tp***)	~~0.2125 weight~~ (***tn***)
23	0.2251 party (*fp*)	0.2212 system (***tp***)	~~0.1767 user~~ (*fn*)	0.2070 battery (*fp*)
24	~~0.2097 function~~ (*fn*)	**0.2209 pdf** (***tp***)	**0.1755 memo** (***tp***)	~~0.2005 touch~~ (*fn*)
25	**0.2088 memo** (***tp***)	~~0.2124 memory~~ (*fn*)	~~0.1742 time~~ (*fn*)	0.1880 fm radio (*fp*)
26	**0.2085 pdf** (***tp***)	**0.2122 memo** (***tp***)	~~0.1718 razr~~ (***tn***)	~~0.1820 command~~ (***tn***)
27	0.2067 powerpoint (***tp***)	0.2120 w800i (*fp*)	~~0.1662 application~~ (*fn*)	~~0.1795 time~~ (*fn*)
28	~~0.2035 memory~~ (*fn*)	0.2096 powerpoint (***tp***)	~~0.1658 at&t~~ (***tn***)	~~0.1701 service~~ (*fn*)
29	~~0.1952 recorder~~ (*fn*)	~~0.2067 function~~ (*fn*)	~~0.1609 tone~~ (***tn***)	~~0.1697 application~~ (*fn*)
30	0.1916 w800i (*fp*)	0.1956 apple (*fp*)	~~0.1592 function~~ (***fn***)	**0.1688 note** (***tp***)

Table 5-10. Judgement applied over the top 30 terms of *dimensions 17* (with and without) and *37* (with and without) after discarding method.

Measurements such as accuracy, recall and precision are calculated to view graphically how these dimensions perform, whose graphs are located in **Annex G: Precision, recall and accuracy graphs of LSI**. Here is presented the precision and recall graph, **Graphic 5-5**, which shows the contrasted behaviour of selected dimensions in these first 30 terms. It is possible to see how *dimension 9* in both performances, with and without *extra reviews*, performs much better than the rest of dimensions, keeping the precision, over 90% and of 100%, respectively during the whole sample. *Dimension 2* also achieves good results too, maintaining over 80% in the entire top 30 terms with and without *extra reviews.*The rest of performances of *dimension 17* and *dimension 37*, are poorer regarding these three measures, as it was supposed.

To finalize the process, it is selected dimensions such as *dimension 2* and *dimension 9* to extract semantic related terms with the initial terms of the *organizer* product feature in the context of the smart-phones. From ***dimension 2***, terms such as *contact*, *number*, *email*, *message*, *name*, *phone book*, *menu*, *list*, *field*, *text*, *search*, *date*, *clock*, *information*, *week*, *type*, *letter* and *entry*, are extracted **without** *extra reviews*, with a precision of 85.71%, accuracy of 80% and a recall of 92.31%. **With** *extra reviews* it is found the same list, but 2 terms more, *option* and *application*, than with them, with a precision and accuracy of 86.67%, and a recall of 100%. From ***dimension 9*** in **both** performances, terms such as *day*, *week*, *month*, *clock*,

menu, *appointment*, *option*, *event*, *reminder*, *date* and *timer*, are extracted **without** *extra reviews* with a precision of 100%, accuracy of 90% and a recall of 86.96%, while **with** them with a precision of 95.24%, accuracy of 90% and a recall over 90.91%. From ***dimension 17***, *voice* and *application* are extracted **without** *extra reviews*, with precision and accuracy of 100%, but a recall of 10%. **With** *extra reviews*, the same list of terms is extracted, but also the term *sound*, with a precision and accuracy of 100%, but recall of 11.11%. Finally, from ***dimension 37*** in performance **with** *extra reviews*, term *clock* is extracted with 100% of precision and accuracy, but only a recall of 6.25%.



Graphic 5-5. Precision and recall graph of selected dimensions of *organizer* on paragraphs.

5.2.3. Multimedia

5.2.3.1. Documents

In this last part of the LSI analysis, it is analyzed the behaviour of LSI with the *multimedia* product feature, taking the same steps described in **Table 4-1** to achieve the goal of extracting the major number of well semantically related terms with the initial terms from the *multimedia* product feature. At first, it is run on documents, considering each review as an entity to input of LSI.

Looking inside **Table 9-7**, one of the best scored dimensions is *dimension 1* in both performances, with and without *extra reviews*, like it occurs with *battery* and *organizer* on documents. In *dimension 1*, although all the initial terms are positive scored, there is any initial term in the top 30 terms when the dimension scores are sorted, the first places of the initial terms are the 32nd and the 76th, and this means that there is a high probability that this dimension is not the *multimedia dimension* it is looked for. The same occurs with dimensions such as *dimension 2*, *dimension 20, dimension 65* and *dimension 93*, which have the initial terms scored much better than the rest, but not good enough to be consider firstly as good as to be analyzed. Moreover, these dimensions have 17, 3, 7 and 11 initial terms negative scored, respectively, from the list of 49 initial terms. Therefore, it can be said that, like it occurs with *battery* and *organizer*, it exists any dimension that defines the *multimedia* product feature in performances of LSI on documents, at any proposed value of the *k* parameter.

5.2.3.2. Sections

After LSI is run in documents, it is run on sections, getting as a result the dimensions marked in **Table 9-8**. These dimensions are selected from all the performances of LSI with the selected values of the *k* parameter. Although there are many dimensions where initial terms are good scored in relation to the places they occupy, it is necessary to select the best ones. The problem is that only one of the performances over all dimensions has more than 4 initial terms, of the list of 49, within the top 30 terms, *dimension 91*, with *extra reviews*, having only 3 terms negative scored. Without them, it only reaches 4 initial terms in the top 30, but up to 7 negative scored values. The rest of selected dimensions reach the same number of terms within the top 30, such as *dimension 5* which does not have any negative scored value in both performances, meanwhile *dimension 6* has 3 and *dimension 11* has 4, without *extra reviews*¸ but they decrease to 2 negative terms with them. Moreover, it has been considered the next initial terms in the sorted dimensions, viewing that there are many initial terms which are not included in the top 30, but they are in the top 40 and top 50.

Following the steps introduced in **Table 4-1**, first of all, it is presented the selected dimensions such as *dimension 5*, without *extra reviews* with *k=50,100,200* and with them all observations of the *k* parameter; *dimension 6*, without *extra reviews* with *k=50,150,200*, but only with *k=50,200* with them; *dimension 11*, without *extra reviews* with *k=50,100,150*, and also only with *k=50,100* with them; and finally, *dimension 91*, with all the possible

performances without *extra reviews*, but only with *k=100* with them. After selecting dimensions, discarding method is applied over the performances which included more dimensions of each dimension to evaluate, e.g. in *dimension 5* without *extra reviews* it is taken the performance with *k=200*. Results of this application of the discarding method are shown in **Annex D: Discarding method applied to LSI dimensions** in the **12.3.1** section. There, it is seen that in *dimension 5* in both performances, it is supposed that the entire top 30 terms are good retrieved, because the discarding method has discarded any term, but later it is seen that it is not correctly done. In *dimension 6*, only 2 terms with *extra reviews* and 3 without them are discarded, meanwhile in *dimension 11* occurs the inverse, 4 terms are discarded with *extra reviews*, but only 3 without them. The major number of extractions happens in *dimension 91*, where it also discards some initial terms such as *fm* and *radio* without *extra reviews*, selecting up to 18 terms of the top 30 terms. With them it improves a little, but it still discards *radio* and the number of discarding terms goes up to 13.

Place	*Dimension 5* without *k=50,100,200*	*Dimension 5* with *k=50,100,150,200*	*Dimension 6* without *k=50,150,200*	*Dimension 6* with *k=50,200*
1	**0.7405 video** (***tp***)	**0.7433 video** (***tp***)	0.5263 music (***tp***)	0.5287 music (***tp***)
2	0.6020 player (***tp***)	0.6079 player (***tp***)	0.4068 player (***tp***)	0.4098 player (***tp***)
3	0.5130 sample (***tp***)	0.5110 sample (***tp***)	0.3686 sound (***tp***)	0.3686 sound (***tp***)
4	0.5032 music (***tp***)	0.5090 music (***tp***)	**0.3446 music player** (***tp***)	**0.3466 music player** (***tp***)
5	0.4340 quality (***tp***)	0.4330 quality (***tp***)	0.3288 noise (***tp***)	0.3259 noise (***tp***)
6	0.4331 interface (***tp***)	0.4305 interface (***tp***)	0.3103 album (***tp***)	0.3126 album (***tp***)
7	0.4081 picture (***tp***)	0.4062 picture (***tp***)	0.2978 second (***tp***)	0.2958 song (***tp***)
8	0.3952 playback (***tp***)	0.4006 playback (***tp***)	0.2956 song (***tp***)	0.2941 second (***tp***)
9	0.3780 album (***tp***)	0.3824 album (***tp***)	0.2893 light (***tp***)	**0.2903 radio** (***tp***)
10	0.3611 lightlow (*fp*)	0.3595 lightlow (*fp*)	**0.2879 radio** (***tp***)	0.2868 light (***tp***)
11	0.3457 lightmedium (*fp*)	0.3442 lightmedium (*fp*)	0.2797 image (***tp***)	0.2782 headphone (***tp***)
12	0.3391 lightmedium lightlow (*fp*)	0.3376 lightmedium lightlow (*fp*)	0.2769 headphone (***tp***)	0.2780 speaker (***tp***)
13	0.3265 flash (***tp***)	0.3265 flash (***tp***)	0.2767 speaker (***tp***)	0.2758 image (***tp***)
14	0.3227 track (***tp***)	0.3241 track (***tp***)	~~0.2740 picture~~ (*fn*)	0.2717 kit (*fp*)
15	0.3197 camera (***tp***)	0.3162 camera (***tp***)	0.2710 kit (*fp*)	~~0.2683 picture~~ (*fn*)
16	0.3127 snapshot (***tp***)	0.3111 snapshot (***tp***)	0.2672 track (***tp***)	0.2667 track (***tp***)
17	0.3052 file (***tp***)	**0.3068 radio** (***tp***)	**0.2656 video** (***tp***)	**0.2651 video** (***tp***)
18	**0.3024 radio** (***tp***)	0.3066 file (***tp***)	0.2630 file (***tp***)	0.2629 file (***tp***)
19	0.2976 k750i (*fp*)	0.2986 k750i (*fp*)	0.2607 n73 (*fp*)	0.2595 n73 (*fp*)
20	0.2943 mode (***tp***)	0.2933 mode (***tp***)	0.2569 playlist (***tp***)	0.2573 playlist (***tp***)
21	0.2865 resolution (***tp***)	**0.2860 audio** (***tp***)	0.2501 equalizer (***tp***)	0.2499 equalizer (***tp***)
22	**0.2826 audio** (***tp***)	0.2848 resolution (***tp***)	0.2441 artist (***tp***)	0.2441 artist (***tp***)
23	0.2800 camera interface (***tp***)	0.2787 camera interface (***tp***)	0.2383 car (*fp*)	0.2390 car (*fp*)
24	0.2749 photo (***tp***)	0.2739 artist (***tp***)	0.2381 the (*fp*)	0.2383 the (*fp*)
25	0.2726 detail (***tp***)	0.2736 photo (***tp***)	0.2375 n80 (*fp*)	0.2364 n80 (*fp*)
26	0.2726 artist (***tp***)	0.2711 detail (***tp***)	~~0.2288 mode~~ (*fn*)	~~0.2251 mode~~ (*fn*)
27	0.2700 sample video (***tp***)	0.2687 sample video (***tp***)	0.2232 format (***tp***)	0.2238 format (***tp***)
28	0.2692 shot (***tp***)	0.2681 media (***tp***)	**0.2143 fm** (***tp***)	**0.2186 fm** (***tp***)
29	0.2667 media (***tp***)	0.2680 shot (***tp***)	0.2124 photo (***tp***)	0.2126 memory (***tp***)
30	**0.2637 multimedia** (***tp***)	**0.2664 multimedia** (***tp***)	~~0.2122 memory~~ (*fn*)	0.2088 station (***tp***)

Table 5-11. Judgement applied over the top 30 terms of *dimensions 5* (with and without) and *6* (with and without) after discarding method.

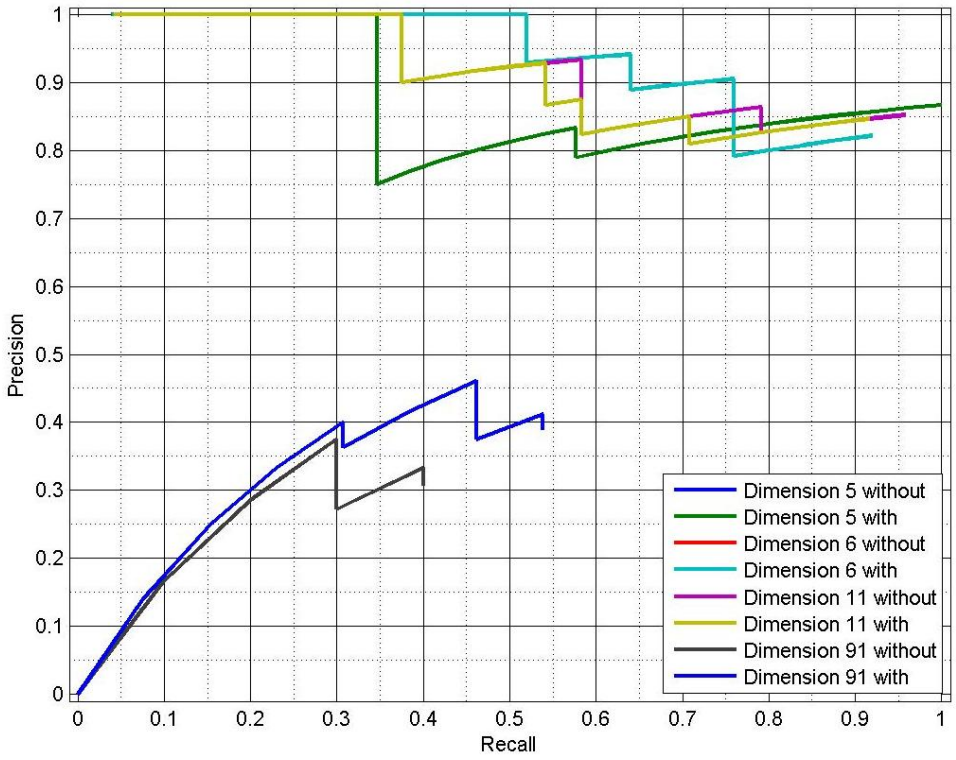
Once the discarding method is applied, the judgement is set over the top 30 terms of the selected dimensions. In general, their performance could be better, because there are terms that are need to be discarded and others which have been wrong discarded. In **Table 5-11** and **Table 5-12** it is presented the entire selected dimensions with the judgement about

the semantic relations of the top 30 terms with the initial terms. The biggest problem is the high appearance of terms related with manufacturers and smart-phone models, which decreases the precision of the sample.

Place	*Dimension 11* without *k=50,100,150*	*Dimension 11* with *k=50,100*	*Dimension 91* without *k=100,150,200*	*Dimension 91* with *k=100*
1	0.5212 music (***tp***)	0.5130 music (***tp***)	0.2721 pixon (*fp*)	0.3272 pixon (*fp*)
2	0.5101 player (***tp***)	0.5055 player (***tp***)	0.2549 aino (*fp*)	0.2558 aino (*fp*)
3	0.4292 album (***tp***)	0.4282 album (***tp***)	~~0.2182 acer~~ (***tn***)	0.2553 renoir (*fp*)
4	0.3893 song (***tp***)	0.3857 song (***tp***)	0.2140 ericsson aino (*fp*)	0.2154 ericsson aino (*fp*)
5	**0.3482 music player** (***tp***)	**0.3456 music player** (***tp***)	0.2120 renoir (*fp*)	0.2126 sony ericsson aino (*fp*)
6	0.3332 artist (***tp***)	0.3300 artist (***tp***)	0.2111 sony ericsson aino (*fp*)	0.2064 acer (*fp*)
7	**0.3137 radio** (***tp***)	0.3107 menu (***tp***)	**0.1924 fm** (***tp***)	**0.1996 divx** (***tp***)
8	0.3065 track (***tp***)	**0.3101 radio** (***tp***)	~~0.1848 omnia~~ (***tn***)	**0.1914 xvid** (***tp***)
9	0.3056 menu (***tp***)	0.3045 track (***tp***)	**0.1815 divx** (***tp***)	**0.1913 fm** (***tp***)
10	0.2939 button (*fp*)	0.2979 button (*fp*)	~~0.1754 wave~~ (***tn***)	~~0.1767 wave~~ (***tn***)
11	0.2788 icon (***tp***)	0.2868 icon (***tp***)	~~0.1724 map~~ (***tn***)	~~0.1757 omnia~~ (***tn***)
12	**0.2691 fm** (***tp***)	**0.2685 fm** (***tp***)	**0.1718 xvid** (***tp***)	**0.1678 radio** (***tp***)
13	0.2634 headphone (***tp***)	0.2590 headphone (***tp***)	~~0.1632 menu~~ (*fn*)	0.1656 c905 (*fp*)
14	0.2469 playlist (***tp***)	0.2450 playlist (***tp***)	0.1588 liquid (*fp*)	0.1652 codec (***tp***)
15	0.2390 genre (***tp***)	0.2420 screen (*fp*)	~~0.1578 samsung wave~~ (***tn***)	~~0.1631 samsung wave~~ (***tn***)
16	0.2375 screen (*fp*)	0.2365 genre (***tp***)	0.1548 acer liquid (*fp*)	~~0.1608 file~~ (*fn*)
17	~~0.2325 iphone~~ (***tn***)	0.2317 k750i (*fp*)	0.1540 c905 (*fp*)	**0.1548 video player** (***tp***)
18	0.2316 k750i (*fp*)	0.2303 shortcut (***tp***)	~~0.1535 ii~~ (***tn***)	~~0.1506 ii~~ (***tn***)
19	0.2312 file (***tp***)	~~0.2281 iphone~~ (***tn***)	**0.1522 radio** (***tp***)	~~0.1475 slider~~ (***tn***)
20	0.2238 shortcut (***tp***)	0.2197 file (***tp***)	~~0.1451 gps~~ (***tn***)	~~0.1463 menu~~ (*fn*)
21	0.2075 home screen (***tp***)	0.2124 home screen (***tp***)	~~0.1444 file~~ (*fn*)	0.1409 liquid (*fp*)
22	0.2049 station (***tp***)	0.2053 home (*fp*)	~~0.1410 fm radio~~ (*fn*)	0.1373 acer liquid (*fp*)
23	~~0.2024 touch~~ (*fn*)	~~0.2022 touch~~ (*fn*)	~~0.1394 setting~~ (*fn*)	~~0.1371 map~~ (***tn***)
24	0.2000 fm radio (***tp***)	0.2014 station (***tp***)	~~0.1394 liquid a1~~ (***tn***)	~~0.1344 setting~~ (*fn*)
25	0.1994 home (*fp*)	0.1953 walkman (***tp***)	~~0.1394 a1~~ (***tn***)	~~0.1313 fm radio~~ (*fn*)
26	0.1941 walkman (***tp***)	0.1949 fm radio (***tp***)	~~0.1387 hd~~ (*fn*)	~~0.1308 hd~~ (*fn*)
27	**0.1929 multimedia** (***tp***)	**0.1923 multimedia** (***tp***)	~~0.1373 omnia hd~~ (***tn***)	0.1283 samsung pixon (*fp*)
28	~~0.1927 blackberry~~ (***tn***)	~~0.1919 blackberry~~ (***tn***)	0.1361 codec (***tp***)	~~0.1255 flash~~ (*fn*)
29	0.1869 display (***tp***)	0.1902 display (***tp***)	0.1346 acer liquid a1 (*fp*)	0.1255 codec support (***tp***)
30	0.1865 play (***tp***)	~~0.1886 theme~~ (*fn*)	~~0.1313 panel~~ (*fn*)	0.1243 a1 (*fp*)

Table 5-12. Judgement applied over the top 30 terms of *dimensions 11* (with and without) and *91* (with and without) after discarding method.

After applying the human review over the top 30 terms, precision, accuracy and recall are calculated numerically and showed graphically in **Annex G: Precision, recall and accuracy graphs of LSI**. Here is added the precision and recall graph **Graphic 5-6**, which show how performs each dimension regarding the *multimedia* product feature on sections scenario. As it was expected, *dimension 91* has the poorest performance of all, because the first terms, which are supposed to be related with the *multimedia* context, mainly are terms related with manufacturers and smart-phone models. Two particular cases occur with *dimension 5* and *dimension 6*, which have almost the same performance with and without *extra reviews*, fact which reveals that there is any improvement in including reviews containing the initial product features all together. One of the most highlights is *dimension 6*, which maintains a precision over 80% until the 24th term, meanwhile *dimension 5* only keeps it until the 11th, but *dimension 11* is the one which remains its precision higher than 80% in the entire top 30 terms with and without *extra reviews*.



Graphic 5-6 Precision and recall graph of selected dimensions of *multimedia* on sections.

As a result, it is considered the extraction of semantic relations of new terms related with the initial terms of the *multimedia* product feature in the context of smart-phones, from all the dimensions analyzed, except from *dimension 91*. Then, from ***dimension 5*** in **both** performances, terms such as *player*, *sample*, *music*, *picture*, *interface*, *quality*, *playback* and *album*, are extracted with a precision of 100%, accuracy also of 100%, but a recall of 34.62%. From ***dimension 6*** in **both** performances too, terms such as *music*, *player*, *sound*, *noise*, *album*, *second*, *song*, *light*, *image*, *headphone*, *speaker*, *track*, *file*, *playlist*, *equalizer* and *artist*, are extracted with a precision 90.48%, accuracy of 86.36% and recall of 76%. Finally, from ***dimension 11*** **without** *extra reviews*, terms such as *music*, *player*, *album*, *song*, *artist*, *menu*, *track*, *icon*, *headphone*, *playlist*, *genre*, *file*, *shortcut*, *home screen*, *station*, *fm radio*, *walkman*, *display* and *play*, are extracted with a precision of 85.19%, accuracy of 83.33% and recall of 95.83%. **With** them, the same terms less ~~*play*~~ are extracted, with a precision of 84.62%, accuracy of 82.76% and recall of 91.67%.

5.2.3.3. Paragraphs

To complete the analysis presented before, the last scenario where LSI is proved is on paragraphs, where reviews are divided in paragraphs and each paragraphs is treated as an entity to be the input of LSI. Starting from the initial terms found in the technical specifications of the *multimedia* product feature in the context of smart-phone reviews. The result of running LSI on paragraphs considering these initial terms is shown in **Table 9-9**. This table contains the dimensions where initial terms are best scored, therefore, from them it is

extracted the best dimensions whose first terms can define the *multimedia* product feature, due to the number of good scored initial terms.

Place	*Dimension 3* without *k=100*	*Dimension 3* with *k=50*	*Dimension 5* without *k=50,100,200*	*Dimension 5* with *k=50,100*
1	**0.9544 video** (***tp***)	**0.9619 video** (***tp***)	0.8356 music (***tp***)	0.7795 music (***tp***)
2	0.5880 player (***tp***)	0.6167 player (***tp***)	0.7169 player (***tp***)	0.6744 player (***tp***)
3	0.5764 resolution (***tp***)	0.5650 resolution (***tp***)	0.6486 sound (***tp***)	0.6067 sound (***tp***)
4	0.5223 interface (***tp***)	0.5140 file (***tp***)	0.5305 picture (***tp***)	0.5648 picture (***tp***)
5	0.5017 file (***tp***)	0.5140 interface (***tp***)	0.4816 quality (***tp***)	**0.4671 video** (***tp***)
6	0.4877 picture (***tp***)	0.4805 picture (***tp***)	**0.4779 video** (***tp***)	0.4550 quality (***tp***)
7	0.4284 image (***tp***)	0.4237 option (***tp***)	0.4450 album (***tp***)	0.4297 album (***tp***)
8	0.4152 option (***tp***)	0.4191 image (***tp***)	**0.4424 music player** (***tp***)	**0.4240 music player** (***tp***)
9	0.3856 music (***tp***)	0.4138 music (***tp***)	0.3806 song (***tp***)	~~0.3765 image~~ (*fn*)
10	~~0.3466 screen~~ (***tn***)	0.3431 album (***tp***)	~~0.3608 image~~ (*fn*)	0.3608 song (***tp***)
11	0.3293 pixel (***tp***)	~~0.3358 screen~~ (***tn***)	0.3563 track (***tp***)	0.3430 mode (***tp***)
12	~~0.3269 samsung~~ (***tn***)	0.3219 pixel (***tp***)	~~0.3370 mode~~ (*fn*)	0.3349 track (***tp***)
13	0.3252 album (***tp***)	0.3135 samsung (*fp*)	~~0.3360 voice~~ (*fn*)	~~0.3275 setting~~ (*fn*)
14	0.3113 page (*fp*)	~~0.3086 page~~ (***tn***)	0.3330 setting (***tp***)	0.3185 noise (***tp***)
15	~~0.3011 application~~ (*fn*)	0.3045 application (***tp***)	0.3315 noise (***tp***)	~~0.3178 voice~~ (*fn*)
16	0.2969 browser (***tp***)	~~0.2944 browser~~ (*fn*)	0.3301 artist (***tp***)	0.3120 artist (***tp***)
17	0.2747 sample (***tp***)	0.2778 playback (***tp***)	~~0.2984 call~~ (***tn***)	~~0.3084 call~~ (***tn***)
18	0.2657 playback (***tp***)	0.2701 sample (***tp***)	~~0.2928 option~~ (*fn*)	~~0.2970 option~~ (*fn*)
19	**0.2607 multimedia** (***tp***)	**0.2678 multimedia** (***tp***)	0.2838 equalizer (***tp***)	0.2835 volume (***tp***)
20	~~0.2575 sony ericsson~~ (***tn***)	0.2535 media (***tp***)	~~0.2805 file~~ (*fn*)	~~0.2710 second~~ (*fn*)
21	~~0.2553 ericsson~~ (***tn***)	**0.2521 music player** (***tp***)	~~0.2705 volume~~ (*fn*)	0.2683 equalizer (***tp***)
22	~~0.2536 sony~~ (***tn***)	~~0.2521 internet~~ (*fn*)	~~0.2675 second~~ (*fn*)	~~0.2657 camera~~ (*fn*)
23	0.2512 internet (***tp***)	**0.2462 divx** (***tp***)	0.2667 playlist (***tp***)	0.2508 playlist (***tp***)
24	0.2453 media (***tp***)	~~0.2442 windows~~ (***tn***)	**0.2587 mp3** (***tp***)	**0.2501 mp3** (***tp***)
25	~~0.2441 windows~~ (***tn***)	0.2368 sony ericsson (*fp*)	~~0.2545 speaker~~ (*fn*)	~~0.2446 menu~~ (*fn*)
26	~~0.2429 lg~~ (***tn***)	~~0.2356 flash~~ (*fn*)	~~0.2416 headset~~ (*fn*)	~~0.2421 speaker~~ (*fn*)
27	~~0.2411 flash~~ (*fn*)	0.2344 ericsson (*fp*)	0.2410 format (***tp***)	~~0.2421 resolution~~ (*fn*)
28	**0.2398 divx** (***tp***)	~~0.2338 lg~~ (***tn***)	~~0.2307 resolution~~ (*fn*)	~~0.2398 file~~ (*fn*)
29	**0.2355 music player** (***tp***)	0.2328 sony (*fp*)	**0.2291 radio** (***tp***)	~~0.2359 light~~ (*fn*)
30	~~0.2239 version~~ (*fn*)	0.2291 song (***tp***)	0.2273 genre (***tp***)	0.2282 format (***tp***)

Table 5-13. Judgement applied over the top 30 terms of *dimensions 3* (with and without) and *5* (with and without) after discarding method.

The first selected dimensions are *dimension 3* and *dimension 5*, because they have up to 4 initial terms within the top 30 in one of the two performances, with and without *extra reviews*, and also they have only 1 initial term negative scored of the list of 49 initial terms. *Dimension 15* in both performances is selected because of having at least 2 terms close to the top of the dimension, which is the initial term *video*. The next initial terms are not too far, but it has up to 10 initial terms negative scored. The last ones are *dimension 39* without *extra reviews*, although it has 12 negative initial terms, it has 4 initial terms in the top 30; and *dimension 13* with *extra reviews*, because it has a reduced number of negative initial terms, only 5, and 3 initial terms within the top 30. There are other potential dimensions susceptible to be selected, such as *dimension 38* with *radio* and *fm* in the 1st and 4th places, but the next initial terms are over the 60th place, or *dimension 26* and *dimension 51*, which both have 3 initial terms in the top 30, but they have more than 20 negative initial terms, but all of them are worse than the selected ones. The selected dimensions are *dimension 3,* without *extra reviews* (*k=100)* and with them (*k=50*), *dimension 5*, without *extra reviews* (*k=50,100,200*) and with them (*k=50,200*), *dimension 15*, without *extra reviews* (*k=100,200*) and with them

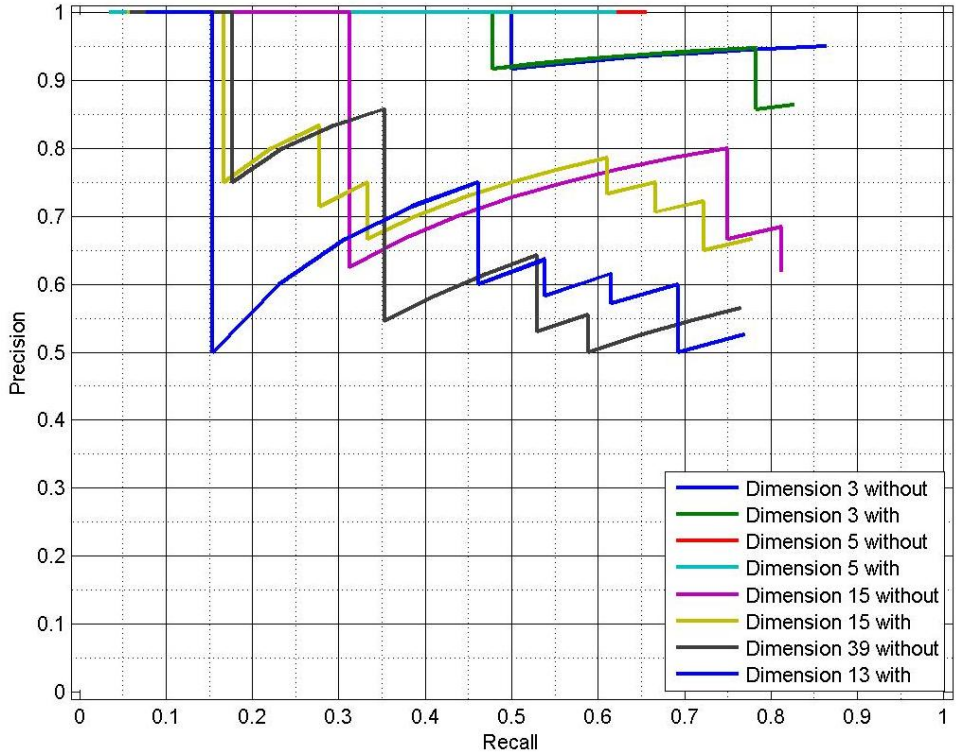
(*k=50,100,200*), *dimension 39* without *extra reviews* (*k=100,150*), and *dimension 13* without *extra reviews* (*k=100,150*).

Place	*Dimension 15* without *k=100,200*	*Dimension 15* with *k=50,100,200*	*Dimension 39* without *k=100,150*	*Dimension 13* with *k=100,150*
1	**0.8736 video** (***tp***)	**0.8543 video** (***tp***)	**0.6561 radio** (***tp***)	0.4986 phone (***tp***)
2	0.5114 headset (***tp***)	0.5251 headset (***tp***)	**0.5462 fm** (***tp***)	**0.4704 video** (***tp***)
3	0.4882 volume (***tp***)	0.4929 volume (***tp***)	0.4620 fm radio (***tp***)	0.4697 samsung (*fp*)
4	0.4244 file (***tp***)	0.4071 screen (*fp*)	0.3592 blackberry (*fp*)	0.4546 card (*fp*)
5	0.4189 resolution (***tp***)	0.4069 file (***tp***)	0.3551 game (***tp***)	~~0.3288 design~~ (***tn***)
6	0.4107 screen (*fp*)	0.4012 resolution (***tp***)	0.3111 station (***tp***)	0.3282 file (***tp***)
7	0.3853 call (*fp*)	~~0.3929 voice~~ (*fn*)	~~0.2998 day~~ (***tn***)	~~0.3180 nokia~~ (***tn***)
8	~~0.3779 voice~~ (*fn*)	~~0.3899 call~~ (***tn***)	0.2928 headphone (***tp***)	0.3162 memory (***tp***)
9	0.3492 rocker (*fp*)	~~0.3526 button~~ (***tn***)	~~0.2668 icon~~ (*fn*)	~~0.3152 lg~~ (***tn***)
10	~~0.3484 button~~ (***tn***)	0.3497 rocker (*fp*)	0.2647 tmobile (*fp*)	0.3143 handset (***tp***)
11	0.3442 sample (***tp***)	0.3364 sample (***tp***)	0.2643 cookie (*fp*)	0.3008 pixel (***tp***)
12	0.3366 pixel (***tp***)	0.3211 page (*fp*)	~~0.2643 button~~ (***tn***)	0.2890 slot (*fp*)
13	~~0.3147 side~~ (***tn***)	0.3179 pixel (***tp***)	0.2534 business (*fp*)	0.2757 cable (*fp*)
14	~~0.3082 page~~ (***tn***)	~~0.3136 side~~ (***tn***)	0.2408 rim (*fp*)	~~0.2549 note~~ (***tn***)
15	0.2990 volume rocker (***tp***)	0.2991 volume rocker (***tp***)	0.2301 antenna (***tp***)	0.2438 microsd (***tp***)
16	0.2914 support (***tp***)	0.2899 support (***tp***)	**0.2292 rds** (***tp***)	~~0.2397 device~~ (*fn*)
17	**0.2898 divx** (***tp***)	**0.2851 divx** (***tp***)	0.2283 theme (***tp***)	0.2379 week (*fp*)
18	0.2556 alarm (***tp***)	~~0.2633 browser~~ (*fn*)	0.2223 at&t (*fp*)	~~0.2295 alarm~~ (*fn*)
19	~~0.2521 browser~~ (*fn*)	0.2540 alarm (***tp***)	~~0.2130 jack~~ (*fn*)	**0.2268 divx** (***tp***)
20	**0.2517 xvid** (***tp***)	0.2497 dial (*fp*)	0.2090 lg cookie (*fp*)	0.2198 sim (*fp*)
21	~~0.2497 samsung~~ (***tn***)	**0.2482 xvid** (***tp***)	0.1953 usa (*fp*)	~~0.2183 display~~ (***tp***)
22	0.2455 dial (*fp*)	~~0.2455 samsung~~ (***tn***)	0.1949 visualization (***tp***)	0.2162 serie (*fp*)
23	0.2445 x10 (*fp*)	0.2444 x10 (*fp*)	~~0.1916 size~~ (*fn*)	~~0.2129 motorola~~ (***tn***)
24	0.2394 xperia (*fp*)	~~0.2419 home screen~~ (*fn*)	0.1907 pearl (*fp*)	0.2114 charger (*fp*)
25	~~0.2392 bluetooth~~ (*fn*)	0.2415 bluetooth (***tp***)	0.1889 tmobile usa (*fp*)	0.2095 cover (*fp*)
26	~~0.2372 home screen~~ (*fn*)	0.2390 xperia (*fp*)	**0.1856 video** (***tp***)	~~0.2044 calendar~~ (***tn***)
27	0.2255 pixel resolution (***tp***)	~~0.2277 home~~ (***tn***)	~~0.1838 date~~ (***tn***)	~~0.2040 manufacturer~~ (***tn***)
28	0.2232 port (*fp*)	~~0.2229 application~~ (*fn*)	~~0.1813 headset~~ (*fn*)	~~0.2003 cell~~ (***tn***)
29	~~0.2226 home~~ (***tn***)	0.2217 port (*fp*)	0.1807 visualization tool (***tp***)	**0.1989 xvid** (***tp***)
30	0.2176 ericsson xperia (*fp*)	0.2176 pixel resolution (***tp***)	0.1807 size visualization (***tp***)	~~0.1984 cell phone~~ (*fn*)

Table 5-14. Judgement applied over the top 30 terms of *dimensions 15* (with and without), *39* (without) and *13* (with) after discarding method.

Firstly, it is applied the discarding method on all the selected dimensions, following the steps in **Table 4-1**. The resultant graphics are located in **Annex D: Discarding method applied to LSI dimensions**. There, it is seen how there are some initial terms which are discarded by this method, although it is not the biggest problem, because in *dimension 3* and *dimension 5* there are a lot of retrieved terms that are wrong discarded, but they do not imply directly a bad performance. This fact may occur due to the high number of initial term of the *multimedia* product feature. It groups a lot of terms whose meanings are quite open and sometimes they may have different interpretations between each others. After application of the discarding method, the human being judgement is applied over these top 30 terms of the selected dimensions, which is showed in **Table 5-13** and **Table 5-14**. There, it is seen though *dimension 13* and *dimension 39* were selected by its potential relation with the *multimedia* context, but there are many mistakes to result a good performance. Precision and recall is presented in **Graphic 5-7**, where it is seen graphically what was thought, that *dimension 3* and *dimension 5*, in both performances, have retrieved almost the entire top 30 terms right, but the discarding method has discarded a lot of them, then, their recall remains not as good as their precision.

Their accuracy is also worse, but their performances in general are the best in this section and almost within the entire LSI performances. *Dimension 15*, in both performances, and *dimension 39* have the same problem with recall, but their precisions are worse than the first dimensions. As it is said, the last one, *dimension 13*, is not good selected to be a *multimedia dimension* at all, comparing with the first ones. Accuracy, precision and recall single graphics of all the dimensions analyzed in this section are located in **Annex G: Precision, recall and accuracy graphs of LSI**.



Graphic 5-7. Precision and recall graph of selected dimensions of *multimedia* on paragraphs.

Finally, based on the obtained results, the dimensions selected which best define the *multimedia* product features considering the initial terms extracted from the specifications are *dimension 3* and *dimension 5*, both in both performances, with and without *extra reviews*. From ***dimension 3*** in performance **without** *extra reviews*, terms such as *player*, *resolution*, *interface*, *file*, *picture*, *image*, *option*, *music*, *pixel*, *album*, *browser*, *sample*, *playback*, *internet* and *media*, are extracted with a precision of 95%, accuracy of 89.66% and recall of 86.36%. **With** them, the same terms are extracted, except ~~*browser*~~ and ~~*internet*~~, but plus *application* and *song*, with a precision of 86.36%, accuracy of 76.67% and recall of 82.61%. From ***dimension 5*** in performance **without** *extra reviews*, terms such as *music*, *player*, *sound*, *picture*, *quality*, *album*, *song*, *track*, *setting*, *noise*, *artist*, *equalizer*, *playlist*, *format* and *genre*, are extracted with a precision of 100%, accuracy of 66.67%, but a recall only of 65.52%. **With** them, the same terms, except ~~*setting*~~ and ~~*genre*~~, but plus *volume* and *mode*, are extracted with the same a precision of 100%, accuracy over 63.33%, but also a recall only over 62.07%. From ***dimension 15***, **without** *extra reviews*, terms such as *headset*, *volume*, *file* and *resolution*, are extracted with 100% of precision and accuracy, with recall of 31.25%. **With** *extra reviews*, only terms like *headset* and *volume* are extracted with 100% of precision and accuracy, but recall of 16.67%. From ***dimension 39*** **without** *extra reviews* only fm radio is extracted with

100% of precision and accuracy, but recall of 11.76%. Finally, from ***dimension 13*** **with** *extra reviews*, only the term *phone* is extracted with 100% of precision and accuracy, but recall of 15.38%.

5.3. PLSI analysis

5.3.1. Battery

5.3.1.1. Documents

In order to follow the steps described in **Table 4-1**, PLSI is run considering the environment described above, on documents, where each review is taken as an entity, obtaining the probabilities corresponding to the different matrices named before as $\widehat{U}$, $\hat{S}$ and $\widehat{V}^T$, but it is only needed the first one, because there it is stored $P(w_j|z_k)$, the probability of a word w_j appearing in a topic z_k. As in LSI (see Latent Semantic Indexing (LSI)), the dimensions which were analysed from the U_k matrix, now are called topics, but they are correctly considered now as groups of words semantically related based on PLSI's performances. Like in **Figure 5-4**, values are organized in vertical vectors.

First, it is taken a look at **Table 5-15**, where it is found a summarization of the best topics found in the $\widehat{U}$ matrix. It is seen that, unlike it happened with LSI, here any topic is repeated as well as the *k* parameter is incremented. In other words, it means that any execution is dependent on other one. It is found that there are different values for each topic on each performance. It is noticed that values are represented by a probability distribution, then, they are set from 0.0 to 1.0. Because of that, there are no negative values, but they are scored to 0.0 instead. In addition, it has been analyzed one by one the entire collection of topics, where at least one of the initial terms was placed in places close to the top. Sometimes there are topics where initial terms are better scored than in others, but they are placed in worse places. Then, as it is done in the LSI analysis, it is preferable considered the placement of the initial terms in the sorted topic. This means that, for example, in **Table 5-15**, *topic 18* with *k=150* and with *extra reviews* and *topic 154* with *k=200* without them, are selected instead of selecting *topic 11* or *topic 19* with *k=50* with *extra reviews*, whose initial terms such as *video playback* or *li-ion* are considerably higher scored than those from the first ones. *Topic 89* with *k=150* and *topic 4* with *k=200*, both without *extra reviews*, are also selected with the same particularity than the other two, *topic 4* has very good scored some initial terms, meanwhile *topic 89* does not.

Once selection is done, the application of the discarding method is carried out, getting the graphics located in **Annex E: Discarding method applied to PLSI** . As it is done with dimensions in LSI analysis, here it is shown a table with the top 30 terms of the four selected topics, where the discarding method and the human judgement are applied in order to extract some conclusions from each performance.

In **Table 5-16** it is appreciated how the discarding method has not discarded the excess of *false positives* (*fp*). However, when it has discarded some terms, they are wrong discarded, i.e. in *topic 18*, where terms are poorly scored, although the four discarded have a strongly semantic relationship from the sight of a human being.

	Without	With
K=50	× **Topic 41** · 56th with 0.4573 – stand-by time · 188th with 0.0048 – talk time · 210th ... → 5 values set to zero	× **Topic 11** · 30th with 0.9636 – video playback · 136th with 0.0221 – battery · 505th ... → 6 values set to zero × **Topic 19** · 27th with 0.8675 – li-ion · 178th ... → 8 values set to zero
K=100	× **Topic 8** · 10th with 0.9957 – stand-by time · 192nd with 0.0132 – battery · 273rd ... → 2 values set to zero × **Topic 26** · 26th with 0.4636 – battery · 204th with 0.1413 – talk time · 480th ... → 2 values set to zero	× **Topic 83** · 36th with 0.0084 – stand-by time · 93rd with 0.0019 – battery · 129th ... → 5 values set to zero × **Topic 98** · 13th with 0.9801 – video playback · 92nd with 0.0123 – battery · 484th ... → 2 values set to zero
K=150	× **Topic 60** · 2nd with 0.0203 – li-ion · 863rd ... → 2 values set to zero ✓ **Topic 89** · 5th with 0.3105 – battery · 180th with 0.0439 – talk time · 477th ... → 4 values set to zero	✓ **Topic 18** · 14th with 0.0099 – stand-by time · 19th with 0.0036 – capacity · 32nd with 0.0007 – talk time · 34th with 0.0006 – music playback · 37th with 0.0004 – battery · 39th ... → NO values set to zero × **Topic 46** · 15th with 0.2037 – battery · 265th ... → 5 values set to zero × **Topic 59** · 1st with 0.0239 – talk time · 698th ... → 3 values set to zero
K=200	✓ **Topic 4** · 3rd with 0.9927 – stand-by time 50th with 0.0105 – talk time · 69th with 0.0079 – battery · X ... → ALL values set to zero ✓ **Topic 154** · 12th with 0.2949 – battery · 48th with 0.1076 – talk time · 884th ... → 3 values set to zero	× **Topic 36** · 27th with 0.2305 – battery · 516th ... → 6 values set to zero × **Topic 65** · 2nd with 0.0179 – battery · 851st ... → 10 values set to zero

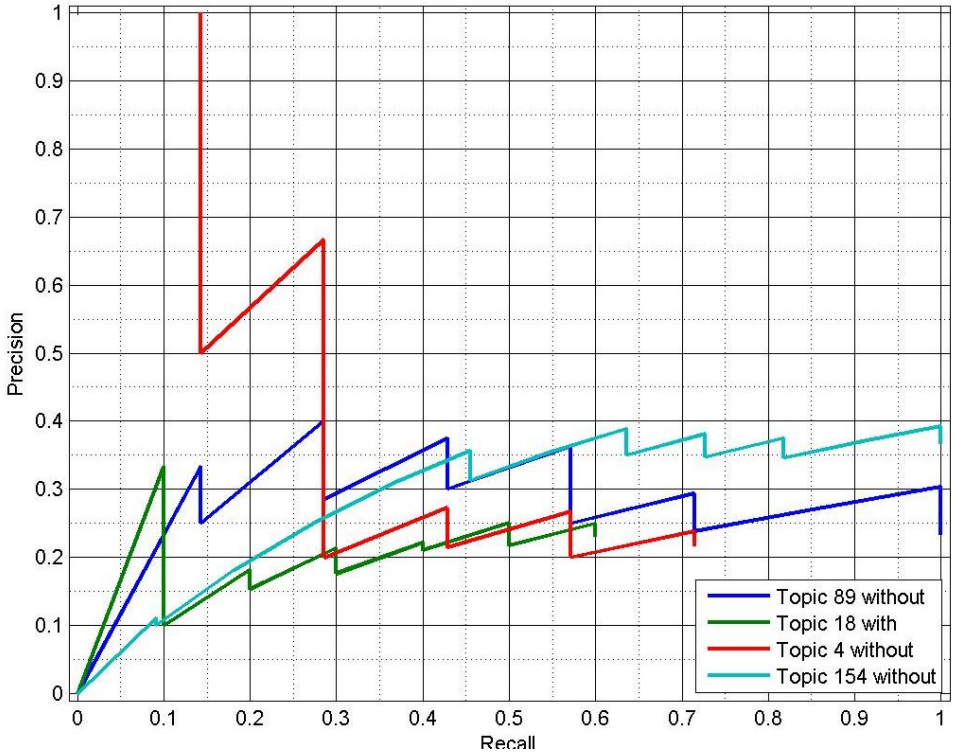
Table 5-15. Best scored dimensions of PLSI considering the *battery* product feature on documents.

After applying the human judgement, it has been proceeded to draw the results in the graphics corresponding to the accuracy, precision and recall located in **Annex H: Precision, recall and accuracy graphs of PSLI**. **Graphic 5-8** shows the precision and recall graphic obtained by the selected topics. There it is seen how all the scores of PLSI on these topics are considerably poor, because only one of them, *topic 4*, starts with a correct term retrieved. But it is the only one which is possible to be extracted from the whole performance of PLSI on documents looking for the *battery topic* which involves the whole battery context and some of its initial terms.

Following the enounced steps in **Table 4-1** and due to the commented poor performances, from ***topic 4*** in performance **without** *extra reviews*, only the term *offline* is extracted with precision of 100%, accuracy of 100% and recall of 14.29%.

Place	*Topic 89* without *k=150*	*Topic 18* with *k=150*	*Topic 4* without *k=200*	*Topic 154* without *k=200*
1	1.0000 fm transmit... (*fp*)	1.0000 argument (*fp*)	0.9942 offline (***tp***)	1.0000 character limit (*fp*)
2	0.5191 studiomelody (*fp*)	0.9999 sony ericsson … (*fp*)	0.9930 plate (*fp*)	0.9859 micoach (*fp*)
3	0.4284 gsm (***tp***)	0.9998 android smartp... (***tp***)	**0.9927 standby time** (***tp***)	0.6882 eartip (*fp*)
4	0.3998 landscape pos... (*fp*)	0.9998 evening (*fp*)	0.9924 linux (*fp*)	0.6567 lg handset (*fp*)
5	**0.3105 battery** (***tp***)	0.9998 sidetop (*fp*)	0.9920 mail (*fp*)	0.6438 capture quality (*fp*)
6	0.2837 pc edit (*fp*)	0.9975 sound quality (*fp*)	0.9889 pub (*fp*)	0.5217 lit (*fp*)
7	0.2205 land (*fp*)	0.9974 xenon (*fp*)	0.9848 crowd (*fp*)	0.4517 sample video (*fp*)
8	0.2077 call (***tp***)	0.9966 website (*fp*)	0.9829 fashion accessory (*fp*)	0.4467 bill (*fp*)
9	0.1965 end (*fp*)	0.9963 video preview (*fp*)	0.9804 innovator (*fp*)	0.4348 gsm (***tp***)
10	0.1808 sound (*fp*)	0.9927 platform (*fp*)	0.9779 qvga screen (*fp*)	0.4007 headset (*fp*)
11	0.1788 indicator (***tp***)	0.6617 flash-light (***tp***)	0.9747 launching (***tp***)	0.3355 charger ac3us... (***tp***)
12	0.1760 detail (*fp*)	0.6560 abundance (*fp*)	0.9648 aim (*fp*)	**0.2940 battery** (***tp***)
13	0.1670 rim (*fp*)	0.0200 cell phone market (*fp*)	0.5419 sample video (*fp*)	0.2854 indicator (***tp***)
14	0.1587 dial (*fp*)	**0.0099 standby time** (***tp***)	0.5405 bill (*fp*)	0.2751 battery life (***tp***)
15	0.1576 indoor picture (*fp*)	~~0.0057 3g~~ (*fn*)	0.4939 display system in... (***tp***)	0.2417 bit (*fp*)
16	0.1507 blackberry (*fp*)	0.0048 partner (*fp*)	0.4673 lit (*fp*)	0.2081 volume (*fp*)
17	0.1474 battery life (***tp***)	0.0048 fingerprint (*fp*)	0.3807 lag (*fp*)	0.1965 power (***tp***)
18	0.1463 top (*fp*)	0.0048 standard audio (*fp*)	0.3367 lg handset (*fp*)	0.1845 hour (***tp***)
19	0.1425 menus (*fp*)	**0.0036 capacity** (***tp***)	0.1665 upload (*fp*)	0.1832 cradle (*fp*)
20	0.1387 picture (*fp*)	0.0033 preset mode (*fp*)	0.0433 headset (*fp*)	0.1709 port (*fp*)
21	0.1384 life (*fp*)	0.0033 producer (***tp***)	0.0316 degree (***tp***)	0.1695 quality (***tp***)
22	0.1377 number (***tp***)	~~0.0031 time~~ (*fn*)	~~0.0284 battery life~~ (*fn*)	0.1665 way (*fp*)
23	0.1345 gps (***tp***)	0.0021 prominence (*fp*)	0.0244 show (*fp*)	0.1604 upload (*fp*)
24	0.1312 keypad (*fp*)	0.0021 i5500 (*fp*)	0.0207 shell (*fp*)	0.1559 switch (***tp***)
25	0.1291 menu (*fp*)	0.0021 hardware reset (*fp*)	~~0.0190 life~~ (***tn***)	0.1548 sound (*fp*)
26	0.1282 voice (*fp*)	~~0.0010 video~~ (*fn*)	~~0.0181 cradle~~ (***tn***)	0.1434 implementation (*fp*)
27	0.1174 nothing (*fp*)	0.0009 notifier (***tp***)	~~0.0172 user~~ (***tn***)	0.1354 performance (***tp***)
28	0.1156 example (*fp*)	~~0.0009 video call~~ (*fn*)	~~0.0169 point~~ (***tn***)	0.1334 device (***tp***)
29	0.1149 key (*fp*)	0.0008 tune (*fp*)	~~0.0156 video~~ (*fn*)	0.1316 conclusion (*fp*)
30	0.1138 earphone (*fp*)	0.0008 proof (*fp*)	~~0.0153 party~~ (***tn***)	0.1289 design (*fp*)

Table 5-16. Judgement applied over the top 30 terms of *topics 89* (without), *18* (with), *4* and *154* (without) after discarding method.



Graphic 5-8. Precision and recall graph of selected topics of *battery* on documents.

5.3.1.2. Sections

It is acted in the same way as in LSI, starting from the initial terms; it is looked for the topics which better describe the *battery product feature* defined by these terms, the ones which have better placed them on their sorted values. In these PLSI performances on sections, it is possible to find many topics whose initial terms are very good scored, but it happens that there are many that are "alone" well scored in highest places, but their topic does not suit the *battery topic*, and there are almost any semantically related term with the *battery context*.

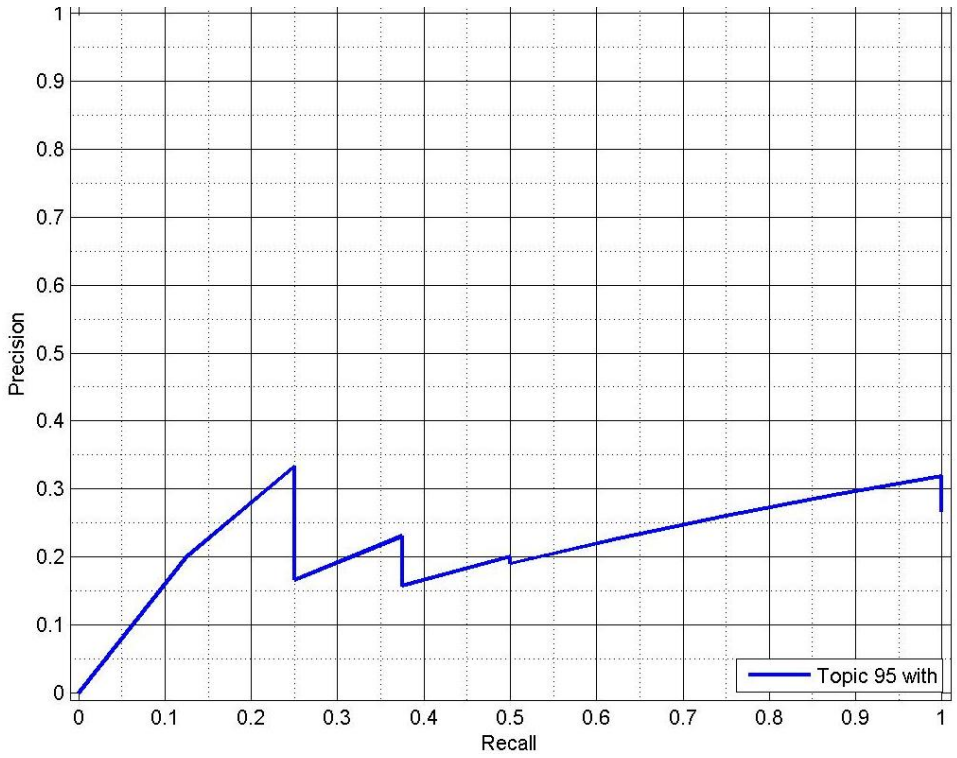
Place	*Topic 95,* with, *k=150*
1	0.9942 combo (*fp*)
2	0.9937 row (*fp*)
3	0.9936 addition (*fp*)
4	0.9930 calendar (*fp*)
5	0.9925 docking (***tp***)
6	0.9922 call (***tp***)
7	0.9919 ability (*fp*)
8	0.9916 page rendering (*fp*)
9	0.9909 menus (*fp*)
10	0.9875 operator (*fp*)
11	0.9865 par (*fp*)
12	0.9843 search (*fp*)
13	0.9800 mode (***tp***)
14	0.9798 samsung app (*fp*)
15	0.9776 representation (*fp*)
16	0.9763 app (*fp*)
17	0.9728 one (*fp*)
18	0.9680 background (*fp*)
19	0.9653 hand (*fp*)
20	**0.9600 battery** (***tp***)
21	0.9569 pixel resolution (*fp*)
22	0.9442 device (***tp***)
23	0.9393 time (***tp***)
24	0.5000 charger headphone port (***tp***)
25	0.5000 micro usb cable ca101 (***tp***)
26	0.4969 b3210 video (*fp*)
27	0.4846 ericsson w350 (*fp*)
28	0.4843 key pad (*fp*)
29	0.3425 endeavor hx1 (*fp*)
30	0.2471 sghm620 (*fp*)

Table 5-17. Judgement applied over the top 30 terms of *topic 95* (with) after discarding method.

Results of the PLSI's execution are located in **Table 10-2** in **Annex B: PLSI's running tables**. It is seen that PLSI's performances on sections are poorer than on documents, unlike it happens with LSI's performances, which improves on shorter scenarios. It only can be selected *topic 95* with *k=150* with *extra reviews*, although it is check that its values are not good at all, it is the best topic to fit the supposed *existence* of a *battery topic*. The application of the discarding method, which discards any term from the top 30 terms of the topic, and the human judgement application are shown in

After seeing how related are the terms of *topic 95*, it is drawn its corresponding, precision, recall and accuracy graphics located in **Annex H: Precision, recall and accuracy**

graphs of PSLI. In GRAPPHIC it is seen precision and recall behaviour, where it is seen how it is impossible to extract any term with a precision higher than 80%.



Graphic 5-9. Precision and recall graph of selected topics of *battery* on sections.

5.3.1.3. Paragraphs

Although it is supposed that PLSI method performs better with shorter contexts like sections and paragraphs, here it is found the poorest performance of PLSI for the *battery product feature*, which is located in **Annex B: PLSI's running tables**. There are topics which seem to have a high semantic relation with the *battery context* like *topic 149* without *extra reviews*, *topics 2*, *42,* 94, 57 and 77 with *extra reviews*, which have the best scored initial term with 1.0, which is the maximum value, but the rest of terms are scored with 0.0, which is the lowest value. This fact means that it is possible to find a remote semantic relation with the *battery context*, but it is very difficult. Only *topic 156* with *extra reviews* could be a candidate to be selected as a *battery topic*, but looking at its top 30 terms it is seen almost any semantically related term with the *battery product features*. Then, as a conclusion of the seen PLSI's performances, it is affirmed that in the battery context, the lower the number of terms semantically related with the initial ones can be found, the shorter the context is.

5.3.2. Organizer

5.3.2.1. Documents

In this part, it is talked about the *organizer* product feature, taking the same steps described in **Table 4-1** to achieve the goal of extracting the major number of well semantically related terms with the initial terms from the *organizer* product feature, satisfying a high precision. The first analysis, as it is done with the *battery* feature, is done on documents, considering each review as an entity to input of PLSI.

Place	*Topic 119* without *k=150*	*Topic 135* without *k=150*	*Topic 176* without *k=200*
1	0.9994 text message (***tp***)	1.0000 server application (***tp***)	0.7616 wifi router app (***tp***)
2	0.9955 profile button (***tp***)	**0.9999 document viewer** (***tp***)	0.5542 edge (*fp*)
3	0.9955 xperia x10 mini ... (*fp*)	0.9956 mobile application (***tp***)	0.4885 top (*fp*)
4	0.9955 netfront (***tp***)	0.9956 beta (***tp***)	0.4631 bottom (*fp*)
5	0.9950 3gs video (*fp*)	0.9950 creation (***tp***)	0.4035 ringtone (***tp***)
6	0.9926 precision (***tp***)	0.9938 designer model (*fp*)	0.3913 reader (***tp***)
7	0.9926 frame rate (*fp*)	0.9929 metal evolution (*fp*)	0.3886 europe asia (*fp*)
8	0.9914 iphone owner (*fp*)	0.9927 search option (***tp***)	0.3881 attendee (***tp***)
9	0.9914 avi (*fp*)	0.9915 onslaught (*fp*)	0.3749 point (*fp*)
10	0.9912 tale (*fp*)	0.9914 iso (*fp*)	0.3749 memory (***tp***)
11	0.9887 portrait keyboard (***tp***)	0.9912 volume level (***tp***)	0.3739 group (***tp***)
12	0.9885 hardware (*fp*)	0.9860 skyfire (***tp***)	0.3698 advertisement ... (*fp*)
13	0.9864 launch icon (***tp***)	0.9837 store (***tp***)	**0.3649 alarm** (***tp***)
14	0.9861 menu home screen (***tp***)	0.9836 york time (*fp*)	0.3431 style (***tp***)
15	0.9835 middle (*fp*)	0.9800 user guide (***tp***)	0.3319 field (***tp***)
16	0.9781 couple (*fp*)	0.9794 load (***tp***)	0.3311 power (*fp*)
17	0.9779 cloudy day (*fp*)	0.9707 range (***tp***)	0.3244 internet (***tp***)
18	0.9778 expansion card (*fp*)	0.9439 web site (***tp***)	0.3123 htc (*fp*)
19	0.9775 genre (***tp***)	0.6374 question (***tp***)	**0.3087 note** (***tp***)
20	**0.9762 calendar** (***tp***)	0.6313 s8500 video (*fp*)	0.2913 htc desire (*fp*)
21	0.9739 format (***tp***)	0.6212 competition (*fp*)	0.2813 part (*fp*)
22	0.9726 sony ericsson jalou (*fp*)	0.5320 stock market (*fp*)	0.2798 n810 (*fp*)
23	0.9643 oz (*fp*)	0.5300 while (*fp*)	0.2730 column view option (***tp***)
24	0.9634 combination (*fp*)	0.4610 daily brief (***tp***)	0.2642 explorer (***tp***)
25	0.9627 fluid (*fp*)	0.3291 fox (*fp*)	0.2585 video (*fp*)
26	0.9585 access (***tp***)	0.3281 turnbyturn voice (*fp*)	0.2507 look (*fp*)
27	0.9520 line (*fp*)	0.2208 xperiax10 (*fp*)	0.2503 wifi (*fp*)
28	0.5486 language (***tp***)	0.1122 creator (***tp***)	0.2479 end (*fp*)
29	0.5319 brass (*fp*)	0.0995 iphone (*fp*)	0.2439 windows (*fp*)
30	0.5280 bird (*fp*)	0.0749 slide (*fp*)	0.2357 matter (***tp***)

Table 5-18. Judgement applied over the top 30 terms of *topic 119* (without), *topic 135* (without) and *topic 176* (without) after discarding method.

After running PLSI on documents, it is extracted the table (see **Table 10-4**) with the topics where the initial terms are best scored. It is possible to see how the best topics have at most 2 initial terms within the top 30 terms, but some of them are not selected to be a potential *organizer* topic. For example, *topic 186* has *task* and *note* within the top 30, and *calendar* very close in the 42nd place, but their scores are very low not to be deleted by the discarding method. Indeed, only *topic 176*, from the selected topics, has more than 1 initial term within its top 30. The rest of them have been selected although they have only a single

initial term well ranked and scored, but the terms that surround it are strongly related with the *organizer* feature, at least to reach a high precision in the first terms.

After applying the discarding method, whose results can be seen in **Annex E: Discarding method applied to PLSI** , it is seen that any term is rejected, then, it is concluded that all the terms are strongly related, semantically or not, is seen with the human judgement application, which is shown in tables **Table 5-18** and **Table 5-19**.

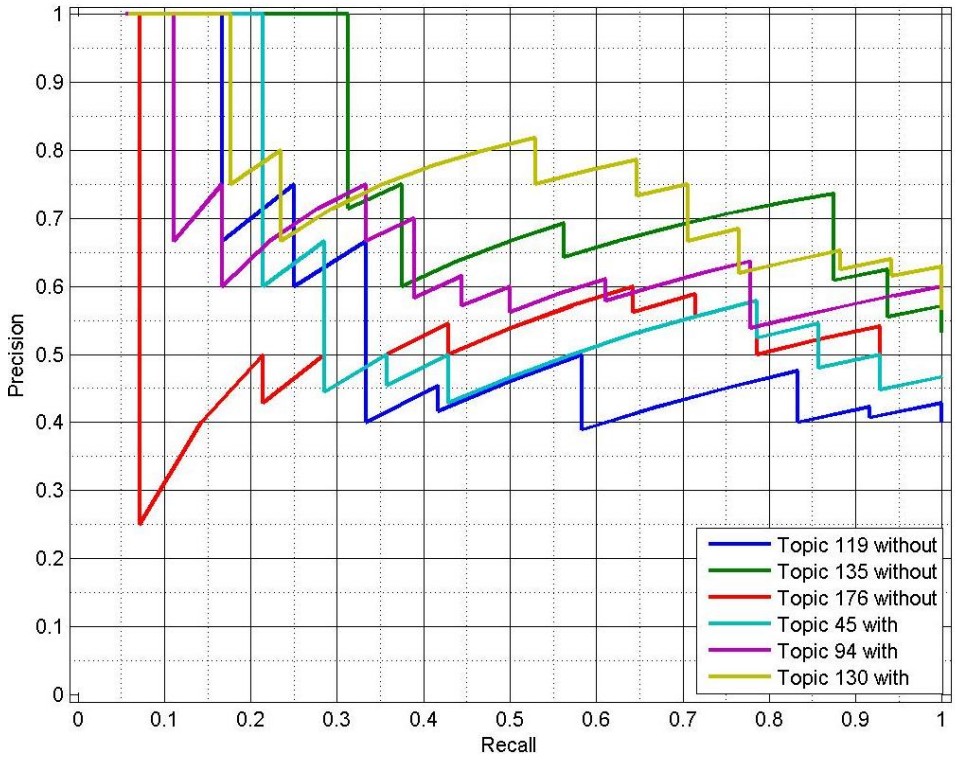
Place	*Topic 45* with *k=50*	*Topic 94* with *k=150*	*Topic 130* with *k=200*
1	0.9682 deadline (***tp***)	0.9999 search box (***tp***)	0.9998 hand operation (***tp***)
2	0.9465 transcriber (***tp***)	0.9999 server application (***tp***)	0.9998 dial option (***tp***)
3	0.8850 internet source (***tp***)	0.9999 gps unit (*fp*)	0.9998 search box (***tp***)
4	0.8330 computer version (*fp*)	0.9999 doubletapping (***tp***)	0.9998 gps unit (*fp*)
5	0.7959 tomtom navigator (*fp*)	0.9999 cycle (*fp*)	0.9998 screen widget (***tp***)
6	0.7494 access (***tp***)	0.9999 samsung app (***tp***)	0.9998 gigabyte gsmart s1205 (*fp*)
7	0.7442 bottom (*fp*)	0.9999 entertainment option (***tp***)	0.9997 phone keypad (***tp***)
8	0.7425 ok (*fp*)	0.9999 screen widget (***tp***)	0.9997 server application (***tp***)
9	0.7196 top (*fp*)	0.9999 gigabyte gsmart s1205 (*fp*)	**0.9997 document viewer** (***tp***)
10	0.7169 visualization window (***tp***)	0.9998 dolphin (***tp***)	0.9997 entertainment option (***tp***)
11	0.6970 lack (*fp*)	0.9997 legibility (*fp*)	0.9996 attaching multimedia (***tp***)
12	0.6966 level (***tp***)	0.9997 cam (*fp*)	0.9996 el (*fp*)
13	0.6758 pc (*fp*)	0.9997 hand operation (***tp***)	0.9996 keystroke (***tp***)
14	0.6721 europe (*fp*)	0.9997 combo (*fp*)	0.9996 doubletapping (***tp***)
15	0.6540 mobile professional device (***tp***)	0.9997 attaching multimedia (***tp***)	0.9996 holster (*fp*)
16	0.6482 page (***tp***)	0.9997 el (*fp*)	0.9996 work phone (***tp***)
17	0.6459 internet (***tp***)	0.9997 keystroke (***tp***)	0.9995 legibility (*fp*)
18	**0.6394 note** (***tp***)	0.9997 zoom capability (***tp***)	0.9995 combo (*fp*)
19	0.6369 zoom (***tp***)	0.9997 network signal (*fp*)	0.9995 numpad (***tp***)
20	0.6360 look (*fp*)	0.9997 work phone (***tp***)	0.9995 camera application (*fp*)
21	0.6245 player (*fp*)	0.9996 dial option (***tp***)	0.9995 network signal (*fp*)
22	0.6236 image (***tp***)	0.9996 email app (***tp***)	0.9994 dolphin (***tp***)
23	0.6176 direction (*fp*)	0.9996 camera application (*fp*)	0.9993 zoom capability (***tp***)
24	0.6164 hsp (*fp*)	0.9995 fire (*fp*)	0.9993 dog (*fp*)
25	0.6137 wifi (*fp*)	0.9995 blurring (*fp*)	0.9993 menu icon (***tp***)
26	0.6116 resolution (***tp***)	0.9995 holster (*fp*)	0.9992 fire (*fp*)
27	0.6112 navigator (*fp*)	0.9994 phone keypad (***tp***)	0.9991 email app (***tp***)
28	0.6095 light (*fp*)	0.9993 numpad (***tp***)	0.9990 cam (*fp*)
29	0.6073 power (*fp*)	**0.9993 document viewer** (***tp***)	0.9990 blurring (*fp*)
30	0.6067 day (***tp***)	0.9993 webkit (***tp***)	0.9990 30fps (*fp*)

Table 5-19. Judgement applied over the top 30 terms of *topic 45* (with), *topic 94* (with) and *topic 130* (with) after discarding method.

Once terms have been evaluated, it has to be drawn the performance of PLSI on documents, represented by these selected topics, regarding the seeking of the *organizer* product feature. Precision, accuracy and recall graphics of these topics are located in **Annex H: Precision, recall and accuracy graphs of PSLI**. In **Graphic 5-10**, it is possible to see how in *topic 135* is the best, because its line in precision and recall's graphic remains high more than the others and due to that, it gets more new term semantically related with the first ones. If it is taken a look at the precision graphic, this topic has less precision, but it is still remains higher than 80% getting the highest recall, like it is seen in the recall graphic.

To summarize the obtained results, from ***topic 119* without** *extra reviews*, terms such as *text message* and *profile button* are extracted with precision and accuracy of 100%, and recall

of 16.67%. From ***topic 135*** **without** *extra reviews*, terms such as *server application*, *mobile application*, *beta* and *creation*, are extracted with precision and accuracy of 83.33%, and recall of 31.25%. From ***topic 176*** **without** *extra reviews*, only the term *wifi router app* is extracted with precision and accuracy of 100%, and recall of 7.14%. From ***topic 45*** **with** *extra reviews*, terms such as *deadline*, *transcriber* and *internet source*, are extracted with precision and accuracy of 100%, and recall of 21.43%. From ***topic 94*** **with** *extra reviews*, terms such as *search box* and *server application*, are extracted with precision and accuracy of 100%, and recall of 11.11%. From ***topic 130*** **with** *extra reviews*, terms such as *hand operation*, *dial option* and *search box*, are extracted with precision and accuracy of 100%, and recall of 17.65%.



Graphic 5-10. Precision and recall graph of selected topics of *organizer* on documents.

5.3.2.2. Sections

PLSI is run on sections, in the same way as on documents, but considering each review as a group of parts, delimited by the titles which appear in it, like “Introduction and design” or “Performance”. It is followed the steps described in **Table 4-1** to test the method in this scenario.

First of all, after running PLSI, topics where initial terms are better scored, are located in **Table 10-5**. There, those topics, which have more potential terms on their first places in the sorted topic, are selected to be the potential *organizer* topic. Although there are topics which have higher scored the initial terms, i.e., *topic 49* without *extra reviews* and with k=100 which has only the 2 initial terms higher than zero within the top 30 terms, but its first terms are more semantically related with the *organizer* ones than in other topics. Other atypical circumstance in these PLSI performances is the appearance of a topic, *topic 60* without *extra*

reviews and k=150, which has 3 initial terms very well placed, but not very well scored (all 3 under 0.1). Moreover, *topic 154* without *extra reviews* and with k=200, has as the top of the sorted topic an *organizer* initial term, having in the same circumstances *topic 27*, but whose performance is lower, because its first terms are not semantically related with the *organizer* feature.

Once topics are selected to be evaluated their behaviours, discarding method is applied getting the graphics stored in the **Annex E: Discarding method applied to PLSI topics**. Results from this application and the human judgement application can be found in the **Table 5-20** and **Table 5-21**, on the top 30 of selected topics, observing that discarding method, as it happens with documents above, has a subtle effect on topics, discarding only 2 terms in *topic 6* (1 wrong discarded) and 5 in *topic 164* (also 1 wrong discarded).

Place	*Topic 28* without *k=50*	*Topic 6* without *k=100*	*Topic 49* without *k=100*	*Topic 60* without *k=150*
1	1.0000 pixel resolution ... (***tp***)	1.0000 error (***tp***)	0.9966 value (***tp***)	0.1384 refinement (***tp***)
2	0.9967 legibility (f*p*)	0.9999 s1205 (*fp*)	0.9966 gigabyte (***tp***)	**0.0963 memo** (***tp***)
3	0.9953 work phone (***tp***)	0.9998 color representation (***tp***)	0.9965 mpeg4 (*fp*)	**0.0913 office** (***tp***)
4	0.9952 doubletapping (***tp***)	0.9995 register (***tp***)	0.9964 equalizer preset (*fp*)	0.0732 gallery (***tp***)
5	0.9949 error (***tp***)	0.9993 user interface (***tp***)	0.9964 tetris (*fp*)	0.0731 anything (*fp*)
6	0.9946 wqvga (***tp***)	0.9984 qwerty keyboard (***tp***)	0.9960 camera button (*fp*)	**0.0674 task** (***tp***)
7	0.9924 touchwiz interface (***tp***)	0.9976 interaction (***tp***)	**0.9958 calculator** (***tp***)	0.0653 contrast (***tp***)
8	0.9921 menu icon (***tp***)	**0.9969 document viewer** (***tp***)	0.9956 doubletapping (***tp***)	0.0653 widget (***tp***)
9	0.9911 touchwiz phone (***tp***)	0.9968 menu icon (***tp***)	0.9954 flash-indoor sample (***tp***)	0.0606 myspace (*fp*)
10	0.9904 fare (*fp*)	0.9954 webos (***tp***)	0.9951 touchwiz interface (***tp***)	0.0514 browser (***tp***)
11	0.9901 dog (*fp*)	0.9940 text message (***tp***)	0.9947 touchscreen (***tp***)	0.0479 art (*fp*)
12	0.9897 customization (***tp***)	0.4993 omniapro b7610... (*fp*)	0.9944 search box (***tp***)	0.0468 entertainment (***tp***)
13	0.9896 map (*fp*)	0.4841 spending time (*fp*)	0.9941 clock (***tp***)	0.0446 unit (***tp***)
14	0.9891 amount (*fp*)	0.3333 ring id (***tp***)	0.9940 sluggishness (*fp*)	0.0405 step (***tp***)
15	0.9888 samsung app (***tp***)	0.3333 bit inconvenient (*fp*)	0.9935 bar (*fp*)	0.0405 lg (*fp*)
16	0.9883 album art (*fp*)	0.2499 rds radio (*fp*)	0.9928 music player (*fp*)	0.0399 degree view (*fp*)
17	0.9878 text message (***tp***)	0.2414 star iii (*fp*)	**0.9923 stopwatch** (***tp***)	0.0398 resolution (***tp***)
18	0.9867 holster (*fp*)	0.2368 disaster (*fp*)	0.9923 question (***tp***)	0.0387 twitter (*fp*)
19	0.9865 line-up (***tp***)	0.1458 sketch (*fp*)	0.9922 menus (***tp***)	0.0387 windows phone (*fp*)
20	0.9820 rest (*fp*)	0.1081 rocket (*fp*)	0.9913 google maps (*fp*)	0.0386 network (*fp*)
21	0.9818 radio (*fp*)	0.0471 treble (*fp*)	0.9913 par (*fp*)	0.0386 exchange (***tp***)
22	0.9812 system (***tp***)	0.0430 email address (***tp***)	0.9898 icon (***tp***)	0.0383 market (*fp*)
23	**0.9789 memo** (***tp***)	0.0164 voice navigation ... (*fp*)	0.9890 list (***tp***)	0.0364 magic (*fp*)
24	0.9767 hour (***tp***)	~~0.0104 function~~ (*fn*)	0.9870 world (*fp*)	0.0350 engine (*fp*)
25	0.9732 screenand (*fp*)	0.0103 fm (*fp*)	0.9856 life (*fp*)	0.0344 introduction (*fp*)
26	0.9732 currency exchange (***tp***)	0.0098 contact (***tp***)	0.9732 case (*fp*)	0.0338 contact view (***tp***)
27	0.9732 mini golf (*fp*)	0.0083 message (***tp***)	0.9705 multimedia (*fp*)	0.0336 thing (*fp*)
28	0.9731 performance thank (*fp*)	0.0079 usa (*fp*)	0.9676 interface (***tp***)	0.0323 adopt (*fp*)
29	0.9731 email protocol (***tp***)	0.0076 new (*fp*)	0.5055 home screen interface (***tp***)	0.0322 competitor (*fp*)
30	0.9731 adobe photoshop... (*fp*)	~~0.0065 latter~~ (***tn***)	0.4997 bp5mnokia charger (*fp*)	0.0319 fresh (*fp*)

Table 5-20. Judgement applied over the top 30 terms of *topic 28* (without), *topic 6* (without), *topic 49* (without) and *topic 60* (without) after discarding method.

After applying steps 3 and 4 from **Table 4-1**, precision, accuracy and recall are calculated. Those graphics are located in **Annex H: Precision, recall and accuracy graphs of PSLI**, but here it is only shown the precision and recall graphic, to understand the differents behaviours of the method by these topics on sections. It is seen in **Graphic 5-11**, how *topic 11* is the topic which fits better the *organizer* feature, remaining and ending with higher precision and recall than the rest of the topics. However, between them there are more remarkable

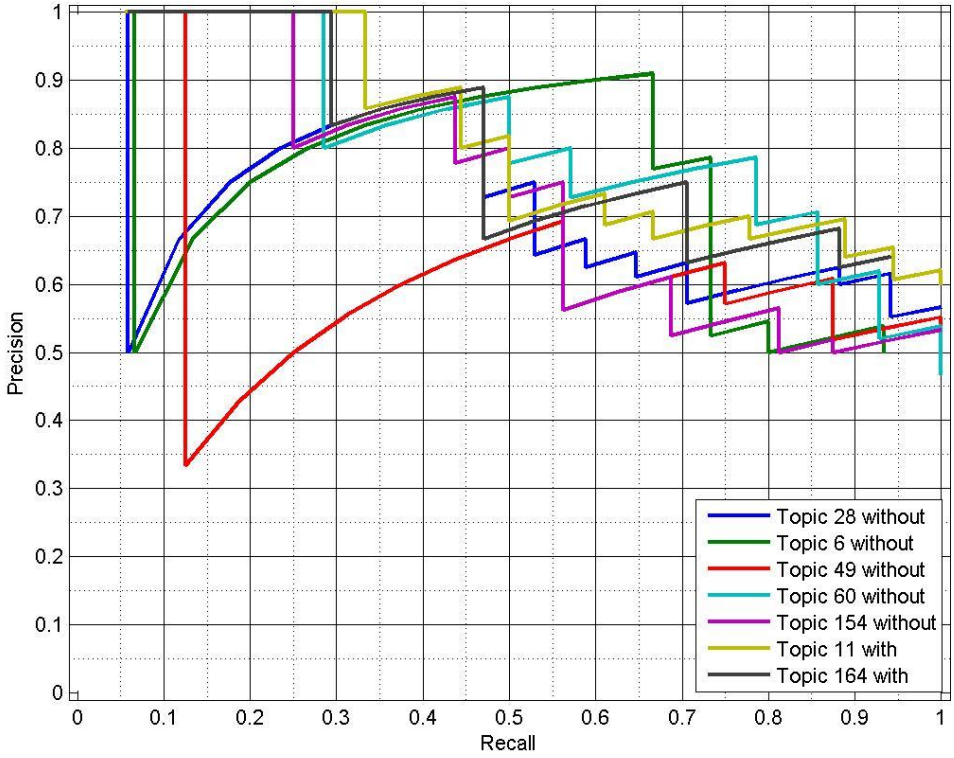
topics such as *topic 60*, *topic 154* and *topic 164*, with maintain a high precision and recall too. Unlike it happens with documents, PLSI performs better with *extra reviews*, where topics have reached more new terms. To summarize the whole performance of PLSI on sections, it is explained below.

Place	*Topic 154* without *k=200*	*Topic 11* with *k=100*	*Topic 164* with *k=200*
1	**0.9969 calendar** (***tp***)	0.9996 search button (***tp***)	1.0000 search file explorer (***tp***)
2	0.9956 qvga resolution (***tp***)	0.9996 flash indoor sample (***tp***)	0.9997 anniversary holiday (***tp***)
3	0.9950 fact (***tp***)	0.9986 application menu (***tp***)	0.9985 calendar contact (***tp***)
4	0.9926 recorder (***tp***)	0.9985 subscription (***tp***)	0.9925 day (***tp***)
5	0.9911 shot (*fp*)	0.9919 server application (***tp***)	0.6979 day event (***tp***)
6	0.9904 example (***tp***)	**0.9916 stopwatch** (***tp***)	0.6667 santa (*fp*)
7	0.9873 capability (***tp***)	0.9886 network signal (*fp*)	0.6660 time duration (***tp***)
8	0.9745 number (***tp***)	0.9880 java (***tp***)	0.5953 program menu (***tp***)
9	0.9195 camera (*fp*)	0.9879 action (***tp***)	0.5091 windows mobile (***tp***)
10	**0.9154 world clock** (***tp***)	0.9870 camera button (*fp*)	0.3743 truffle (*fp*)
11	0.5038 stigma (*fp*)	0.9865 screen widget (***tp***)	0.3589 endeavor hx1 (*fp*)
12	0.4753 screen play (***tp***)	0.9814 music player (*fp*)	0.3547 request (*fp*)
13	0.4569 battery endurance (*fp*)	0.9812 dog (*fp*)	0.3229 privacy (***tp***)
14	0.4149 smartphone today (*fp*)	0.9758 touch (***tp***)	0.2913 priority (***tp***)
15	0.4135 dpad joystick (*fp*)	0.9715 user (***tp***)	0.2830 image file (***tp***)
16	0.3321 camera software (*fp*)	0.9526 plastic (*fp*)	**0.1391 task** (***tp***)
17	0.3109 confirmation key (***tp***)	0.9517 page (***tp***)	0.0541 score (*fp*)
18	0.1957 note application (***tp***)	0.9503 nothing (*fp*)	0.0218 wave (*fp*)
19	0.1101 player fm (*fp*)	0.9310 name (***tp***)	0.0217 s60 phone (*fp*)
20	0.0936 htc solution (*fp*)	0.9152 device (***tp***)	0.0215 internet connection (***tp***)
21	0.0926 brick (*fp*)	0.8950 hand (*fp*)	0.0156 week (***tp***)
22	0.0681 pc sync (***tp***)	0.7731 second (***tp***)	~~0.0123 samsung wave~~ (***tn***)
23	0.0553 home office (***tp***)	0.5219 email system (***tp***)	0.0121 question (***tp***)
24	0.0520 mp3 file (*fp*)	0.5000 micro usb ... (*fp*)	0.0114 head (*fp*)
25	0.0392 acer liquid a1 (*fp*)	0.5000 bp5mnokia (*fp*)	0.0098 card (*fp*)
26	0.0307 samsung pixon (*fp*)	0.4996 mini qwerty keyboard (***tp***)	**0.0085 converter** (***tp***)
27	0.0191 privacy (***tp***)	0.4959 key row (*fp*)	~~0.0072 appointment~~ (*fn*)
28	0.0176 head (*fp*)	0.4924 sim windows phone (*fp*)	~~0.0064 phone start~~ (***tn***)
29	0.0158 bold (***tp***)	0.4561 interface phone (***tp***)	~~0.0048 prior~~ (***tn***)
30	0.0140 color code (***tp***)	0.3569 process power (*fp*)	~~0.0046 day week~~ (***tn***)

Table 5-21. Judgement applied over the top 30 terms of *topic 154* (without), *topic 11*(with) and *topic 164* (with) after discarding method.

Without *extra reviews*, all the selected topics have achieved some results, ones more than others, and the same has occurred **with** them. From *topic 28*, term *pixel resolution width* is extracted with precision and accuracy of 100%, but recall of 5.88%. From ***topic 6*** also only the term *error* is extracted with 100% of precision and accuracy, but a recall of 6.67%. From ***topic 49***, terms such as *value* and *gigabyte* are extracted with 100% of precision and accuracy, but a recall of 12.50%. Better results are given by ***topic 60***, whose terms like *refinement*, *gallery*, *contrast* and *widget*, are extracted with 87.50% of precision and accuracy, with a recall of 50%. One more is got from ***topic 154***, still **without** *extra reviews*, terms such as qvga resolution, fact, recorder, example, number and capability, are extracted with a precision and recall of 87.50%, but a recall of 473.75%. **With** *extra reviews*, from ***topic 11***, terms such as *search button*, *flash indoor sample*, *application menu*, *subscription*, *server application*, *java*, *action* and *screen widget*, are extracted with 81.82% of precision and accuracy and recall of 50%. Finally, from ***topic 164***, terms such as *search file explorer*, *anniversary holiday*, *calendar*

contact, *day*, *day event*, *time duration*, *program menu* and *windows mobile*, are extracted with 88.89% of precision and accuracy, but a recall of 47.06%. These two last topics are the best performances of PLSI on sections, differing only on a higher precision (*topic 11*) and a lower recall (*topic 164*).



Graphic 5-11. Precision and recall graph of selected topics of *organizer* on sections.

5.3.2.3. Paragraphs

In this part, performance of PLSI run on paragraphs is covered, where it happens something similar like with the *battery* product feature, but because *organizer* feature has more initial terms. In **Annex B: PLSI's running tables**, it is possible to see how *organizer* initial terms are better scored than *battery* ones on paragraphs performance. Although in performances with and without *extra reviews* with k=50 there is any topic susceptible to be an *organizer* topic. However, in performances with k=100 and k=150 is found topics with initial terms within their top 30 terms. Only *topic 26* and *topic 107* are selected from these last performances respectively, without and with *extra reviews*, respectively. Despite of not being the best topics selected, because there are topics with better scored initial terms, they are those whose performance is the one which achieve higher precision and recall to get more terms.

After selecting topics, the application of the discarding method is done and its results can be seen in **Annex E: Discarding method applied to PLSI topics**. There it is possible to see how any terms is discarded, then, it is supposed that the entire top 30 terms are strongly related between each other, because they have one of the five highest scores given in PLSI

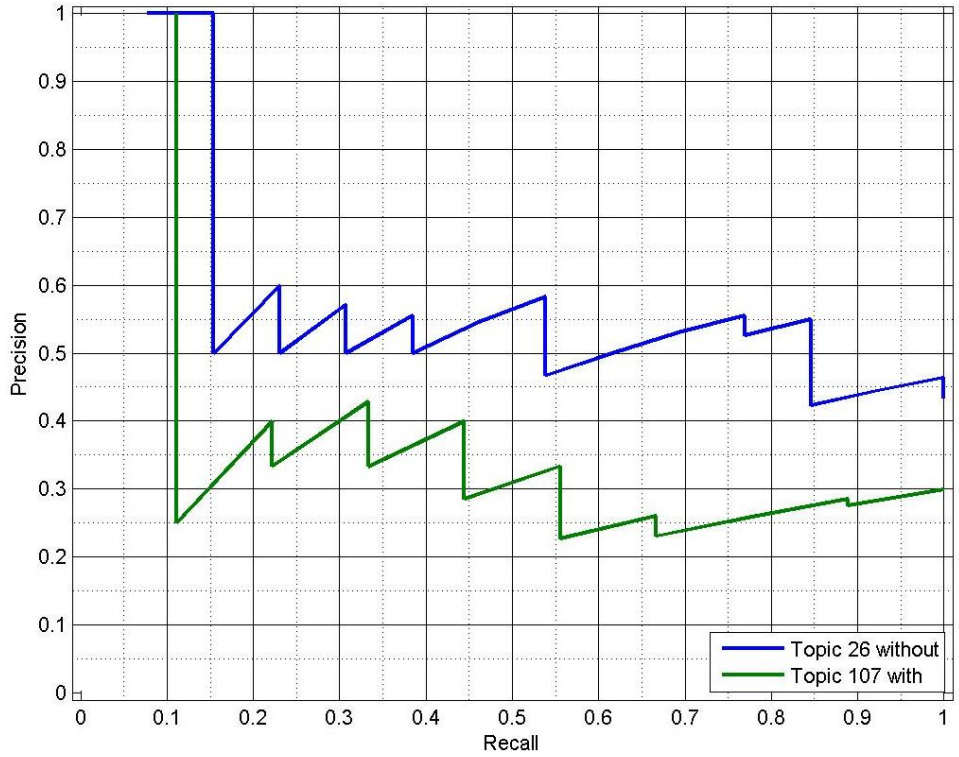
performance. Human judgement is applied next and its results can be viewed in the following **Table 5-22**.

Place	*Topic 26* without *k=100*	*Topic 107* with *k=150*
1	1.0000 suggestion (***tp***)	1.0000 sns support (***tp***)
2	1.0000 symbol (***tp***)	1.0000 chicken (*fp*)
3	1.0000 mini sample video (*fp*)	1.0000 device capture video (*fp*)
4	1.0000 handset sport (*fp*)	1.0000 customer support (*fp*)
5	1.0000 household (***tp***)	1.0000 text field (***tp***)
6	1.0000 powermat (*fp*)	1.0000 armani b7620 (*fp*)
7	1.0000 filename (***tp***)	1.0000 gallery (***tp***)
8	1.0000 msn messenger (*fp*)	1.0000 mean compact (*fp*)
9	1.0000 qcif resolution (***tp***)	1.0000 assassin (*fp*)
0	1.0000 park lot (*fp*)	1.0000 alarm sound (***tp***)
11	**1.0000 currency convert** (***tp***)	1.0000 win (*fp*)
12	1.0000 software application (***tp***)	1.0000 satellite data update (*fp*)
13	1.0000 fifa (*fp*)	1.0000 earphones user (*fp*)
14	1.0000 ericsson satio sample (*fp*)	1.0000 mono (*fp*)
15	1.0000 ingenuity (*fp*)	1.0000 easy keyboard (***tp***)
16	1.0000 preview window (***tp***)	1.0000 mobile excel (*fp*)
17	1.0000 nokia ovi suite (***tp***)	1.0000 game platform (*fp*)
18	1.0000 window mobile profess...(***tp***)	1.0000 realaudio (*fp*)
19	1.0000 s7220 (*fp*)	1.0000 n95samsung sghg... (*fp*)
20	1.0000 level up (***tp***)	1.0000 area code (*fp*)
21	1.0000 sepia (*fp*)	1.0000 missed call (*fp*)
22	1.0000 push page (*fp*)	1.0000 the contact (*fp*)
23	1.0000 i8910 hd (*fp*)	1.0000 application support (***tp***)
24	1.0000 friendliness (*fp*)	0.9999 arm (*fp*)
25	1.0000 power adapter (*fp*)	0.9999 variant (*fp*)
26	1.0000 window mount (*fp*)	0.9999 reject (*fp*)
27	1.0000 sound file (***tp***)	0.9999 sending (***tp***)
28	1.0000 single icon (***tp***)	0.9999 column (***tp***)
29	1.0000 factory set (*fp*)	0.9999 spb benchmark (*fp*)
30	1.0000 develop community (*fp*)	**0.9998 pdf** (***tp***)

Table 5-22. Judgement applied over the top 30 terms of *topic 26* (without) and *topic 107* (with) after discarding method.

Following the steps announced in **Table 4-1** to check each performance on each scenario, precision, accuracy and recall graphics are drawn and located in **Annex H: Precision, recall and accuracy graphs of PSLI**. To see how it has achieved the goal of getting the major number of new terms semantically related with the *organizer* feature, GRAPHIC shows precision vs. recall values. It is seen how *topic 26* performs better than *topic 107*, but both acts considerably poorer than the rest seen on documents and sections. It supports the idea of that PLSI performs better on documents than LSI, but poorer in paragraphs, where they are too much and too reduced to find clearly topics related with the product features proposed in this research.

To summarize the obtained results, it is affirmed that from ***topic 26* without** *extra reviews*, terms such as *suggestion* and *symbol* are extracted with 100% of precision and accuracy, but 15.38% of recall. Meanwhile, from ***topic 107* with** *extra reviews*, only *sns support* term is extracted with 100% of precision and accuracy, but 11.11% of recall.



Graphic 5-12. Precision and recall graph of selected topics of *organizer* on paragraphs.

5.3.3. Multimedia

5.3.3.1. Documents

As it is done before, PLSI is run on documents scenario, where reviews are taken as an entity, regarding in this entire chapter the *multimedia* product feature. Taking the initial terms collected from the technical specifications, the steps enounced in **Table 4-1**, where the first step to take is to select the most relevant topics susceptible to be the *multimedia* topic, which is supposed to exist, at least one.

Place	*Topic 56* without *k=150*	*Topic 25* without *k=200*	*Topic 45* without *k=200*
1	0.9999 sample picture (***tp***)	0.9982 frontfacing (***tp***)	1.0000 fm transmitter advertisement (***tp***)
2	0.9996 pioneer (***tp***)	0.9961 profile button (***tp***)	1.0000 video share (***tp***)
3	0.9994 transmitter (***tp***)	0.9961 netfront (***tp***)	0.9899 question (***tp***)
4	0.9988 entry method (***tp***)	0.9928 precision (***tp***)	0.9866 wallpaper (***tp***)
5	0.9978 human (*fp*)	0.9927 fashion device (*fp*)	0.9862 s8500 video (*fp*)
6	0.9966 nuvifone g60 (*fp*)	0.9914 basic (*fp*)	0.9859 8gb (*fp*)
7	0.9948 event entry (*fp*)	0.9904 reason (*fp*)	0.9842 carl zeiss tessar optics (***tp***)
8	0.9939 codec (***tp***)	**0.9852 avi** (***tp***)	0.9787 store (***tp***)
9	0.9929 set menus (*fp*)	0.9744 camcorder (***tp***)	0.9718 xperiax10 (*fp*)
10	0.9925 ring (***tp***)	0.9738 webkit browser (***tp***)	0.9682 difference (*fp*)
11	0.9900 quadband umts (*fp*)	0.9737 middle (*fp*)	0.9614 half (*fp*)
12	0.0615 map (*fp*)	0.8648 calculator (*fp*)	0.9380 web site (***tp***)
13	~~0.0237 app~~ (*fn*)	0.7598 fast (*fp*)	0.3143 lock (***tp***)
14	0.0234 samsung wave (*fp*)	0.7174 iphone 3gsapple ... (*fp*)	0.2829 creator (*fp*)
15	0.0208 sender (***tp***)	0.6467 whistle (*fp*)	0.1885 prism (*fp*)
16	0.0206 platform experience (*fp*)	0.6452 diva folder (*fp*)	0.1824 music control (***tp***)
17	~~0.0186 platform~~ (***tn***)	0.6088 bulk (*fp*)	0.1643 touch (***tp***)
18	0.0180 samsung wave s8500 (*fp*)	0.5643 037 (*fp*)	0.1484 soulb (*fp*)
19	0.0158 wave s8500 (*fp*)	0.5094 android user (*fp*)	0.1451 size visualization (***tp***)
20	0.0155 wave (*fp*)	0.5084 wire (*fp*)	0.1376 rim (*fp*)
21	~~0.0152 lg~~ (***tn***)	0.5000 facebook integration (***tp***)	0.1359 visualization tool (***tp***)
22	~~0.0148 samsung~~ (***tn***)	0.4313 personalization option (***tp***)	0.1169 app (***tp***)
23	0.0147 rubber (*fp*)	0.4043 price (*fp*)	0.1006 capture interface (***tp***)
24	~~0.0137 shot~~ (*fn*)	0.2403 rubber (*fp*)	0.0875 gsm (*fp*)
25	0.0135 lg mini (*fp*)	0.2044 motion (***tp***)	0.0852 tool (***tp***)
26	0.0122 webkit browser (***tp***)	0.1331 europe asia (*fp*)	0.0841 video review (***tp***)
27	**0.0121 fm** (***tp***)	0.1038 implementation (*fp*)	0.0819 environment (***tp***)
28	~~0.0121 appearance~~ (*fn*)	0.0975 life (*fp*)	0.0796 review (*fp*)
29	0.0120 price (*fp*)	0.0746 attendee (*fp*)	**0.0794 fm** (***tp***)
30	~~0.0117 tapping~~ (*fn*)	0.0675 touch (***tp***)	0.0729 quality (***tp***)

Table 5-23. Judgement applied over the top 30 terms of *topic 56* (without), *topic 25* (without) and *topic 45* (without) after discarding method.

In **Table 10-7** it is seen all the preselected topics, where *multimedia* initial terms are better scored, although there are more, it includes a big representation. For example, despite *topic 156,* in performance with *extra reviews* with k=200, seems to be the best which fits better the *multimedia* feature, where initial terms are lower scored than the rest of the topics. However, all the topics selected have only one term within their top 30 terms, but they performs better after applying discarding method and human judgement. Moreover, *topic 20* with *extra reviews* with k=100, which has up to 3 initial terms within the top 30 terms of the sorted topic, is a clear example of misspelling, where there are many potential terms, but

there are some mistakes on their writing and it is probably due to its lower occurrence that PLSI has scored lower and placed in this topic.

Place	*Topic 38* with *k=50*	*Topic 50* with *k=50*	*Topic 41* with *k=100*
1	0.1178 musician (***tp***)	1.0000 prominence (*fp*)	0.0586 house photo (***tp***)
2	0.1175 wm ppc phone (***tp***)	1.0000 notifier (***tp***)	0.0567 home screen skin (***tp***)
3	0.1153 s7070 review (*fp*)	1.0000 i5500 (*fp*)	0.0560 censor (***tp***)
4	0.1144 business meet (*fp*)	1.0000 proof (*fp*)	0.0540 base melody (***tp***)
5	0.1140 slider form (*fp*)	1.0000 hardware reset (*fp*)	0.0539 htc touch pro (*fp*)
6	0.1138 samsung sghl770 review (*fp*)	0.9857 partner (*fp*)	0.0539 color version (***tp***)
7	0.1137 whisper mode (***tp***)	0.9857 standard audio (***tp***)	0.0530 ac3user (*fp*)
8	0.1122 test people (*fp*)	0.9840 flow (*fp*)	0.0528 meridian player (***tp***)
9	0.1110 user feedback (*fp*)	0.9730 fingerprint (*fp*)	0.0528 touch serie (*fp*)
10	0.1093 expert (*fp*)	0.8796 tune (***tp***)	0.0525 pocket device (***tp***)
11	**0.1090 eaac** (***tp***)	0.6287 preset mode (***tp***)	0.0525 updown (*fp*)
12	0.1087 camera hold (***tp***)	0.6240 producer (*fp*)	0.0523 &raquo (*fp*)
13	0.1065 mall (*fp*)	**0.0903 streaming** (***tp***)	0.0522 baby (*fp*)
14	0.1060 microsoft office package (*fp*)	0.0600 matter (*fp*)	0.0522 rearrangement (*fp*)
15	0.1054 samsung beatb m3510 (*fp*)	0.0404 cell phone market (*fp*)	0.0522 quality image (***tp***)
16	0.1054 face recognition (***tp***)	0.0212 sound quality (***tp***)	0.0520 mp camera (***tp***)
17	0.1048 media menu house photo (***tp***)	0.0210 sidetop (*fp*)	0.0520 xscale 300mhz (*fp*)
18	0.1047 opt (*fp*)	0.0204 xenon (*fp*)	0.0519 counterfeiter (*fp*)
19	0.1044 sensory (*fp*)	0.0201 sony ericsson xperia … (*fp*)	0.0519 support site (***tp***)
20	0.1035 tweeter (***tp***)	0.0194 video preview (***tp***)	0.0518 cable carrying (***tp***)
21	0.1021 call cost (*fp*)	0.0192 platform (*fp*)	0.0516 music button (***tp***)
22	0.1019 voip service (***tp***)	0.0191 website (***tp***)	0.0515 sony ericsson w705 (*fp*)
23	0.1012 demo version (***tp***)	0.0189 even (*fp*)	0.0515 power save (*fp*)
24	0.1009 autorotation function (***tp***)	0.0184 argument (*fp*)	0.0514 unrest (*fp*)
25	0.1006 synchronisation (***tp***)	~~0.0127 carrier~~ (***tn***)	0.0513 pop gd510 video (*fp*)
26	0.1003 carlzeiss (*fp*)	0.0124 flash-light (***tp***)	0.0513 phone software (***tp***)
27	0.1003 meet memo (*fp*)	0.0116 abundance (*fp*)	0.0513 program thank (*fp*)
28	0.1002 photo flash (***tp***)	~~0.0109 tone~~ (*fn*)	0.0513 qtek (*fp*)
29	0.1000 multimedia studio ... (*fp*)	~~0.0092 film~~ (*fn*)	0.0512 headset ac adapter (*fp*)
30	0.0992 media option (***tp***)	~~0.0085 ring~~ (*fn*)	**0.0512 midi** (***tp***)

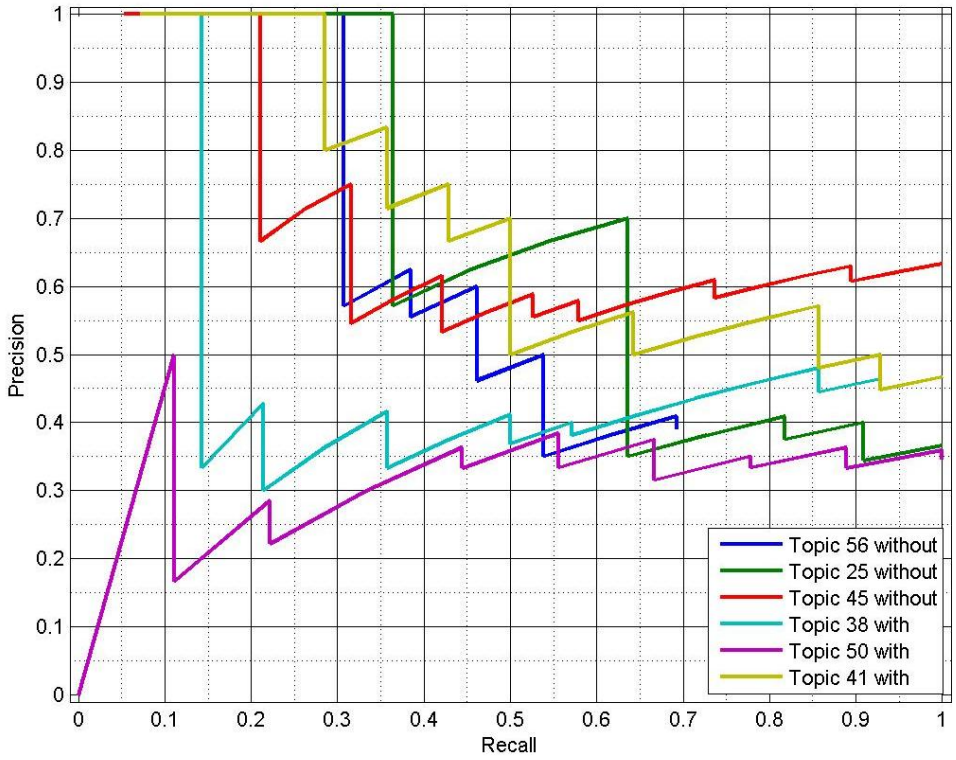
Table 5-24. Judgement applied over the top 30 terms of *topic 38* (with), *topic 56* (with) and *topic 41* (with) after discarding method.

Following the steps named before, the discarding method is applied, where only in *topic 50* and *topic 56* of the selected topics, are some terms discarded. Results of this application are located in **Annex E: Discarding method applied to PLSI topics**. In both applications there are terms wrong discarded and this fact can be seen in **Table 5-23** and **Table 5-24**, where the human judgement is applied in order to show the achievements got by PLSI on these circumnstances.

In **Graphic 5-13** can be appreciated how *topic 25* is the one which has the best beginning, althouth its ending is almost the worst, due to the wrong discarded terms. In general, all of them begin with high values of precision, except *topic 50*, whose first term is not semantically related with any of the *multimedia* feature, and this not relationship causes the poorest performance in this section. It is also remarkable that *topic 45* almost ends its performance with a 70% of precision.

To summarize the obtained results, from topics **without** *extra reviews*, like ***topic 56***, terms such as *sample picture*, *pioneer*, *transmitter* and *entry method*, are extracted with 100%

of precision and accuracy, but a 30.77% of recall. From ***topic 25***, terms such as *frontfacing*, *profile button*, *netfront* and *precision*, are extracted with a precision and accuracy of 100%, with a recall of 36.36%. From ***topic 45***, still **without** *extra reviews*, terms such as *fm transmitter advertisement*, *video share*, *question* and *wallpaper*, are extracted with 100% of precision and accuracy, but only a 21.15% of recall. Finally, from topics **with** *extra reviews*, from ***topic 38***, only terms *musician* and *wm ppc phone* are extracted with 100% of precision and accuracy, but a recall of 14.29%. However, from ***topic 41***, terms such as *house photo*, *home screen skin*, *censor* and *base melody*, are extracted with precision and accuracy of 100%, but a recall of 28.57%, concluding that PLSI perform in the same way topics with and without *extra reviews* on documents looking for the *multimedia* topic, which defines the *multimedia* product feature.



Graphic 5-13. Precision and recall graph of selected topics of *multimedia* on documents.

5.3.3.2. Sections

In this part, as it is done before, reviews are considered as groups of sections clearly delimited by titles. Then, after running PLSI on sections scenario, the table **Table 10-8** shows the most relevant topics found in its performances with each value of the *k* parameter. As it is assumed in this research, steps described in **Table 4-1** are followed and to start them, the best topics have been selected in the table. Although it seems that there are more initial terms better placed in topics that have not been selected, like it happens with topics such as *topic 39*, *topic 79*, *topic 122* and *topic 14*, with k=50,100,150,200 respectively. They have from 2 up to 4 initial terms within the top 30 terms of the sorted topic. However, they all have scores given by PLSI too low to be considered such strong relations between them and the rest of terms in the topic. Otherwise, only *topic 49*, in performance without *extra reviews* with k=100, has more

than 1 initial term in the first 30 terms, meanwhile the rest have only 1. Moreover, from selected topics, *topic 47* has only 1 term scored higher than zero, from the list of 42 initial terms of the *multimedia* product feature.

Place	*Topic 31* without *k=50*	*Topic 49* without *k=100*	*Topic 59* without *k=100*	*Topic 107* without *k=200*
1	1.0000 search button (***tp***)	0.9966 value (***tp***)	1.0000 ohm speaker (***tp***)	0.9974 viewer (***tp***)
2	1.0000 page rendering (***tp***)	0.9966 gigabyte (*fp*)	0.9968 docking (***tp***)	0.9974 equalizer preset (***tp***)
3	1.0000 equalizer preset (***tp***)	**0.9965 mpeg4** (***tp***)	0.9968 mpeg4 video (***tp***)	0.9972 google (*fp*)
4	0.9998 el (*fp*)	0.9964 equalizer preset (***tp***)	0.9964 search button (***tp***)	**0.9970 mpeg4** (***tp***)
5	0.9998 document (*fp*)	0.9964 tetris (***tp***)	0.9961 email service (*fp*)	0.9968 microphone (***tp***)
6	0.9998 chip (*fp*)	0.9960 camera button (***tp***)	0.9960 file browser (***tp***)	0.9954 letter (*fp*)
7	**0.9997 mpeg4** (***tp***)	0.9958 calculator (*fp*)	0.9959 customization (***tp***)	0.9951 wqvga (***tp***)
8	0.9994 equalizer (***tp***)	0.9956 doubletapping (***tp***)	0.9956 cable stereo headset (***tp***)	0.9941 choice (*fp*)
9	0.9992 file browser (***tp***)	0.9954 flash-indoor ... (***tp***)	0.9956 row (*fp*)	0.9939 par (*fp*)
10	0.9991 action (***tp***)	0.9951 touchwiz interface (***tp***)	0.9953 gps (*fp*)	0.9934 nothing (*fp*)
11	0.9991 screen widget (***tp***)	0.9947 touchscreen (***tp***)	0.9951 advantage (*fp*)	0.9928 rest (*fp*)
12	0.9990 camera button (***tp***)	0.9944 search box (***tp***)	**0.9949 radio** (***tp***)	0.9906 service (***tp***)
13	0.9990 email service (*fp*)	0.9941 clock (*fp*)	0.9947 community (*fp*)	0.9877 preset mode (***tp***)
14	0.9987 search box (***tp***)	0.9940 sluggishness (*fp*)	0.9944 letter (*fp*)	0.9864 sluggishness (*fp*)
15	0.9987 company (*fp*)	0.9935 bar (*fp*)	0.9940 fan (***tp***)	0.9571 player (***tp***)
16	0.9985 earpiece (***tp***)	**0.9928 music player** (***tp***)	0.9938 distinction (*fp*)	0.5000 outdoor ... (*fp*)
17	0.9984 render (***tp***)	0.9923 stopwatch (*fp*)	0.9938 pixel resolution (***tp***)	0.4878 satio sample (*fp*)
18	0.9983 recorder (***tp***)	0.9923 question (***tp***)	0.9923 facebook (***tp***)	0.4868 spinnin (*fp*)
19	0.9982 combo (*fp*)	0.9922 menus (*fp*)	0.9918 portrait (***tp***)	0.4250 ericsson aino video (***tp***)
20	0.9978 advantage (*fp*)	0.9913 google maps (*fp*)	0.9915 keypad (***tp***)	0.3561 photo album (***tp***)
21	0.9977 day (*fp*)	0.9913 par (*fp*)	0.9915 service (***tp***)	0.3066 phone stand (*fp*)
22	0.9975 gigabyte ... (*fp*)	0.9898 icon (***tp***)	0.9898 front (*fp*)	0.2772 viewfindercamera (*fp*)
23	0.9972 clock (*fp*)	0.9890 list (***tp***)	0.9861 lot (*fp*)	0.2497 headphones-usb (***tp***)
24	0.9971 vice (*fp*)	0.9870 world (*fp*)	0.9811 computer (*fp*)	0.1989 sghg800 (*fp*)
25	0.9969 nothing (*fp*)	0.9856 life (*fp*)	0.9804 call (*fp*)	0.1518 hotswap (*fp*)
26	0.9968 degree (*fp*)	0.9732 case (*fp*)	0.9774 second (***tp***)	0.1254 recognizer ... (*fp*)
27	0.9964 cam (***tp***)	**0.9705 multimedia** (***tp***)	0.9768 hand (*fp*)	0.0502 hue (*fp*)
28	0.9961 life (*fp*)	0.9676 interface (***tp***)	0.9656 picture (***tp***)	0.0402 satio (*fp*)
29	0.9959 pixel resolution (***tp***)	0.5055 home ... (*fp*)	0.9593 voice (***tp***)	0.0271 lightlow lightdark (*fp*)
30	0.9953 service (***tp***)	0.4997 bp5mnokia charger (*fp*)	0.5023 end camera phone (*fp*)	0.0108 exchange (*fp*)

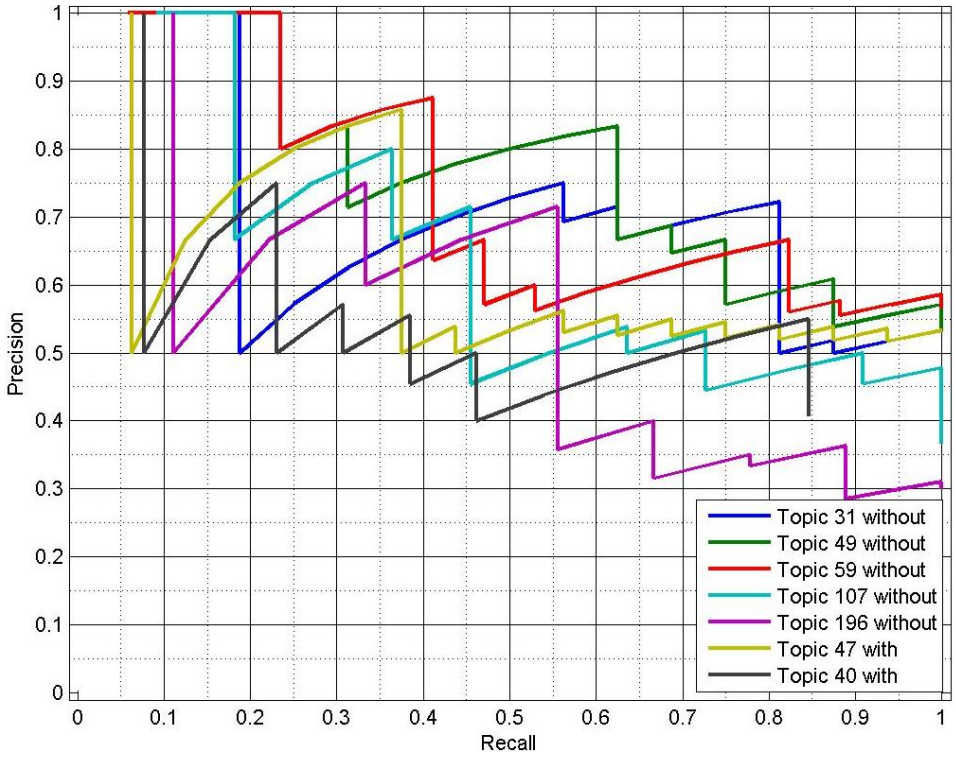
Table 5-25. Judgement applied over the top 30 terms of *topic 31* (without), *topic 49* (without), *topic 59* (without) and *topic 47* (with) after discarding method.

Once topics are selected, next step consists in the application of the discarding method, which results are placed in **Annex E: Discarding method applied to PLSI topics**, where it is possible to check that only in *topic 40,* with *extra reviews* and with k=200, are 3 terms discarded, 2 of them wrong discarded, but it means that the entire rest of terms are supposed to be strongly related between each others.

After applying discarding method, human judgement, considering the relationship between top 30 terms and any of the initial terms related to the *multimedia* feature, is applied. Its results are shown in **Table 5-25** and **Table 5-26**, from where it can be supposed a kind of performance, but following the steps, precision, accuracy and recall are calculated for all the selected topics, locating their results in **Annex H: Precision, recall and accuracy graphs of PSLI**. In this section, it is only included the precision vs. recall graphic, where it can be seen graphically how each topic has a different behaviour.

Place	*Topic 196* without *k=200*	*Topic 47* with *k=150*	*Topic 40* with *k=200*
1	0.9976 camera button (***tp***)	0.9956 flash-indoor (***tp***)	1.0000 flash-indoor sample (***tp***)
2	0.9968 tmobil usa (*fp*)	0.9907 server application (*fp*)	1.0000 piece (*fp*)
3	0.9964 rendering (***tp***)	**0.9871 mpeg4** (***tp***)	0.9997 search (***tp***)
4	0.9948 tetris (***tp***)	0.9843 microusb port (***tp***)	0.9991 color representation (***tp***)
5	0.9926 fare (*fp*)	0.9587 organizer alarm (***tp***)	0.9906 company (*fp*)
6	**0.9920 music player** (***tp***)	0.9525 portrait (***tp***)	0.9726 fire (*fp*)
7	0.9896 video recorder (***tp***)	0.8412 screen interface (***tp***)	**0.9689 video player** (***tp***)
8	0.9770 port (*fp*)	0.8357 voltage (*fp*)	0.9336 lot (*fp*)
9	0.9712 end device (*fp*)	0.8147 motorola v3 (*fp*)	0.8995 song (***tp***)
10	0.9712 senseme ... (*fp*)	0.7889 linux os (*fp*)	0.5065 bp5mnokia (*fp*)
11	0.9712 active standby (*fp*)	0.7884 v60 (*fp*)	0.5000 micro usb … (*fp*)
12	0.9712 front one (*fp*)	0.7509 htc touch dual (*fp*)	0.5000 headphone carrying (***tp***)
13	0.9712 samsung ... (*fp*)	0.7009 interference condition (***tp***)	0.4690 word fileexcel... (*fp*)
14	0.9712 beat video (*fp*)	0.6876 lifebattery (*fp*)	0.4153 quality software (*fp*)
15	0.9712 quality image (***tp***)	0.6853 photometry (***tp***)	0.2173 consumer group (*fp*)
16	0.9712 phase (*fp*)	0.6795 document folder (***tp***)	0.1854 search application (***tp***)
17	0.9712 show note (*fp*)	0.6716 ericsson w810 (*fp*)	0.1813 headset (***tp***)
18	0.9712 t600 (*fp*)	0.6677 internet link (***tp***)	0.1204 trial (***tp***)
19	0.9712 horizontal scroll (*fp*)	0.6588 mms implementation (*fp*)	0.0704 sound (***tp***)
20	0.9712 stereo earphone (***tp***)	0.6326 device control (***tp***)	0.0610 voice clarity (***tp***)
21	0.9712 itap (*fp*)	0.6316 studioimage ... (*fp*)	0.0521 fm radiofm (*fp*)
22	0.9712 repetition option (***tp***)	0.6284 tv-set (***tp***)	0.0315 tmobile usa (*fp*)
23	0.9712 wifi switch (*fp*)	0.6266 changeable (*fp*)	0.0083 generation (*fp*)
24	0.9712 pixon video (*fp*)	0.6159 data capability (***tp***)	0.0067 addition (*fp*)
25	0.9711 camera … (*fp*)	0.5938 feet jabra bt250025 (*fp*)	~~0.0061 art~~ (*fn*)
26	0.5067 artist genre... (*fp*)	0.5902 windows mobile home-screen (***tp***)	~~0.0060 introduction~~ (***tn***)
27	0.5000 sghf400 video (*fp*)	0.5860 headset ac adapter (*fp*)	0.0059 pixon (*fp*)
28	0.4962 stigma (*fp*)	0.5857 connection manager (***tp***)	0.0058 subscription (*fp*)
29	0.2500 pause key (***tp***)	0.5834 stigma (*fp*)	0.0054 map (*fp*)
30	0.1587 cookie kp500 (*fp*)	0.5791 play-pause key (***tp***)	~~0.0050 browser~~ (*fn*)

Table 5-26. Judgement applied over the top 30 terms of *topic 107* (without), *topic 196* (without) and *topic 40* (with) after discarding method.



Graphic 5-14. Precision and recall graph of selected topics of *multimedia* on sections.

In **Graphic 5-14** is viewed how *topic 59* is the single topic that remains with high precision more than 20% of recall, and it increases once again until more than 40%, while the rest goes down quickly. *Topic 59* also ends with the highest values of recall and precision, while *topic 196* starts normally, but ends with the poorest results.

Summarizing obtained results, first from topics **without** *extra reviews*, like *topic 31*, from which terms such as search button, page rendering and equalizer preset, are extracted with 100% of precision and accuracy, but a recall of 18.75%. From *topic 49*, only the term value is extracted with 100% of precision and accuracy, but only a 6.25% of recall. However, from *topic 59*, terms such as ohm speaker, docking, mpeg4 video, search button, file browser, customization and cable stereo headset, are extracted with precision and accuracy of 87.50%, and recall of 41.18%. From *topic 107*, terms like viewer and equalizer preset are extracted with 100% of precision and accuracy, but 18.18% of recall. From *topic 196*, only the term camera button is extracted with 100% of precision and accuracy, but only 11.11% of recall. From topics **with** *extra reviews*, like *topic 47*, only the term flash-indoor is extracted with 100% of precision and accuracy, but a poor recall of 6.25%, and also from *topic 40*, the single term flash-indoor sample is extracted with 100% of precision and accuracy, but recall of 7.69%. It is seen how the inclusion of *extra reviews* does not improve anything on the seeking of the *multimedia* topic, because best performances have been found without them.

5.3.3.3. Paragraphs

Finally, to end the detailed analysis of PLSI, like it is done in the rest of product features such as *battery* and *organizer*, it is run on paragraphs, where the *multimedia* topic is looked for. The process carried out is the one defined by the steps of **Table 4-1**. After running PLSI on this scenario, it is collected a group of preselected topics which may satisfy the minimum conditions to be analyzed. These topics are located in **Table 10-9**. Like it happens before, there are topics, i.e. *topic 35* or *topic 100* without *extra reviews* and with k=150, which have a single initial term within the top 30 terms of the sorted topics, but placed in the 1st or the 2nd place. As a exception, *topic 153* has been selected without having any *multimedia* initial term in the top 30, but it is considered that its performance may increase the rest ones. Otherwise, *topic 83*, in the same situation of *topic 153*, but also with more initial terms closer to the top 30 terms of the sorted topic, have not been selected due to the lower scores of these initial terms, which indicate that higher places do not have better scores to be as relevant as to be strongly related between each others.

Once topics have been selected, discarding method is run over these topics, getting some results and locating them in **Annex E: Discarding method applied to PLSI topics**. There, it is seen that how any term is discarded, then, it is supposed that every term have a strong semantic relation with the rest of the top 30 of their topic. After that, without changing terms retrieved by PLSI, the human judgement is applied in order to separate between those terms which make sense in a supposed *multimedia* topic, which defines and extends the *multimedia* product feature, and those that do not. In **Table 5-27** it is possible to se how scores given by PLSI are almost all 1.0000, this may mean that they are strongly related or it is just a coincidence.

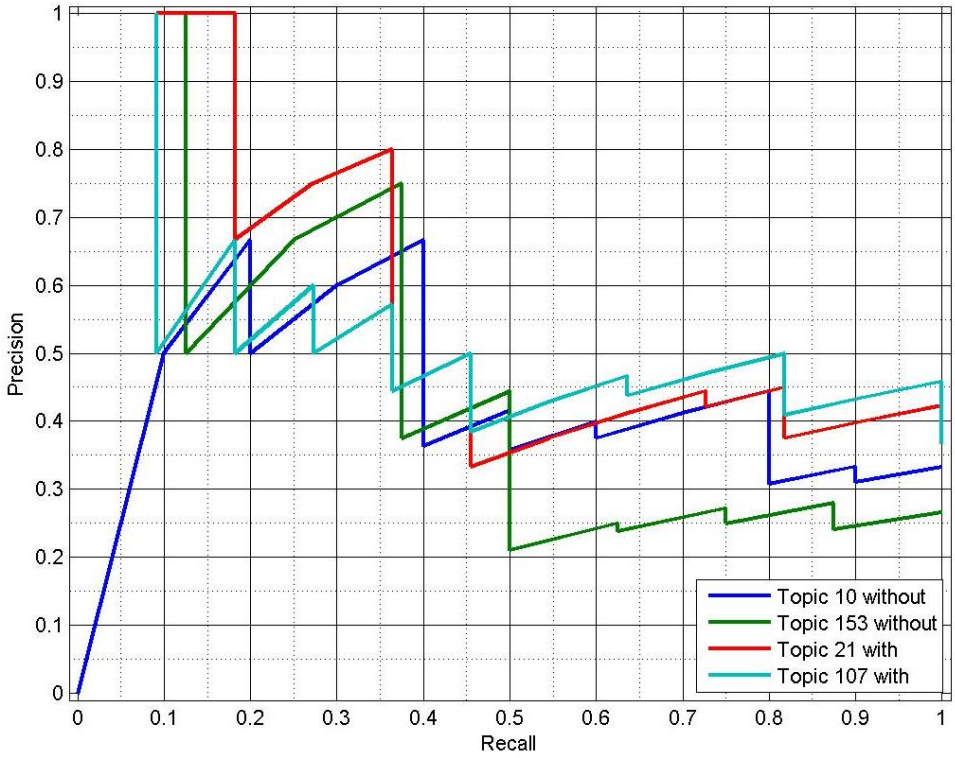
Place	*Topic 10* without *k=100*	*Topic 153* without *k=200*	*Topic 21* with *k=100*	*Topic 107* with *k=150*
1	1.0000 ip lcd display (*fp*)	1.0000 audio file (***tp***)	1.0000 msn messenger (***tp***)	1.0000 sns support (***tp***)
2	**1.0000 avi** (***tp***)	1.0000 repetition ... day (*fp*)	1.0000 xpressmusic ... (***tp***)	1.0000 chicken (*fp*)
3	1.0000 filter option (***tp***)	1.0000 memo option (***tp***)	1.0000 phone home (*fp*)	1.0000 device capture video (***tp***)
4	1.0000 device today (*fp*)	1.0000 play-pause key (***tp***)	1.0000 mb video (***tp***)	1.0000 customer support (*fp*)
5	1.0000 data cable (***tp***)	1.0000 map interface (*fp*)	1.0000 jack adapter (***tp***)	1.0000 text field (***tp***)
6	1.0000 android ui (***tp***)	1.0000 advertisement ... (*fp*)	1.0000 eten x610 (*fp*)	1.0000 armani b7620 (*fp*)
7	1.0000 navigation... (*fp*)	1.0000 delivery package (*fp*)	1.0000 ii video review (*fp*)	1.0000 gallery (***tp***)
8	1.0000 fear (*fp*)	1.0000 htc mogul (*fp*)	1.0000 navigation device (*fp*)	1.0000 mean compact (*fp*)
9	1.0000 convenience key (*fp*)	1.0000 size resolution ratio (***tp***)	1.0000 samsung s7330 (*fp*)	1.0000 assassin (*fp*)
10	1.0000 software problem (*fp*)	1.0000 phone module (*fp*)	1.0000 white balance (***tp***)	1.0000 alarm sound (***tp***)
11	1.0000 contact name (*fp*)	1.0000 studio-media file (*fp*)	1.0000 better (*fp*)	1.0000 win (*fp*)
12	1.0000 java version (***tp***)	1.0000 themer (*fp*)	1.0000 playerfm radioth (*fp*)	1.0000 satellite data update (*fp*)
13	1.0000 lend (*fp*)	0.9999 patch (*fp*)	1.0000 operating speed (*fp*)	1.0000 earphones-user (*fp*)
14	1.0000 steeper (*fp*)	0.9999 generation ipod (*fp*)	1.0000 matrix (*fp*)	1.0000 mono (***tp***)
15	1.0000 interconnection (***tp***)	0.9999 touch cruise ... (*fp*)	1.0000 haring (*fp*)	1.0000 easy keyboard (***tp***)
16	1.0000 dna (*fp*)	0.9999 end model (*fp*)	**1.0000 3gpp** (***tp***)	1.0000 mobileexcel (*fp*)
17	1.0000 mode audio player (***tp***)	0.9999 modeldimension ... (*fp*)	1.0000 internal memory (***tp***)	1.0000 game platform (***tp***)
18	1.0000 menu background (***tp***)	0.9998 status etc (*fp*)	**1.0000 m4a** (***tp***)	**1.0000 realaudio** (***tp***)
19	1.0000 wayfinder (*fp*)	0.9998 calendar source (*fp*)	1.0000 nighttime (*fp*)	1.0000 n95samsung ... (*fp*)
20	1.0000 heck (*fp*)	0.9998 video dj (***tp***)	1.0000 carl zeiss lens (***tp***)	1.0000 area code (*fp*)
21	1.0000 menuth interface (*fp*)	0.9997 playerfm radiovideo (*fp*)	1.0000 my screen (*fp*)	1.0000 missed call (*fp*)
22	1.0000 sahara (*fp*)	0.9997 dj photo (***tp***)	1.0000 sghg800 review (*fp*)	1.0000 the contact (*fp*)
23	1.0000 today we (*fp*)	0.9993 feet jabra bt25025 (*fp*)	1.0000 ericsson w880 (*fp*)	1.0000 application support (***tp***)
24	1.0000 mr (*fp*)	0.9992 quest (*fp*)	1.0000 today price (*fp*)	0.9999 arm (***tp***)
25	1.0000 lg renoir ... (*fp*)	0.9992 microphone quality (***tp***)	1.0000 micro sd card (***tp***)	0.9999 variant (*fp*)
26	1.0000 bottom right (*fp*)	0.9990 market nich (*fp*)	1.0000 supertooth light (***tp***)	0.9999 reject (*fp*)
27	1.0000 interval (***tp***)	0.9990 press release (*fp*)	1.0000 management ... (*fp*)	0.9999 send (*fp*)
28	1.0000 sprite (*fp*)	0.9988 reflex (*fp*)	1.0000 k850canon ... (*fp*)	0.9999 column (*fp*)
29	1.0000 theiphone (*fp*)	0.9986 s7550 (*fp*)	1.0000 offspring (*fp*)	0.9999 spb benchmark (*fp*)
30	1.0000 cable slot (***tp***)	0.9986 repeat mode (***tp***)	1.0000 word fileexcel ... (*fp*)	0.9998 pdf (*fp*)

Table 5-27. Judgement applied over the top 30 terms of *topic 10* (without), *topic 153* (without), *topic 21* (with) and *topic 107* (with) after discarding method.

Precision, accuracy and recall have been calculated in order to see graphically how the four topics satisfy the conditions of being the *multimedia* topic what is looked for. These graphics are located in **Annex H: Precision, recall and accuracy graphs of PSLI**, but here it is included the precision vs. recall graphic, where results are shown clrearer. In Graphic 5-15, it is seen how *topic 21* performs better than the rest, although its performance is poorer than the rest analyzed before. However, *topic 10* has the worst beginning, because of its precision does not remain high, then, any term can be extracted from it.

Summarizing results obtained in these topics, firstly, from *topic 153* **without** *extra reviews*, only the term audio file is extracted with 100% of precision and accuracy, but with only a recall of 12.50%. From *topic 21*, in performance **with** *extra reviews*, terms such as msn messenger and xpressmusic sample video, are extracted with 100% of precision and accuracy, with recall of 18.18%. Finally, from *topic 107*, also **with** *extra reviews*, only the terms sns support is extracted with 100% of precision and accuracy, but a poor recall of 9.09%. It can be concluded that all topics are far away from the *multimedia* meaning, because few terms have been extracted, and also the better performances have been with *extra reviews*. It confirms partially the supposition that PLSI performs better on documents, and sections, because they

are a reduced number of elements, and also they have more terms inside to design a distribution which models the separation of topics.



Graphic 5-15 Precision and recall graph of selected topics of *multimedia* on paragraphs.

5.4. LDA analysis

5.4.1. Battery

5.4.1.1. Documents

As it is done before with LSI and PLSI, to carry out the analysis, the LDA method is run firstly on the documents scenario, where the *battery* topic is looked for from the results given. This LDA implementation, as the PLSI one, returns two different matrices, the first one relating terms contained in documents with the topics set, and the second one relating these same topics with the documents used in the training, as it is seen in **Figure 5-5**. The most important matrix in this case is the first one, where the probability of finding a term in a topic, is stored. Then, following the steps described in **Table 4-1**, the topics which have the initial terms of the *battery* product feature, in this case, best scored, are chosen to be a potential *battery* topic. Then they are processed and finally it is seen their performance in the current context.

	Without	With
K=50	✓ **Topic 3** · 8th with 0.0123 – battery · 27th with 0.0054 – talk time · 883rd ... → 4 values set to zero	× **Topic 2** · 16th with 0.0081 – battery · 34th with 0.0048 – talk time · 621st ... → 8 values set to zero × **Topic 4** · 9th with 0.0113 – battery · 105th ... → 2 values set to zero
K=100	× **Topic 1** · 15th with 0.0453 – radio · 57th ... → 6 values set to zero × **Topic 56** · 9th with 0.0453 – radio · 80th ... → 6 values set to zero	✓ **Topic 1** · 18th with 0.0071 – battery · 28th with 0.0047 – talk time · 485th ... → 5 values set to zero ✓ **Topic 44** · 21st with 0.0081 – battery · 88th ... → 9 values set to zero
K=150	× **Topic 75** · 16th with 0.0164 – battery · X ... → ALL values set to zero × **Topic 127** · 9th with 0.0148 – battery · 187th ... → 6 values set to zero	× **Topic 4** · 4th with 0.0141 – battery · 155th ... → 6 values set to zero × **Topic 70** · 13th with 0.0086 – battery · 112th ... → 9 values set to zero ✓ **Topic 134** · 19th with 0.0107 – battery · X ... → ALL values set to zero
K=200	× **Topic 168** · 16th with 0.0081 – battery · X ... → ALL values set to zero × **Topic 170** · 14th with 0.0125 – battery · 66th ... → 6 values set to zero	× **Topic 87** · 16th with 0.0099 – battery · X ... → ALL values set to zero × **Topic 152** · 5th with 0.0198 – battery · X ... → ALLL values set to zero

Table 5-28. Best scored topics of LDA considering the *battery* product feature on documents.

First of all, topics from the whole group of performances are selected. These performances refer to the different values given to the *k* parameter, which are k=50, 100, 150 and 200. As it is seen in **Table 5-28** (located in **Annex C: LDA's running tables**), there are

several topics susceptible to be *battery* topics. Although here there is only a small representation, initial terms in LDA topics on documents scenario are better scored than in LSI (because there is any term within the top 30 terms) and in PLSI, where there are few of them, but not as good as here. Due to that, up to four topics are selected to be analyzed. It is possible to appreciate that, like in PLSI's performances on documents; the best scored initial term always is the *battery* one, because in a big scenario its strength is higher than the rest of them. Following it, *talk time* remains as the second product feature of the *battery* initial group.

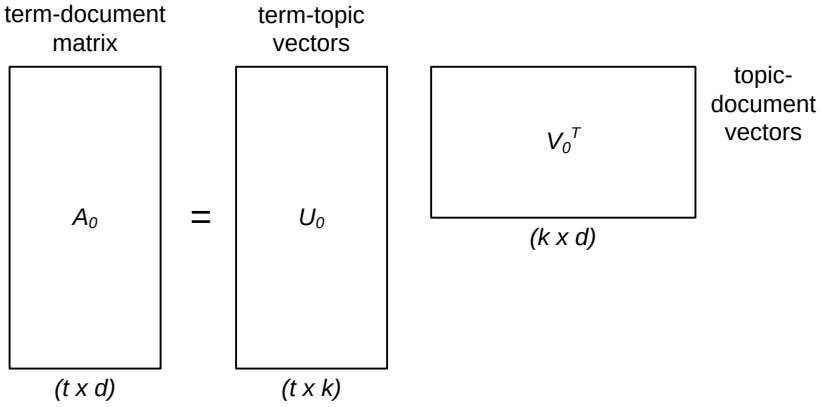
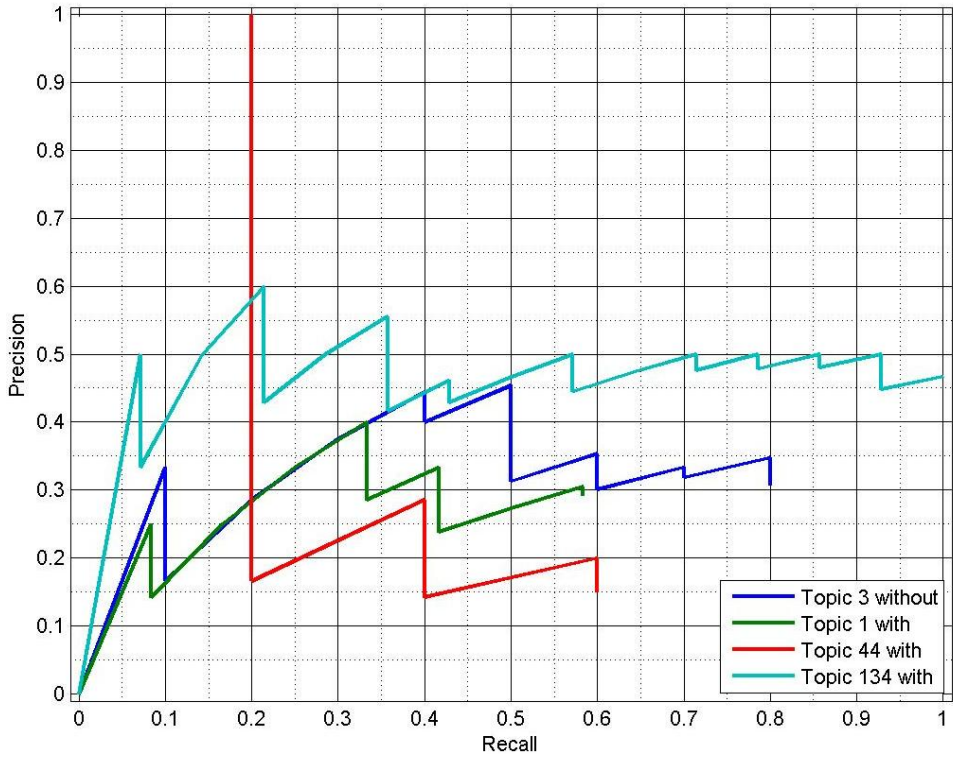


Figure 5-5.Matricial representation of the LDA term-topic and topic-document probabilities.

Place	*Topic 3* without *k=50*	*Topic 1* with *k=100*	*Topic 44* with *k=100*	*Topic 134* with *k=150*
1	0.0544 headset (*fp*)	0.0375 headset (*fp*)	0.0524 phone (***tp***)	0.2698 nokia (*fp*)
2	0.0209 ear (*fp*)	0.0151 ear (*fp*)	0.0165 number (*fp*)	0.0523 device (***tp***)
3	0.0162 time (***tp***)	0.0144 button (*fp*)	0.0162 motorola (*fp*)	0.0273 clip (*fp*)
4	0.0142 jabra (*fp*)	0.0137 time (***tp***)	0.0161 display (*fp*)	0.0205 hour (***tp***)
5	~~0.0139 device~~ (*fn*)	0.0131 jabra (*fp*)	0.0146 razr (*fp*)	0.0197 power (***tp***)
6	0.0129 design (*fp*)	~~0.0126 device~~ (*fn*)	0.0104 pc (*fp*)	0.0175 plastic (*fp*)
7	0.0126 button (*fp*)	~~0.0123 phone~~ (*fn*)	~~0.0103 key~~ (***tn***)	0.0170 module (*fp*)
8	**0.0123 battery** (***tp***)	0.0108 design (*fp*)	~~0.0103 button~~ (***tn***)	0.0167 time (***tp***)
9	0.0122 call (***tp***)	0.0101 noise (*fp*)	~~0.0101 menu~~ (***tn***)	0.0166 power button (***tp***)
10	0.0117 hour (***tp***)	~~0.0090 quality~~ (*fn*)	0.0097 call (***tp***)	0.0160 design (*fp*)
11	~~0.0113 phone~~ (*fn*)	0.0089 hour (***tp***)	0.0097 name (*fp*)	0.0152 side (*fp*)
12	0.0102 volume (*fp*)	0.0089 talk (***tp***)	~~0.0096 picture~~ (***tn***)	0.0140 way (*fp*)
13	0.0102 quality (***tp***)	0.0085 call (***tp***)	0.0095 keypad (*fp*)	0.0139 gps (***tp***)
14	0.0095 sound (*fp*)	0.0082 sound (*fp*)	0.0095 player (*fp*)	0.0134 place (*fp*)
15	0.0092 voice (*fp*)	0.0075 bluetooth (*fp*)	~~0.0095 camera~~ (***tn***)	0.0126 playback (***tp***)
16	0.0087 bluetooth (*fp*)	0.0074 voice (*fp*)	0.0095 v3i (*fp*)	0.0121 minute (***tp***)
17	0.0084 noise (*fp*)	0.0074 volume (*fp*)	0.0087 color (*fp*)	0.0116 bluetooth (*fp*)
18	~~0.0076 nokia~~ (***tn***)	**0.0071 battery** (***tp***)	0.0087 light (*fp*)	0.0113 people (*fp*)
19	0.0071 microphone (*fp*)	0.0069 gadget (*fp*)	~~0.0084 option~~ (***tn***)	**0.0107 battery** (***tp***)
20	~~0.0068 key~~ (***tn***)	0.0065 technology (*fp*)	0.0082 slvr (*fp*)	0.0106 contrast (***tp***)
21	0.0065 talk (***tp***)	0.0064 people (*fp*)	**0.0081 battery** (***tp***)	0.0100 dimension (*fp*)
22	0.0064 gadget (*fp*)	0.0060 emporia (*fp*)	~~0.0081 message~~ (*fn*)	0.0099 indicator (***tp***)
23	0.0064 technology (*fp*)	0.0058 life (*fp*)	0.0081 side (*fp*)	0.0098 mode (*fp*)
24	0.0063 people (*fp*)	0.0054 end (*fp*)	~~0.0073 quality~~ (*fn*)	0.0097 manufacturer (***tp***)
25	0.0060 size (***tp***)	0.0053 test (***tp***)	0.0069 multimedia (*fp*)	0.0087 front (*fp*)
26	0.0060 end (*fp*)	~~0.0052 size~~ (*fn*)	0.0067 memory (*fp*)	0.0086 performance (***tp***)
27	**0.0054 talk time** (***tp***)	~~0.0047 performance~~ (*fn*)	~~0.0067 sound~~ (***tn***)	0.0086 whole (*fp*)
28	0.0053 head (*fp*)	**0.0047 talk time** (***tp***)	0.0067 card (*fp*)	0.0082 back (*fp*)
29	0.0047 environment (*fp*)	~~0.0044 feature~~ (***tn***)	0.0066 melody (*fp*)	0.0081 introduction (*fp*)
30	0.0045 use (*fp*)	0.0043 use (*fp*)	~~0.0061 contact~~ (***tn***)	0.0077 test (***tp***)

Table 5-29. Judgement applied over the top 30 terms of *topic 3* (without), *topic 1* (with), *topic 44* (with) and *topic 134* (with) after discarding method.

After selecting topics from the results of LDA's performances, the discarding method is applied on them. How terms are discarded can be checked in **Annex F: Discarding method applied to LDA topics** where the respective graphics are located. It is seen that, e.g. terms like *phone* or *device* which are related with the *battery* semantic initial group, are every time discarded due to their high scores in the rest of topics, but causing a *false negative* (*fn*) which decreases the LDA's performance. Once the automatic "correct" terms are selected within the top 30 terms, human judgement is applied in order to relate terms semantically with the *battery* initial ones. Results of both applications are found in **Table 5-29**.



Graphic 5-16. Precision and recall graph of selected topics of *battery* on documents.

Although it seemed to be good LDA performances, they are better than PLSI ones, but not good enough to extract significant conclusions. The unique topic which starts and remains a little with high precision is *topic 44*. Therefore, only from ***topic 44*** in performance **with** *extra reviews*, only the term *phone* is extracted with precision of 100%, accuracy of 100% and recall of 20%.

5.4.1.2. Sections

Now, on sections it is followed the same process described in the steps of **Table 4-1**. LDA improves a lot the PLSI performance on sections, because on this scenario with PLSI there is only one topic selected to be a *battery* topic, but here, like in the rest of product features analysis, there is up to eight topics selected. It is seen that exists like a pattern in the topic

selection, because in this case *topic 3* is selected in more than one performance of the *k* parameter, with and without *extra reviews*. It does not occur like in LSI where topics are repeated as the dimensions (or topics) increase. This time it is observed such a phenomena proper of the method which tends to group similar groups of terms around the same contiguous topics, this time it is found three of the eight best scored topics of the *battery* product feature located in the *topic 3* and another one in the *topic 2*. Although there are another topics with initial terms, such as *battery*, better scored or placed in better places than the ones selected, topics which can better represent the *battery* product feature are selected from the previous selection, which can be seen in **Table 11-2** in **Annex C: LDA's running tables**. Once topics are selected, next step consist on applying the discarding method (**Annex F: Discarding method applied to LDA topics**), which like it occurs on documents, often discard terms like *phone* or *device*, causing poorer performances.

Place	*Topic 3* without *k=100*	*Topic 122* without *k=150*	*Topic 106* without *k=200*	*Topic 3* with *k=50*
1	0.0411 performance (***tp***)	0.1009 hour (***tp***)	0.1176 hour (***tp***)	0.0378 hour (***tp***)
2	0.0406 hour (***tp***)	**0.0879 battery** (***tp***)	**0.0889 battery** (***tp***)	0.0373 performance (***tp***)
3	~~0.0363 phone~~ (*fn*)	0.0711 time (***tp***)	0.0779 talk (***tp***)	0.0347 quality (***tp***)
4	0.0351 voice (*fp*)	0.0569 talk (***tp***)	0.0693 time (***tp***)	0.0339 sound (*fp*)
5	0.0351 sound (*fp*)	0.0492 performance (***tp***)	0.0498 performance (***tp***)	~~0.0319 phone~~ (*fn*)
6	0.0343 quality (***tp***)	0.0452 quality (***tp***)	0.0491 voice (*fp*)	0.0318 time (***tp***)
7	0.0334 time (***tp***)	0.0418 day (***tp***)	0.0464 quality (***tp***)	**0.0293 battery** (***tp***)
8	**0.0333 battery** (***tp***)	**0.0360 talk time** (***tp***)	0.0450 day (***tp***)	0.0289 voice (*fp*)
9	0.0214 end (*fp*)	0.0359 voice (*fp*)	**0.0429 talk time** (***tp***)	0.0193 end (*fp*)
10	0.0213 talk (***tp***)	~~0.0329 phone~~ (*fn*)	0.0355 end (*fp*)	0.0191 talk (***tp***)
11	0.0165 volume (*fp*)	0.0313 end (*fp*)	~~0.0339 phone~~ (*fn*)	0.0149 volume (*fp*)
12	0.0159 call (***tp***)	0.0251 life (*fp*)	0.0323 stand-by (***tp***)	0.0137 call (***tp***)
13	0.0150 day (***tp***)	0.0227 stand-by (***tp***)	0.0258 caller (*fp*)	0.0133 day (***tp***)
14	0.0140 test (***tp***)	0.0223 battery life (***tp***)	0.0247 manufacturer (***tp***)	0.0130 test (***tp***)
15	**0.0128 talk time** (***tp***)	0.0203 call (***tp***)	~~0.0189 call~~ (*fn*)	~~0.0120 device~~ (*fn*)
16	0.0110 noise (*fp*)	0.0198 caller (*fp*)	~~0.0132 device~~ (*fn*)	**0.0114 talk time** (***tp***)
17	~~0.0106 device~~ (*fn*)	0.0188 test (***tp***)	~~0.0125 sound~~ (***tn***)	0.0107 noise (*fp*)
18	0.0103 conversation (*fp*)	0.0145 minute (***tp***)	0.0114 people (*fp*)	0.0095 conversation (*fp*)
19	0.0091 signal (*fp*)	~~0.0112 volume~~ (***tn***)	0.0107 test (***tp***)	0.0080 music (*fp*)
20	0.0087 caller (*fp*)	0.0111 signal (*fp*)	~~0.0101 mode~~ (***tn***)	0.0076 manufacturer (***tp***)
21	0.0083 stand-by (***tp***)	0.0101 network (***tp***)	~~0.0094 line~~ (***tn***)	0.0076 people (*fp*)
22	0.0079 result (*fp*)	~~0.0100 handset~~ (*fn*)	0.0089 playback (***tp***)	0.0075 caller (*fp*)
23	0.0070 life (*fp*)	0.0092 charger (***tp***)	~~0.0083 problem~~ (***tn***)	0.0074 signal (*fp*)
24	0.0068 battery life (***tp***)	0.0082 line (*fp*)	~~0.0079 volume~~ (***tn***)	0.0072 life (*fp*)
25	0.0067 reception (*fp*)	0.0080 area (*fp*)	~~0.0069 bit~~ (***tn***)	0.0071 stand-by (***tp***)
26	0.0066 environment (*fp*)	0.0078 speaker (*fp*)	0.0066 loudspeaker (*fp*)	0.0069 line (*fp*)
27	0.0066 line (*fp*)	0.0073 reception (*fp*)	~~0.0063 noise~~ (***tn***)	0.0069 result (*fp*)
28	~~0.0063 manufacturer~~ (*fn*)	0.0064 usage (***tp***)	0.0063 charger (***tp***)	0.0066 environment (*fp*)
29	0.0061 bit (*fp*)	~~0.0061 problem~~ (***tn***)	~~0.0061 minute~~ (*fn*)	0.0060 bit (*fp*)
30	~~0.0060 problem~~ (***tn***)	0.0060 use (*fp*)	0.0060 average (*fp*)	0.0060 problem (*fp*)

Table 5-30. Judgement applied over the top 30 terms of *topic 3* (without), *topic 122* (without), *topic 106* (without) and *topic 3* (with) after discarding method.

To extract some conclusions from the topics selected after applying the discarding method, the human judgement is carried out on the top 30 terms of each selected topic, deciding which of them are correct discarded or not discarded, and correct returned or not. Results of this application are placed in **Table 5-30** and **Table 5-31**.

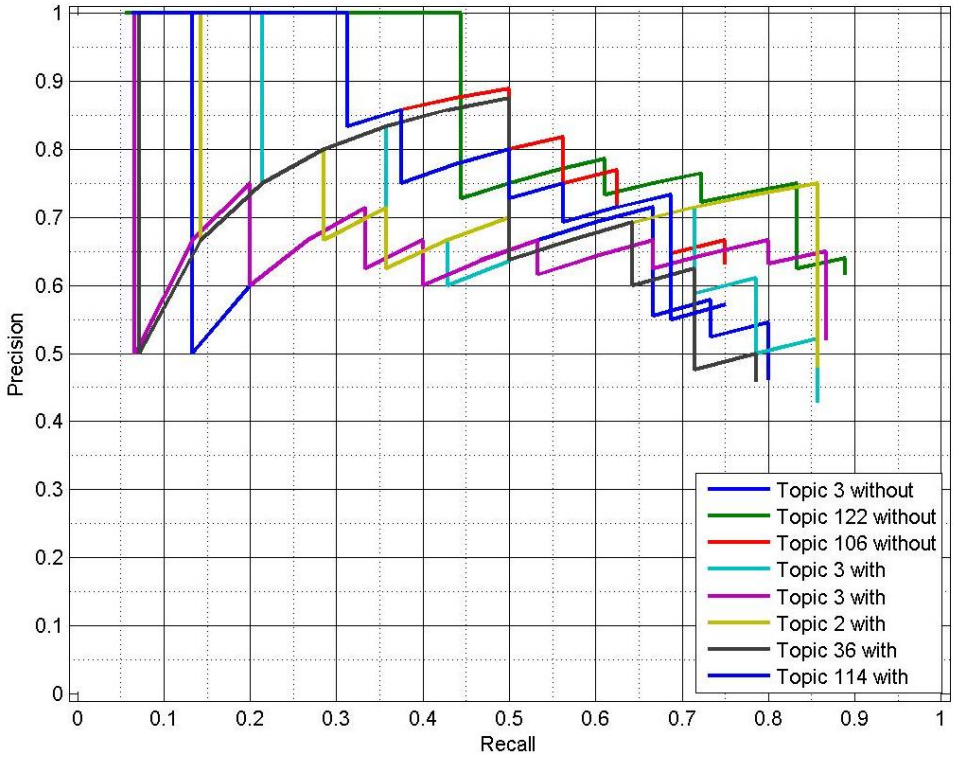
As it is explained above, it occurs that when an initial term like *battery* in *topic 36* with *extra reviews* and k=150 is scored with a high probability and placed at the top of the sorted topic, the topic does not correspond to expectations, because its performance is poorer than the rest. In **Annex I: Precision, recall and accuracy graphs of LDA** it is possible to see how the only term that remains within the 80% of the threshold set is this one, *battery*. Then it is concluded that any term is extracted from *topic 36*. Otherwise, from the rest of topics, more or less new semantically related terms with the *battery* product feature are extracted and summarized below.

Place	*Topic 3* with *k=100*	*Topic 2* with *k=150*	*Topic 36* with *k=150*	*Topic 114* with *k=200*
1	0.0397 hour (***tp***)	0.0435 performance (***tp***)	**0.0602 battery** (***tp***)	0.1001 hour (***tp***)
2	0.0378 sound (*fp*)	0.0435 hour (***tp***)	0.0523 life (*fp*)	0.0718 time (***tp***)
3	0.0368 performance (***tp***)	0.0399 voice (*fp*)	0.0427 battery life (***tp***)	0.0627 performance (***tp***)
4	0.0339 quality (***tp***)	0.0363 quality (***tp***)	0.0366 time (***tp***)	**0.0600 battery** (***tp***)
5	0.0317 voice (*fp*)	0.0363 time (***tp***)	0.0310 handset (***tp***)	0.0581 talk (***tp***)
6	~~0.0291 phone~~ (*fn*)	0.0360 sound (*fp*)	0.0267 day (***tp***)	0.0570 voice (*fp*)
7	0.0290 time (***tp***)	~~0.0360 phone~~ (*fn*)	0.0256 quality (***tp***)	0.0474 quality (***tp***)
8	**0.0249 battery** (***tp***)	**0.0288 battery** (***tp***)	~~0.0250 phone~~ (*fn*)	0.0449 end (*fp*)
9	0.0228 end (*fp*)	0.0258 end (*fp*)	0.0243 hour (***tp***)	~~0.0416 phone~~ (*fn*)
10	0.0199 talk (***tp***)	0.0247 talk (***tp***)	0.0230 signal (*fp*)	**0.0329 talk time** (***tp***)
11	0.0174 volume (*fp*)	0.0173 call (***tp***)	0.0171 area (*fp*)	0.0266 day (***tp***)
12	0.0138 test (***tp***)	0.0160 volume (*fp*)	0.0162 recept (*fp*)	0.0210 sound (*fp*)
13	0.0129 call (***tp***)	0.0153 test (***tp***)	0.0161 performance (***tp***)	0.0203 stand-by (***tp***)
14	0.0127 noise (*fp*)	**0.0140 talk time** (***tp***)	0.0159 network (***tp***)	~~0.0200 call~~ (*fn*)
15	**0.0126 talk time** (***tp***)	~~0.0139 device~~ (*fn*)	0.0156 speaker (*fp*)	0.0188 caller (*fp*)
16	~~0.0124 device~~ (*fn*)	0.0122 day (***tp***)	0.0131 issue (*fp*)	~~0.0187 device~~ (*fn*)
17	0.0113 day (***tp***)	0.0095 manufacturer (***tp***)	~~0.0128 call~~ (*fn*)	0.0186 manufacturer (***tp***)
18	0.0112 conversation (*fp*)	0.0093 stand-by (***tp***)	0.0124 charger (***tp***)	0.0178 test (***tp***)
19	0.0084 caller (*fp*)	0.0091 result (*fp*)	0.0118 blast (*fp*)	0.0177 line (*fp*)
20	~~0.0082 headset~~ (***tn***)	0.0088 conversation (*fp*)	~~0.0113 volume~~ (***tn***)	~~0.0126 volume~~ (***tn***)
21	0.0080 stand-by (***tp***)	0.0084 caller (*fp*)	0.0105 wing (*fp*)	0.0125 problem (*fp*)
22	0.0078 result (*fp*)	0.0083 line (*fp*)	0.0105 bar (*fp*)	0.0122 bit (*fp*)
23	0.0076 manufacturer (***tp***)	0.0071 bit (*fp*)	0.0091 signal strength (*fp*)	0.0106 people (*fp*)
24	0.0072 environment (*fp*)	~~0.0071 problem~~ (***tn***)	0.0090 ii (*fp*)	~~0.0092 mode~~ (***tn***)
25	0.0071 people (*fp*)	0.0070 noise (*fp*)	0.0086 minute (***tp***)	~~0.0078 minute~~ (*fn*)
26	~~0.0071 music~~ (***tn***)	~~0.0066 people~~ (***tn***)	0.0082 end (*fp*)	~~0.0069 noise~~ (***tn***)
27	0.0068 line (*fp*)	0.0062 average (*fp*)	0.0077 strength (*fp*)	0.0068 lag (*fp*)
28	0.0063 bit (*fp*)	0.0060 signal (*fp*)	~~0.0074 voice~~ (***tn***)	~~0.0068 result~~ (***tn***)
29	~~0.0061 problem~~ (***tn***)	~~0.0059 level~~ (***tn***)	~~0.0072 setting~~ (***tn***)	0.0065 mah (***tp***)
30	0.0053 average (*fp*)	0.0057 loudspeaker (*fp*)	~~0.0067 test~~ (*fn*)	~~0.0064 environment~~ (***tn***)

Table 5-31. Judgement applied over the top 30 terms of *topic 3* (with), *topic 36* (with), *topic 2* (with) and *topic 114* (with) after discarding method.

From ***topic 3*** **without** *extra reviews*, terms such as *performance* and *hour* are extracted with 100% of precision and accuracy, but 13.33% of recall. From ***topic 122*** **without** *extra reviews*, terms such as *hour*, *time*, *talk*, *performance*, *quality* and *day* are extracted with 100% of precision and accuracy and 44.44% of recall. From ***topic 106*** **without** *extra reviews*, terms such as *hour*, *talk*, *time*, *performance*, *quality*, *day* and *stand-by* are extracted with 81.82% of precision, 75% of accuracy and 56.25% of recall. From ***topic 3*** **with** *extra reviews* and **k=50**, terms such as *performance*, *quality* and *day* are extracted with 100% of precision and accuracy and 21.43% of recall. From ***topic 3*** **with** *extra reviews* and **k=100**, only the term *hour* is extracted with 100% of precision and accuracy and 6.67% of recall. From ***topic 2*** **with** *extra reviews*, terms such as *performance* and *hour* are extracted with 100% of precision and

accuracy, but 14.29% of recall. From ***topic 114*** **with** *extra reviews* too, terms such as *hour*, *time*, *performance*, *talk*, *quality* and *day* are extracted with 85.71% of precision and accuracy and 37.50% of recall. Summarizing these results it is realized that terms extracted from different terms on different performances converge to the same end group, which means that LDA achieve the best results and also they are correlated between each others.



Graphic 5-17. Precision and recall graph of selected topics of *battery* on sections.

5.4.1.3. Paragraphs

Finally, *battery* topics are looked for on the paragraphs scenario, where the same steps described in **Table 4-1** are followed. In **Table 11-3** it is found a set of topics from the LDA's performances with the different values of *k* parameter. Many differences are found between LDA's performances and the LSI and PLSI ones, because where these last ones almost found anything, from LDA's performances up to eight topics are selected to be potential *battery* topics. The initial term that appears most is the *battery* one, followed by the *talk time* one, but LDA also includes the *capacity* one within some of the top 30 terms of the sorted topics.

Once the topics are selected, the discarding method (**Annex F: Discarding method applied to LDA topics**) is applied, and following it, the human judgement, is set on the top 30 terms of the selected topics. Results of both applications are seen in **Table 5-32** and **Table 5-33**, where is it possible to how *topic 62*, instead of discarding more terms to discard wrong retrieved terms, almost all remain giving a poorer performance. However, in *topic 43*, although there are many discarded terms, in general more of the fifty percent are good discarded, but it decreases its potential new terms. The poorest performance is found on *topic 29*, where the four discarded terms are wrong removed and there are some early *false positive* that make its

precision go down quickly. The entire collection of graphics extracted from these performances is located in **Annex I: Precision, recall and accuracy graphs of LDA**, but here it is only showed the precision vs. recall graphic, **Graphic 5-18**, to see how they perform in comparation with each other and between the rest of method's performances. A summarization of the semantic related terms and their respective topics is explained below.

Place	*Topic 26* without *k=50*	*Topic 29* without *k=100*	*Topic 43* without *k=100*	*Topic 62* without *k=100*
1	0.1148 hour (***tp***)	**0.2064 battery** (***tp***)	0.2055 hour (***tp***)	0.1005 cable (***tp***)
2	**0.1138 battery** (***tp***)	0.1169 life (*fp*)	0.1268 time (***tp***)	0.0776 usb (*fp*)
3	0.0836 time (***tp***)	0.0688 battery life (***tp***)	0.0983 talk (***tp***)	0.0747 charger (***tp***)
4	0.0826 day (***tp***)	0.0505 day (***tp***)	**0.0703 battery** (***tp***)	0.0485 usb cable (*fp*)
5	0.0575 talk (***tp***)	0.0472 charger (***tp***)	**0.0606 talk time** (***tp***)	0.0440 user (*fp*)
6	0.0426 life (*fp*)	0.0399 usage (***tp***)	0.0546 day (***tp***)	0.0421 box (***tp***)
7	**0.0332 talk time** (***tp***)	~~0.0269 phone~~ (*fn*)	0.0434 stand-by (***tp***)	0.0407 guide (*fp*)
8	0.0290 minute (***tp***)	0.0217 mah (***tp***)	0.0380 minute (***tp***)	0.0394 adapter (***tp***)
9	0.0252 battery life (***tp***)	0.0197 use (*fp*)	0.0317 manufacturer (***tp***)	0.0323 headset (*fp*)
10	~~0.0251 phone~~ (*fn*)	0.0181 stand (*fp*)	~~0.0240 device~~ (*fn*)	0.0290 cd (*fp*)
11	0.0225 stand-by (***tp***)	0.0158 power (***tp***)	~~0.0206 phone~~ (*fn*)	0.0243 manual (*fp*)
12	0.0190 manufacturer (***tp***)	0.0137 handset (***tp***)	0.0175 test (***tp***)	0.0221 card (*fp*)
13	~~0.0184 device~~ (*fn*)	~~0.0131 time~~ (*fn*)	0.0171 mode (*fp*)	0.0218 user guide (*fp*)
14	0.0181 test (***tp***)	0.0129 performance (***tp***)	0.0155 data (*fp*)	0.0213 software (*fp*)
15	0.0132 charger (***tp***)	0.0110 test (***tp***)	0.0120 result (*fp*)	0.0194 package (*fp*)
16	0.0128 performance (***tp***)	**0.0108 capacity** (***tp***)	~~0.0113 nokia~~ (***tn***)	0.0187 wall (*fp*)
17	0.0127 mode (*fp*)	~~0.0095 call~~ (*fn*)	0.0099 half (*fp*)	0.0169 headphone (*fp*)
18	0.0115 usage (***tp***)	0.0086 shame (*fp*)	0.0082 performance (***tp***)	0.0166 case (*fp*)
19	0.0099 data (*fp*)	0.0080 mah battery (***tp***)	~~0.0080 call~~ (*fn*)	0.0165 stereo (*fp*)
20	0.0085 result (*fp*)	0.0080 battery performance (***tp***)	0.0065 maximum (*fp*)	0.0160 wall charger (***tp***)
21	0.0082 mah (***tp***)	0.0075 wifi (***tp***)	0.0061 plan (*fp*)	~~0.0154 nokia~~ (***tn***)
22	0.0079 half (*fp*)	0.0071 battery time (***tp***)	0.0060 network (***tp***)	0.0140 microsd (*fp*)
23	0.0077 use (*fp*)	0.0071 brightness (***tp***)	0.0058 rate (*fp*)	0.0134 microsd card (*fp*)
24	~~0.0064 call~~ (*fn*)	0.0062 condition (*fp*)	~~0.0050 internet~~ (***tn***)	0.0114 bluetooth (*fp*)
25	~~0.0060 handset~~ (*fn*)	0.0061 charger (***tp***)	~~0.0050 music~~ (***tn***)	**0.0114 battery** (***tp***)
26	0.0048 power (***tp***)	0.0057 minute (***tp***)	~~0.0050 lot~~ (***tn***)	0.0109 stereo headset (*fp*)
27	0.0046 specification (*fp*)	0.0052 work (*fp*)	0.0045 talktime (***tp***)	0.0104 pouch (*fp*)
28	0.0043 wifi (***tp***)	0.0052 indicator (***tp***)	0.0041 playback (***tp***)	~~0.0101 connection~~ (***tn***)
29	0.0042 talktime (***tp***)	0.0049 energy (***tp***)	~~0.0038 case~~ (***tn***)	0.0100 computer (*fp*)
30	0.0040 rate (*fp*)	~~0.0045 smartphone~~ (***tn***)	~~0.0036 fact~~ (***tn***)	0.0095 accessory (***tp***)

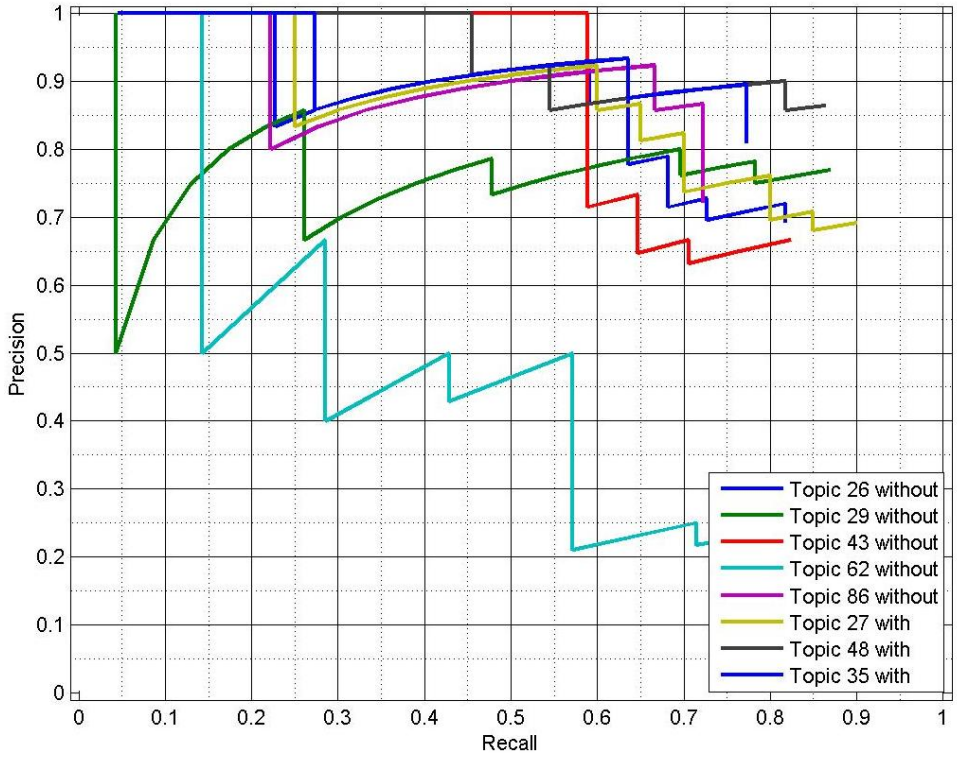
Table 5-32. Judgement applied over the top 30 terms of *topic 26* (without), *topic 29* (without), *topic 43* (without) and *topic 62* (without) after discarding method.

From ***topic 26* without** *extra reviews*, terms such as *hour*, *time*, *day*, *talk*, *minute*, *battery life*, *stand-by*, *manufacturer*, *test*, *charger*, *performance* and *usage* are extracted with 87.50% of precision, 77.78% of accuracy and 63.64% of recall. From ***topic 43* without** *extra reviews*, terms such as *hour*, *time*, *talk*, *day*, *stand-by*, *minute*, *manufacturer* and *test* are extracted with 100% of precision, 83.33% of accuracy and 58.82% of recall. From ***topic 62* without** *extra reviews*, only the term *cable* is extracted with 100% of precision and accuracy, but only a 14.29% of recall. From ***topic 86* without** *extra reviews*, terms such as *hour*, *time*, *talk*, *day*, *battery life*, *minute*, *stand-by*, *test*, *manufacturer*, *mah* and *usage* are extracted with 86.67% of precision, 76.19% of accuracy and 72.22% of recall. From ***topic 27* with** *extra reviews*, terms such as *hour*, *time*, *day*, *talk*, *minute*, *battery life*, *stand-by*, *test*, *manufacturer*, *charger*, *performance* and *usage* are extracted with 82.35% of precision, 73.68% of accuracy and 70% of recall. From ***topic 48* without** *extra reviews*, terms such as *hour*, *time*, *talk*, *day*, *stand-by*, *manufacturer*, *minute*, *test*, *mah*, *performance*, *charger*, *network*, *battery performance*, *playback*, *talktime* and *mah battery* are extracted with 86.36% of precision, 82.14% of accuracy

and 86.36% of recall. From ***topic 35* without** *extra reviews*, terms such as *hour*, *time*, *talk*, *day*, *battery life*, *stand-by*, *minute*, *charger*, *test*, *performance*, *usage*, *manufacturer*, *mah*, *playback* and *min* are extracted with 89.47% of precision, 76.92% of accuracy and 77.27% of recall.

Place	*Topic 86* without *k=150*	*Topic 27* with *k=50*	*Topic 48* with *k=100*	*Topic 35* with *k=150*
1	0.1563 hour (***tp***)	0.1175 hour (***tp***)	0.1927 hour (***tp***)	0.1099 hour (***tp***)
2	0.1170 time (***tp***)	**0.1149 battery** (***tp***)	0.1149 time (***tp***)	**0.0883 battery** (***tp***)
3	**0.1143 battery** (***tp***)	0.0795 time (***tp***)	0.0876 talk (***tp***)	0.0807 time (***tp***)
4	0.0702 talk (***tp***)	0.0795 day (***tp***)	**0.0859 battery** (***tp***)	0.0515 talk (***tp***)
5	0.0580 life (*fp*)	0.0574 talk (***tp***)	**0.0607 talk time** (***tp***)	0.0405 day (***tp***)
6	0.0547 day (***tp***)	0.0405 life (*fp*)	0.0504 day (***tp***)	**0.0399 talk time** (***tp***)
7	**0.0537 talk time** (***tp***)	**0.0332 talk time** (***tp***)	0.0454 stand-by (***tp***)	0.0397 life (*fp*)
8	0.0440 battery life (***tp***)	0.0252 minute (***tp***)	0.0289 manufacturer (***tp***)	0.0328 battery life (***tp***)
9	0.0307 minute (***tp***)	0.0248 battery life (***tp***)	0.0254 minute (***tp***)	~~0.0262 phone~~ (*fn*)
10	0.0255 stand-by (***tp***)	~~0.0230 phone~~ (*fn*)	~~0.0228 phone~~ (***tn***)	0.0213 stand-by (***tp***)
11	~~0.0222 phone~~ (*fn*)	0.0210 stand-by (***tp***)	~~0.0215 device~~ (*fn*)	0.0205 minute (***tp***)
12	0.0193 test (***tp***)	0.0182 test (***tp***)	0.0193 test (***tp***)	~~0.0172 device~~ (*fn*)
13	0.0189 manufacturer (***tp***)	0.0174 manufacturer (***tp***)	0.0144 mode (*fp*)	0.0154 charger (***tp***)
14	~~0.0182 device~~ (*fn*)	~~0.0164 device~~ (*fn*)	0.0135 mah (***tp***)	0.0138 test (***tp***)
15	~~0.0126 quality~~ (*fn*)	0.0139 charger (***tp***)	0.0134 performance (***tp***)	0.0115 performance (***tp***)
16	0.0102 mah (***tp***)	0.0111 mode (*fp*)	0.0123 data (*fp*)	0.0113 usage (***tp***)
17	~~0.0093 nokia~~ (***tn***)	0.0111 performance (***tp***)	~~0.0103 nokia~~ (***tn***)	0.0109 manufacturer (***tp***)
18	~~0.0090 voice~~ (***tn***)	0.0100 data (*fp*)	0.0096 charger (***tp***)	~~0.0109 volume~~ (***tn***)
19	~~0.0088 mode~~ (***tn***)	0.0091 usage (***tp***)	~~0.0062 handset~~ (*fn*)	~~0.0105 quality~~ (*fn*)
20	0.0088 specification (*fp*)	0.0089 use (*fp*)	0.0062 network (***tp***)	~~0.0097 call~~ (*fn*)
21	0.0082 usage (***tp***)	0.0087 result (*fp*)	0.0055 battery performance (***tp***)	~~0.0082 mode~~ (***tn***)
22	0.0075 party (*fp*)	0.0082 mah (***tp***)	**0.0055 capacity** (***tp***)	0.0081 end (*fp*)
23	~~0.0074 people~~ (***tn***)	~~0.0069 nokia~~ (***tn***)	0.0055 playback (***tp***)	0.0077 mah (***tp***)
24	0.0074 result (*fp*)	0.0065 power (***tp***)	~~0.0046 lot~~ (***tn***)	~~0.0069 music~~ (***tn***)
25	~~0.0073 call~~ (*fn*)	0.0053 specification (*fp*)	0.0046 talktime (***tp***)	0.0062 playback (***tp***)
26	~~0.0070 use~~ (***tn***)	0.0047 achievement (*fp*)	~~0.0043 smartphone~~ (***tn***)	0.0061 min (***tp***)
27	~~0.0070 volume~~ (***tn***)	0.0046 wifi (***tp***)	0.0040 rate (*fp*)	0.0061 result (*fp*)
28	~~0.0067 end~~ (***tn***)	~~0.0045 music~~ (***tn***)	0.0039 mah battery (***tp***)	~~0.0058 network~~ (*fn*)
29	~~0.0058 power~~ (*fn*)	0.0043 half (*fp*)	~~0.0036 usage~~ (*fn*)	0.0058 specification (*fp*)
30	0.0046 course (*fp*)	0.0042 talktime (***tp***)	~~0.0034 internet~~ (***tn***)	~~0.0056 party~~ (***tn***)

Table 5-33. Judgement applied over the top 30 terms of *topic 86* (without), *topic 27* (with), *topic 48* (with) and *topic 35* (with) after discarding method.



Graphic 5-18. Precision and recall graph of selected topics of *battery* on paragraphs.

5.4.2. Organizer

5.4.2.1. Documents

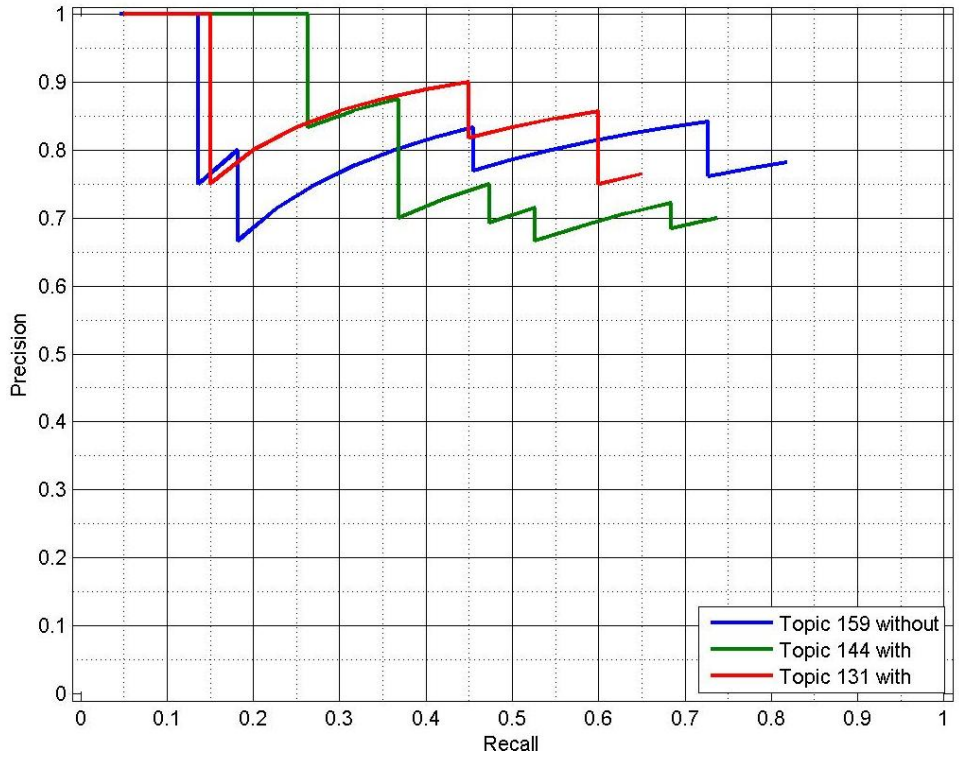
As the analysis is proceeded with the *battery* product feature, the same process described in **Table 4-1** is followed in this point of the project, but now with the *organizer* product feature, which has more initial terms and they have improved their performances in methods like LSI and PLSI. Then in this chapter it is cover the behaviour of the initial terms on the documents scenario firstly.

Looking into **Table 11-4** it is seen that there are fewer topics selected than in PLSI performance on documents, but the ones selected seems to have stronger relation with the *organizer* initial product features. Here, there are only three topics to be analyzed as the potential *organizer* topics to extract from them new semantically related terms which increase the initial semantic group. However, *topic 159* has up to 8 initial terms, of the list of 16, within the top 30 terms.

Place	*Topic 159* without *k=200*	*Topic 144* with *k=150*	*Topic 131* with *k=200*
1	0.0322 time (***tp***)	0.0732 phone (***tp***)	0.0620 phone (***tp***)
2	0.0256 option (***tp***)	0.0300 message (***tp***)	0.0250 message (***tp***)
3	0.0242 clock (***tp***)	0.0224 picture (***tp***)	0.0160 list (***tp***)
4	0.0212 world (*fp*)	0.0211 number (***tp***)	~~0.0157 samsung~~ (***tn***)
5	0.0206 phonebook (***tp***)	0.0198 menu (***tp***)	~~0.0140 time~~ (*fn*)
6	0.0201 music (*fp*)	~~0.0185 samsung~~ (***tn***)	~~0.0136 option~~ (*fn*)
7	~~0.0197 screen~~ (***tn***)	0.0178 display (*fp*)	~~0.0134 picture~~ (*fn*)
8	**0.0188 calendar** (***tp***)	0.0178 text (***tp***)	~~0.0125 camera~~ (***tn***)
9	~~0.0187 picture~~ (*fn*)	0.0178 time (***tp***)	0.0125 thing (*fp*)
10	0.0187 message (***tp***)	0.0154 feature (*fp*)	~~0.0119 feature~~ (***tn***)
11	0**.0172 alarm** (***tp***)	~~0.0131 key~~ (***tn***)	~~0.0119 menu~~ (*fn*)
12	**0.0166 world clock** (***tp***)	0.0125 light (*fp*)	~~0.0116 button~~ (***tn***)
13	**0.0165 converter** (***tp***)	0.0123 ring (***tp***)	0.0116 sound (***tp***)
14	**0.0161 organizer** (***tp***)	~~0.0116 screen~~ (***tn***)	0.0107 phonebook (***tp***)
15	0.0150 one (*fp*)	0.0113 color (***tp***)	0.0107 voice (***tp***)
16	0.0132 day (***tp***)	~~0.0113 contact~~ (*fn*)	**0.0100 note** (***tp***)
17	0.0130 resolution (***tp***)	0.0110 call (*fp*)	~~0.0098 screen~~ (***tn***)
18	**0.0121 task** (***tp***)	**0.0104 alarm** (***tp***)	0.0096 setting (***tp***)
19	0.0116 end (***tp***)	0.0098 game (*fp*)	0.0087 voice note (***tp***)
20	~~0.0115 camera~~ (***tn***)	~~0.0093 motorola~~ (***tn***)	0.0087 way (*fp*)
21	**0.0115 calculator** (***tp***)	~~0.0091 camera~~ (***tn***)	**0.0085 calendar** (***tp***)
22	~~0.0114 display~~ (***tn***)	0.0091 service (***tp***)	0.0085 record (***tp***)
23	~~0.0113 phone~~ (*fn*)	0.0090 text message (***tp***)	0.0078 entry (***tp***)
24	~~0.0111 image~~ (*fn*)	~~0.0088 sound~~ (*fn*)	~~0.0076 number~~ (*fn*)
25	**0.0111 note** (***tp***)	~~0.0088 voice~~ (*fn*)	~~0.0075 text~~ (*fn*)
26	0.0108 lot (*fp*)	0.0079 minute (***tp***)	~~0.0073 mode~~ (*fn*)
27	0.0108 music player (*fp*)	~~0.0076 email~~ (*fn*)	0.0072 t809 (*fp*)
28	0.0105 voice (***tp***)	0.0074 mms (*fp*)	~~0.0067 slider~~ (***tn***)
29	0.0105 text (***tp***)	0.0071 style (***tp***)	0.0067 cell (*fp*)
30	~~0.0095 memory~~ (*fn*)	~~0.0067 option~~ (*fn*)	0.0066 picture message (***tp***)

Table 5-34. Judgement applied over the top 30 terms of *topic 159* (without), *topic 144* (with) and *topic 131* (with) after discarding method.

Once topics are selected, it is possible to apply the discarding method in order to remove those terms which have not a strong semantic relation with the rest of the top 30 terms of the sorted topic. Results of this application are located in **Annex F: Discarding method applied to LDA topics**. Then, human judgement is applied on results of the discarding method, to set how remaining terms are semantically related with the *organizer* ones. In **Table 5-34** it is seen how in *topic 159*, instead of having 8 initial terms within its top 30 terms, there are many unexpected results *(fp*) and missing results (*fn*) which cause a poor performance.



Graphic 5-19. Precision and recall graph of selected topics of *organizer* on documents.

Finally, to summarize the entire topic's performance, they are showed graphically in a precision vs. recall graphic, in **Graphic 5-19**, which shows how the only topic which overcomes the 80% of recall at the end of the top 30 terms is the named *topic 159*, in spite of having a good performance, because its beginning is no good at it was supposed due to its high content of initial terms. Rest of graphics are located in **Annex I: Precision, recall and accuracy graphs of LDA**.

From ***topic 159* without** *extra reviews*, terms such as *time*, *option* and *clock* are extracted with 100% of precision and accuracy, but a 13.64% of recall. From ***topic 144* with** *extra reviews*, terms such as *phone*, *message*, *picture*, *number*, *menu*, *text* and *time* are extracted with 87.50% of precision and accuracy, and 36.84% of recall. From ***topic 131* with** *extra reviews*, terms such as *phone*, *message* and *list* are extracted with 100% of precision and accuracy, but a 15% of recall.

5.4.2.2. Sections

Organizer product feature is analyzed on LDA's performances, now on sections. Although LSI's performances overcome the PLSI's performance on sections, LDA improves their results, at least getting a high number of *organizer* initial product features within the top 30 values. This does not mean that LDA's performances have better precision, accuracy or recall, because they are checked in this chapter, but without applying any action over the data, it is showed to be stronger related than the rest analyzed before.

Then, following the steps described in **Table 4-1**, the best topics from the wider selection made in **Table 11-5** are selected and analyzed in this section. They all seem to be capable topics to represent the *organizer* product feature, but results of applying the discarding method (**Annex F: Discarding method applied to LDA topics**) and the human judgement relative to the semantic connexion between the *organizer* initial terms and the remaining terms after the discarding method are shown in **Table 5-35** and **Table 5-36**.

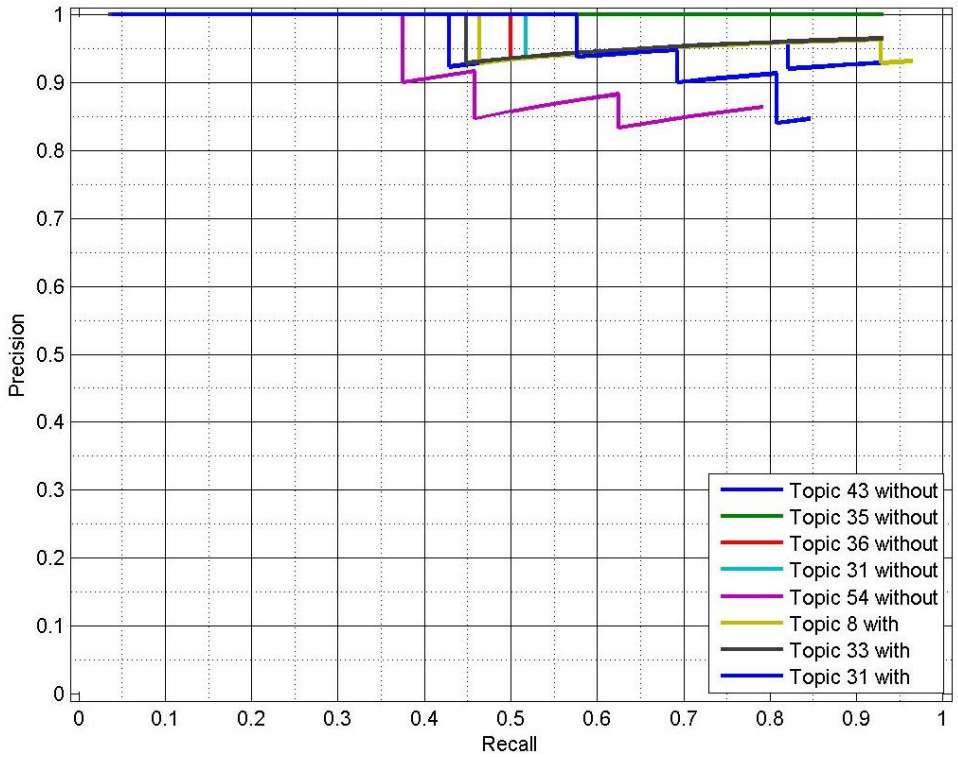
To group the results graphically, the precision vs. recall graphic is presented in **Graphic 5-20**, while the precision, accuracy and recall single graphics are located in **Annex I: Precision, recall and accuracy graphs of LDA**.

Place	*Topic 43* without *k=50*	*Topic 35* without *k=100*	*Topic 36* without *k=100*	*Topic 31* without *k=200*
1	**0.0494 alarm** (***tp***)	**0.0508 organizer** (***tp***)	0.0365 time (***tp***)	**0.0557 alarm** (***tp***)
2	**0.0428 calendar** (***tp***)	**0.0500 calendar** (***tp***)	**0.0338 alarm** (***tp***)	**0.0533 calendar** (***tp***)
3	**0.0391 organizer** (***tp***)	**0.0489 alarm** (***tp***)	0.0302 event (***tp***)	**0.0424 organizer** (***tp***)
4	0.0351 option (***tp***)	0.0423 option (***tp***)	0.0256 day (***tp***)	0.0384 option (***tp***)
5	**0.0322 note** (***tp***)	**0.0403 note** (***tp***)	**0.0246 calendar** (***tp***)	0.0368 day (***tp***)
6	0.0277 day (***tp***)	**0.0299 calculator** (***tp***)	0.0182 date (***tp***)	**0.0328 note** (***tp***)
7	0.0262 time (***tp***)	0.0276 day (***tp***)	0.0177 option (***tp***)	~~0.0311 time~~ (*fn*)
8	0.0244 clock (***tp***)	0.0250 clock (***tp***)	0.0159 view (***tp***)	**0.0285 calculator** (***tp***)
9	**0.0231 calculator** (***tp***)	0.0227 menu (***tp***)	~~0.0159 function~~ (*fn*)	0.0253 clock (***tp***)
10	0.0210 menu (***tp***)	**0.0205 task** (***tp***)	0.0157 clock (***tp***)	**0.0234 task** (***tp***)
11	0.0192 event (***tp***)	0.0196 time (***tp***)	0.0141 hour (***tp***)	0.0232 week (***tp***)
12	0.0175 function (***tp***)	0.0196 world (***tp***)	0.0132 end (***tp***)	~~0.0209 function~~ (*fn*)
13	0.0169 world (*fp*)	0.0195 function (***tp***)	**0.0130 note** (***tp***)	0.0205 event (***tp***)
14	**0.0161 task** (***tp***)	**0.0159 memo** (***tp***)	**0.0120 organizer** (***tp***)	0.0186 month (***tp***)
15	0.0160 voice (***tp***)	**0.0155 world clock** (***tp***)	0.0115 start (***tp***)	~~0.0177 menu~~ (*fn*)
16	~~0.0151 phone~~ (*fn*)	~~0.0151 phone~~ (*fn*)	0.0111 mode (*fp*)	0.0146 appointment (***tp***)
17	**0.0133 world clock** (***tp***)	0.0150 week (***tp***)	**0.0106 task** (***tp***)	~~0.0144 voice~~ (*fn*)
18	0.0133 memory (***tp***)	**0.0138 converter** (***tp***)	0.0101 default (***tp***)	**0.0144 memo** (***tp***)
19	0.0130 week (***tp***)	0.0132 timer (***tp***)	~~0.0098 phone~~ (*fn*)	**0.0124 world clock** (***tp***)
20	**0.0120 converter** (***tp***)	0.0124 event (***tp***)	0.0090 week (***tp***)	0.0115 world (*fp*)
21	**0.0116 memo** (***tp***)	0.0123 month (***tp***)	~~0.0087 menu~~ (*fn*)	0.0110 reminder (***tp***)
22	0.0111 month (***tp***)	0.0105 voice (***tp***)	0.0087 voice (***tp***)	**0.0109 converter** (***tp***)
23	0.0097 timer (***tp***)	~~0.0105 application~~ (*fn*)	~~0.0087 way~~ (***tn***)	0.0109 date (***tp***)
24	0.0094 file (***tp***)	0.0103 appointment (***tp***)	**0.0082 calculator** (***tp***)	0.0104 unit (***tp***)
25	0.0085 view (***tp***)	**0.0095 stopwatch** (***tp***)	0.0081 month (***tp***)	~~0.0101 view~~ (*fn*)
26	0.0081 card (*fp*)	0.0085 unit (***tp***)	0.0080 end time (***tp***)	0.0099 timer (***tp***)
27	0.0077 date (***tp***)	~~0.0080 one~~ (***tn***)	0.0076 scheduler (***tp***)	0.0098 record (***tp***)
28	0.0075 application (***tp***)	0.0077 record (***tp***)	0.0075 entry (***tp***)	**0.0094 stopwatch** (***tp***)
29	0.0073 manager (***tp***)	0.0074 reminder (***tp***)	0.0075 field (***tp***)	0.0086 unit converter (***tp***)
30	0.0073 appointment (***tp***)	0.0071 list (***tp***)	0.0070 list (***tp***)	0.0083 organizer function (***tp***)

Table 5-35. Judgement applied over the top 30 terms of *topic 43* (without), *topic 35* (without), *topic 36* (without) and *topic 31* (without) after discarding method.

Place	*Topic 54* without *k=200*	*Topic 8* with *k=50*	*Topic 33* with *k=100*	*Topic 31* with *k=200*
1	0.0646 file (***tp***)	**0.0407 alarm** (***tp***)	**0.0478 alarm** (***tp***)	0.0625 time (***tp***)
2	0.0300 manager (***tp***)	**0.0385 calendar** (***tp***)	**0.0451 calendar** (***tp***)	0.0522 day (***tp***)
3	~~0.0298 phone~~ (*fn*)	**0.0331 organizer** (***tp***)	**0.0420 organizer** (***tp***)	**0.0482 alarm** (***tp***)
4	0.0269 menu (***tp***)	0.0313 option (***tp***)	0.0372 option (***tp***)	0.0439 clock (***tp***)
5	**0.0265 note** (***tp***)	**0.0302 note** (***tp***)	**0.0350 note** (***tp***)	0.0395 date (***tp***)
6	**0.0260 alarm** (***tp***)	0.0228 day (***tp***)	0.0286 day (***tp***)	**0.0349 calendar** (***tp***)
7	0.0235 file manager (***tp***)	0.0221 time (***tp***)	0.0258 time (***tp***)	0.0234 week (***tp***)
8	0.0170 symbian (***tp***)	**0.0209 calculator** (***tp***)	**0.0253 calculator** (***tp***)	0.0220 event (***tp***)
9	0.0166 quickoffice (***tp***)	0.0196 clock (***tp***)	0.0225 clock (***tp***)	0.0206 hour (***tp***)
10	~~0.0164 option~~ (*fn*)	0.0175 menu (***tp***)	0.0214 menu (***tp***)	0.0190 start (***tp***)
11	0.0157 smartphone (***tp***)	0.0167 function (***tp***)	0.0190 event (***tp***)	**0.0177 note** (***tp***)
12	~~0.0157 time~~ (*fn*)	0.0162 event (***tp***)	**0.0177 task** (***tp***)	0.0165 view (***tp***)
13	0.0148 city (*fp*)	**0.0145 task** (***tp***)	0.0173 function (***tp***)	0.0158 month (***tp***)
14	**0.0146 office** (***tp***)	0.0141 world (*fp*)	0.0154 world (*fp*)	0.0147 way (***tp***)
15	~~0.0143 application~~ (*fn*)	~~0.0136 phone~~ (*fn*)	0.0145 week (***tp***)	0.0144 field (***tp***)
16	0.0141 type (***tp***)	0.0122 voice (***tp***)	0.0127 month (***tp***)	0.0144 location (*fp*)
17	0.0131 e90 (*fp*)	**0.0113 world clock** (***tp***)	**0.0124 memo** (***tp***)	0.0142 list (***tp***)
18	~~0.0121 lot~~ (***tn***)	0.0112 week (***tp***)	**0.0123 converter** (***tp***)	~~0.0127 option~~ (*fn*)
19	~~0.0116 way~~ (***tn***)	**0.0112 memo** (***tp***)	**0.0122 world clock** (***tp***)	0.0122 end (***tp***)
20	~~0.0108 one~~ (***tn***)	0.0111 memory (***tp***)	~~0.0118 phone~~ (*fn*)	0.0121 appointment (***tp***)
21	**0.0105 converter** (***tp***)	**0.0109 converter** (***tp***)	~~0.0107 file~~ (*fn*)	0.0114 world (*fp*)
22	0.0099 document (***tp***)	0.0098 month (***tp***)	0.0103 timer (***tp***)	0.0106 entry (***tp***)
23	0.0093 clock (***tp***)	0.0085 timer (***tp***)	0.0098 appointment (***tp***)	**0.0103 world clock** (***tp***)
24	~~0.0091 interface~~ (*fn*)	0.0079 file (***tp***)	0.0093 voice (***tp***)	~~0.0091 menu~~ (*fn*)
25	**0.0091 calendar** (***tp***)	0.0075 view (***tp***)	0.0083 manager (***tp***)	**0.0083 task** (***tp***)
26	0.0090 n76 (*fp*)	0.0075 application (***tp***)	0.0081 view (***tp***)	0.0074 zone (*fp*)
27	**0.0084 organizer** (***tp***)	0.0071 appointment (***tp***)	0.0077 date (***tp***)	~~0.0071 number~~ (*fn*)
28	**0.0080 calculator** (***tp***)	0.0066 record (***tp***)	**0.0075 stopwatch** (***tp***)	0.0067 space (*fp*)
29	**0.0076 to-do** (***tp***)	0.0064 card (*fp*)	0.0074 list (***tp***)	~~0.0067 minute~~ (*fn*)
30	**0.0075 pdf** (***tp***)	**0.0063 stopwatch** (***tp***)	0.0073 type (***tp***)	0.0066 time zone (***tp***)

Table 5-36. Judgement applied over the top 30 terms of *topic 54* (without), *topic 8* (with), *topic 33* (with) and *topic 31* (with) after discarding method.



Graphic 5-20. Precision and recall graph of selected topics of *organizer* on sections.

Looking at the graphic presented above, it is seen that any of the topics decrease their precision under the 80%, but they do not reach the complete recall with the first 30 terms, because the discarding method understands that their terms are not as correlated as they are. *Topic 35* seems to be the best performance of LDA in this section, although, it is checked that from *topic 36* more new terms are extracted, anyway, in the following summarization it is seen how they perform.

From ***topic 43*** **without** *extra reviews*, terms such as *option*, *day*, *time*, *clock*, *menu*, *event*, *function*, *voice*, *memory*, *week*, *month*, *timer*, *file*, *view*, *day*, *application*, *manager* and *appointment* are extracted with 93.10% of precision, 90% accuracy, but a 96.43% of recall. From ***topic 35*** **without** *extra reviews*, terms such as *option*, *day*, *clock*, *menu*, *time*, *function*, *week*, *timer*, *event*, *month*, *voice*, *appointment*, *unit*, *record*, *reminder* and *list* are extracted with 100% of precision, 93.33% accuracy, but a 93.10% of recall. From ***topic 36*** **without** *extra reviews*, terms such as *time*, *event*, *day*, *date*, *option*, *view*, *clock*, *hour*, *end*, *start*, *default*, *week*, *voice*, *month*, *end time*, *scheduler*, *entry*, *field* and *list* are extracted with 96.15% of precision, 86.67% accuracy, but a 89.29% of recall. From ***topic 31*** **without** *extra reviews*, terms such as *option*, *day*, *clock*, *week*, *event*, *month*, *appointment*, *reminder*, *date*, *unit*, *timer*, *record*, *unit converter* and *organizer function* are extracted with 96% of precision, 80% accuracy, but a 82.76% of recall. From ***topic 54*** **without** *extra reviews*, terms such as *file*, *manager*, *menu*, *file manager*, *symbian*, *quickoffice*, *smartphone*, *clock*, *type* and *document* are extracted with 86.36% of precision, 73.33% accuracy, but a 79.17% of recall. From ***topic 8*** **with** *extra reviews*, terms such as *option*, *day*, *time*, *clock*, *menu*, *function*, *event*, *voice*, *week*, *memory*, *month*, *timer*, *file*, *view*, *application*, *appointment* and *record* are extracted with 93.10% of precision, 90% accuracy, but a 96.43% of recall. From ***topic 33*** **with** *extra reviews*, terms such as *option*, *day*, *time*, *clock*, *menu*, *event*, *function*, *week*, *month*, *timer*, *appointment*, *voice*, *manager*, *view*, *date*, *list* and *type* are extracted with 96.43% of precision, 90% accuracy, but a 93.10% of recall. From ***topic 31*** **with** *extra reviews*, terms such as *time*, *day*, *clock*, *date*, *week*, *event*, *hour*, *start*, *view*, *month*, *field*, *list*, *end*, *appointment*, *entry* and *time zone* are extracted with 84.62% of precision, 73.33% accuracy, but a 84.62% of recall.

5.4.2.3. Paragraphs

Ending with the *organizer* chapter, better performances than in PLSI and LSI are found, where some topics reach up to 11 initial terms within their top 30 terms in the sorted topic. In spite of having problems grouping semantically related terms (as it occurs with PLSI on almos each performance on paragraphs scenario), LDA's measures remains high, and similar topics are found almost in the same positions. For example, from table **Table 11-6** are selected topics such as *topic 3* and *topic 4*, but the first one is selected on performances with k=100 without *extra reviews*, and k=50, 100 with them included. Then it is noticed than LDA group not only terms into topics, even though semantically related topics close to each other and in similar places. To carry out the same analysis applied on the rest of scenarios and performances, steps described in **Table 4-1** are followed to extract some conclusion regarding the precision, accuracy and recall of the different extraction of potential semantically related terms with the initial product features relative to the *organizer* topic.

Place	*Topic 4* without *k=50*	*Topic 3* without *k=100*	*Topic 60* without *k=150*	*Topic 3* with *k=50*
1	**0.0481 calendar** (***tp***)	0.0656 day (***tp***)	0.1209 day (***tp***)	**0.0471 calendar** (***tp***)
2	**0.0422 alarm** (***tp***)	**0.0488 calendar** (***tp***)	**0.0846 calendar** (***tp***)	**0.0429 alarm** (***tp***)
3	0.0373 day (***tp***)	0.0424 week (***tp***)	0.0661 week (***tp***)	0.0380 day (***tp***)
4	**0.0355 note** (***tp***)	**0.0366 task** (***tp***)	0.0600 event (***tp***)	**0.0359 note** (***tp***)
5	0.0315 option (***tp***)	0.0363 event (***tp***)	0.0489 month (***tp***)	0.0308 option (***tp***)
6	0.0288 clock (***tp***)	0.0339 month (***tp***)	0.0464 appointment (***tp***)	**0.0291 task** (***tp***)
7	0.0278 time (***tp***)	~~0.0266 option~~ (*fn*)	0.0454 option (***tp***)	0.0282 clock (***tp***)
8	**0.0246 task** (***tp***)	0.0264 time (***tp***)	0.0447 time (***tp***)	0.0277 time (***tp***)
9	0.0225 week (***tp***)	0.0244 date (***tp***)	**0.0208 alarm** (***tp***)	0.0227 week (***tp***)
10	**0.0221 calculator** (***tp***)	0.0230 appointment (***tp***)	0.0203 priority (***tp***)	**0.0227 calculator** (***tp***)
11	**0.0183 organizer** (***tp***)	**0.0155 note** (***tp***)	0.0203 reminder (***tp***)	**0.0178 organizer** (***tp***)
12	0.0172 world (*fp*)	0.0150 reminder (***tp***)	**0.0191 stopwatch** (***tp***)	0.0178 month (***tp***)
13	0.0171 menu (***tp***)	~~0.0133 menu~~ (*fn*)	0.0184 view (***tp***)	0.0177 event (***tp***)
14	0.0169 event (***tp***)	0.0126 view (***tp***)	0.0179 year (***tp***)	0.0173 world (*fp*)
15	0.0165 month (***tp***)	0.0119 list (***tp***)	0.0176 meeting (***tp***)	0.0155 menu (***tp***)
16	0.0143 date (***tp***)	0.0105 year (***tp***)	0.0170 date (***tp***)	0.0140 date (***tp***)
17	0.0125 appointment (***tp***)	**0.0104 alarm** (***tp***)	0.0166 location (*fp*)	0.0127 appointment (***tp***)
18	**0.0124 world clock** (***tp***)	0.0098 start (***tp***)	0.0166 hour (***tp***)	**0.0123 world clock** (***tp***)
19	~~0.0122 function~~ (*fn*)	0.0096 priority (***tp***)	**0.0157 note** (***tp***)	~~0.0117 function~~ (*fn*)
20	**0.0110 converter** (***tp***)	0.0089 category (***tp***)	~~0.0143 menu~~ (*fn*)	**0.0115 converter** (***tp***)
21	0.0092 reminder (***tp***)	0.0079 meeting (***tp***)	0.0140 start (***tp***)	0.0095 reminder (***tp***)
22	0.0089 list (***tp***)	0.0076 location (*fp*)	0.0105 timer (***tp***)	0.0089 list (***tp***)
23	**0.0083 memo** (***tp***)	0.0073 recurrence (*fp*)	0.0098 scheduler (***tp***)	0.0079 timer (***tp***)
24	0.0077 timer (***tp***)	0.0071 subject (***tp***)	~~0.0094 contact~~ (*fn*)	**0.0076 memo** (***tp***)
25	0.0076 start (***tp***)	0.0068 day week (***tp***)	0.0093 entry (***tp***)	0.0075 view (***tp***)
26	0.0071 view (***tp***)	0.0067 sensitive (*fp*)	0.0092 end (***tp***)	0.0074 start (***tp***)
27	0.0070 location (*fp*)	0.0060 entry (***tp***)	0.0092 field (***tp***)	0.0070 location (*fp*)
28	~~0.0068 phone~~ (*fn*)	0.0059 anniversary (***tp***)	0.0087 birthday (***tp***)	0.0062 unit (***tp***)
29	0.0057 priority (***tp***)	0.0057 item (*fp*)	0.0086 duration (***tp***)	0.0057 year (***tp***)
30	0.0056 field (***tp***)	0.0056 scheduler (***tp***)	0.0084 minute (***tp***)	0.0057 priority (***tp***)

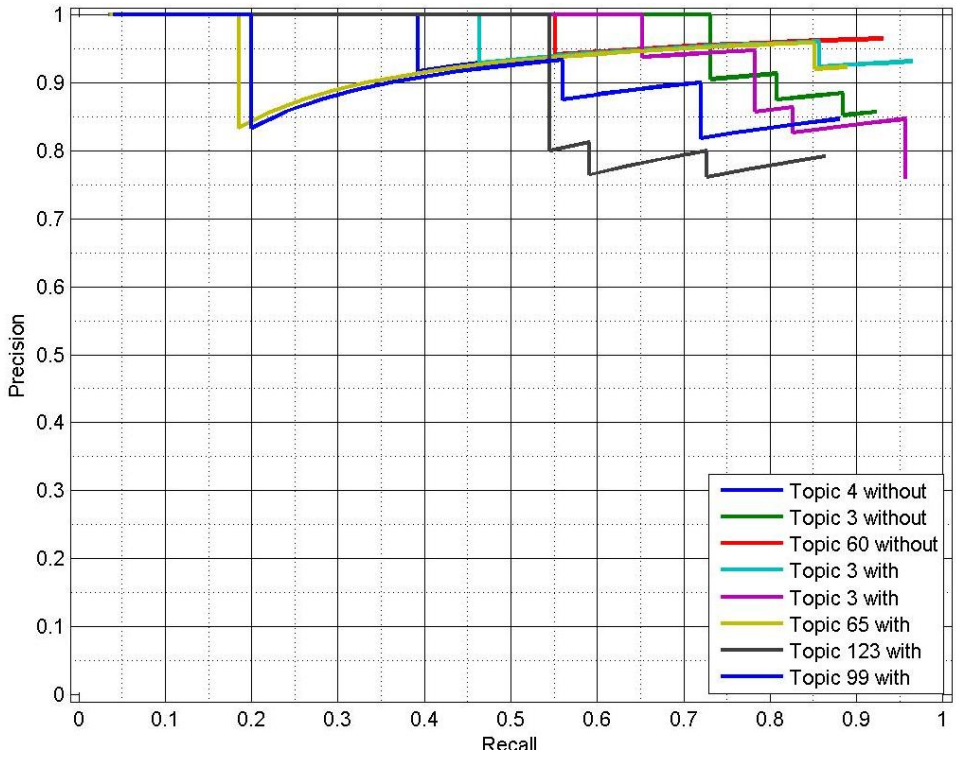
Table 5-37. Judgement applied over the top 30 terms of *topic 4* (without), *topic 8* (without), *topic 60* (without) and *topic 3* (with) after discarding method.

Once topics are selected, the discarding method is applied in order to omit those terms which have a poor semantic relation with the original *organizer* ones, because when terms overcome the threshold of having better scores in more than 5 topics than in the current topic, it is supposed that there is a weak relation with the topic in question. Results of this application are found in **Annex F: Discarding method applied to LDA topics**. Then, human judgement is applied on discarding method results, to discriminate how many terms of the top 30 terms of each sorted topic are actually semantically related with the *organizer* initial product features. **Table 5-37** and **Table 5-38** collect both procedures and show each term with its behaviour in the retrieval.

In **Graphic 5-21**, the precision vs. recall graphic of the topic's performance, it is possible to see how the worst performances are from *topic 3* with k=100 and *topic 123*, both with *extra reviews*, and also both do not remain with a precision superior to 80%. Meanwhile *topic 60* without *extra reviews* and *topic 3* with *them* and with k=50 remain with high precision and almost the complete recall at the end of the top 30 terms. The rest of graphics relatives to results obtained from these performances are located in **Annex I: Precision, recall and accuracy graphs of LDA**. A summarization of these results is done below.

Place	*Topic 3* with *k=100*	*Topic 65* with *k=100*	*Topic 123* with *k=150*	*Topic 99* with *k=200*
1	0.0608 day (***tp***)	**0.1081 alarm** (***tp***)	0.1768 event (***tp***)	0.1602 day (***tp***)
2	0.0508 week (***tp***)	0.0788 clock (***tp***)	**0.1634 calendar** (***tp***)	0.0926 appointment (***tp***)
3	**0.0466 task** (***tp***)	**0.0607 note** (***tp***)	**0.0501 alarm** (***tp***)	0.0899 week (***tp***)
4	0.0429 month (***tp***)	**0.0537 calculator** (***tp***)	**0.0411 organizer** (***tp***)	**0.0626 calendar** (***tp***)
5	0.0418 date (***tp***)	**0.0500 calendar** (***tp***)	0.0387 function (***tp***)	0.0584 option (***tp***)
6	0.0341 time (***tp***)	0.0452 world (*fp*)	0.0365 entry (***tp***)	0.0539 location (*fp*)
7	**0.0282 calendar** (***tp***)	0.0436 option (***tp***)	**0.0324 note** (***tp***)	0.0499 time (***tp***)
8	0.0223 option (***tp***)	**0.0422 organizer** (***tp***)	**0.0250 to-do** (***tp***)	**0.0401 task** (***tp***)
9	0.0219 appointment (***tp***)	**0.0339 world clock** (***tp***)	0.0215 view (***tp***)	**0.0370 note** (***tp***)
10	0.0196 view (***tp***)	0.0328 menu (***tp***)	0.0201 list (***tp***)	0.0344 date (***tp***)
11	0.0164 list (***tp***)	**0.0308 converter** (***tp***)	~~0.0184 time~~ (*fn*)	0.0342 reminder (***tp***)
12	0.0146 category (***tp***)	0.0287 function (***tp***)	~~0.0175 option~~ (*fn*)	0.0329 menu (***tp***)
13	0.0145 priority (***tp***)	0.0267 time (***tp***)	0.0175 day (***tp***)	0.0316 start (***tp***)
14	~~0.0137 menu~~ (*fn*)	**0.0265 memo** (***tp***)	**0.0170 currency** (***tp***)	0.0289 year (***tp***)
15	0.0115 start (***tp***)	0.0242 timer (***tp***)	0.0152 information (*fp*)	0.0259 month (***tp***)
16	0.0106 reminder (***tp***)	0.0205 application (***tp***)	~~0.0141 mode~~ (*fn*)	0.0202 address (*fp*)
17	0.0100 pan (*fp*)	**0.0202 task** (***tp***)	0.0137 etc (*fp*)	0.0199 subject (***tp***)
18	0.0097 year (***tp***)	**0.0157 stopwatch** (***tp***)	~~0.0136 screen~~ (***tn***)	0.0193 end (***tp***)
19	**0.0095 note** (***tp***)	0.0148 unit (***tp***)	0.0132 course (*fp*)	**0.0191 alarm** (***tp***)
20	**0.0090 alarm** (***tp***)	0.0107 number (***tp***)	~~0.0127 way~~ (***tn***)	~~0.0173 contact~~ (*fn*)
21	0.0089 recurrence (*fp*)	~~0.0105 one~~ (***tn***)	**0.0122 task** (***tp***)	~~0.0141 type~~ (*fn*)
22	0.0076 out (*fp*)	0.0099 list (***tp***)	0.0121 click (*fp*)	0.0121 view (***tp***)
23	0.0073 subject (***tp***)	0.0096 tool (***tp***)	0.0120 calendar entry (***tp***)	0.0082 repetition (*fp*)
24	0.0071 glance (*fp*)	0.0094 day (***tp***)	**0.0117 calculator** (***tp***)	0.0057 business (*fp*)
25	0.0071 duration (***tp***)	~~0.0088 interface~~ (*fn*)	0.0110 currency converter (***tp***)	0.0049 tool (***tp***)
26	0.0067 holiday (***tp***)	**0.0087 countdown** (***tp***)	0.0108 latter (*fp*)	0.0045 copy (***tp***)
27	0.0066 sort (***tp***)	~~0.0086 phone~~ (*fn*)	0.0106 clock (***tp***)	**0.0040 office** (***tp***)
28	0.0065 in (*fp*)	0.0086 menus (*fp*)	~~0.0098 type~~ (*fn*)	~~0.0032 list~~ (*fn*)
29	0.0065 in out (*fp*)	0.0080 countdown timer (***tp***)	0.0096 birthday (***tp***)	~~0.0031 menus~~ (***tn***)
30	0.0065 item (*fp*)	~~0.0070 name~~ (*fn*)	0.0081 meeting (***tp***)	0.0026 tool menu (***tp***)

Table 5-38. Judgement applied over the top 30 terms of *topic 3* (with), *topic 65* (with), *topic 123* (with) and *topic 99* (with) after discarding method.



Graphic 5-21. Precision and recall graph of selected topics of *organizer* on paragraphs.

From ***topic 4*** **without** *extra reviews*, terms such as *day*, *option*, *clock*, *time*, *week*, *menu*, *event*, *month*, *date*, *appointment*, *reminder*, *list*, *timer*, *start*, *view*, *priority* and *field* are extracted with 92.86% of precision, 86.67% accuracy, and a XX% of recall. From ***topic 3*** **without** *extra reviews*, terms such as *day*, *week*, *event*, *month*, *time*, *date*, *appointment*, *reminder*, *view*, *list*, *year*, *start*, *priority*, *category*, *meeting*, *subject*, *day week*, *entry*, *anniversary* and *scheduler* are extracted with 85.71% of precision, 80% accuracy, and a 92.31% of recall. From ***topic 60*** **without** *extra reviews*, terms such as *day*, *week*, *event*, *month*, *appointment*, *option*, *time*, *priority*, *reminder*, *view*, *year*, *meeting*, *date*, *hour*, *start*, *timer*, *scheduler*, *entry*, *end*, *field*, *birthday*, *duration* and *minute* are extracted with 96.43% of precision, 90% accuracy, and a 93.10% of recall. From ***topic 3*** **with** *extra reviews* with **k=50**, terms such as *day*, *option*, *clock*, *time*, *week*, *month*, *event*, *menu*, *date*, *appointment*, *reminder*, *list*, *timer*, *view*, *start*, *unit*, *year* and *priority* are extracted with 93.10% of precision, 90 % accuracy, and a 96.43% of recall. From ***topic 3*** **with** *extra reviews* with **k=100**, terms such as *day*, *week*, *month*, *date*, *timer*, *option*, *appointment*, *view*, *list*, *category*, *priority*, *start*, *reminder*, *year*, *subject*, *duration*, *holiday* and *sort* are extracted with 84.62% of precision, 81.48% accuracy, and a 95.65% of recall. From ***topic 65*** **with** *extra reviews*, terms such as *clock*, *option*, *menu*, *function*, *time*, *timer*, *application*, *unit*, *number*, *list*, *tool*, *day* and *countdown timer* are extracted with 92.31% of precision, 86.21 % accuracy, and a 88.89 % of recall. From ***topic 123*** **with** *extra reviews*, terms such as *event*, *function*, *entry*, *view*, *list* and *day* are extracted with 100% of precision, 85.71% accuracy, and a 54.55% of recall. From ***topic 99*** **with** *extra reviews*, terms such as *day*, *appointment*, *week*, *option*, *time*, *date*, *reminder*, *menu*, *start*, *year*, *month*, *subject*, *end*, *view*, *tool*, *copy* and *tool menu* are extracted with 84.62% of precision, 76.67% accuracy, and a 88% of recall.

5.4.3. Multimedia

5.4.3.1. Documents

To finalize with the LDA analysis, it is run looking for the *multimedia* product features semantically related with the initial ones. In this first section, the analysis is done on the documents scenario in performances with k=50, 100, 150, 200, like it is done with LSI and PLSI. Topics where *multimedia* initial product features are found best scored within the top 30 terms of the sorted topics are found in **Table 11-7**, from where, following the steps described in **Table 4-1**, the selection of topics is done there too.

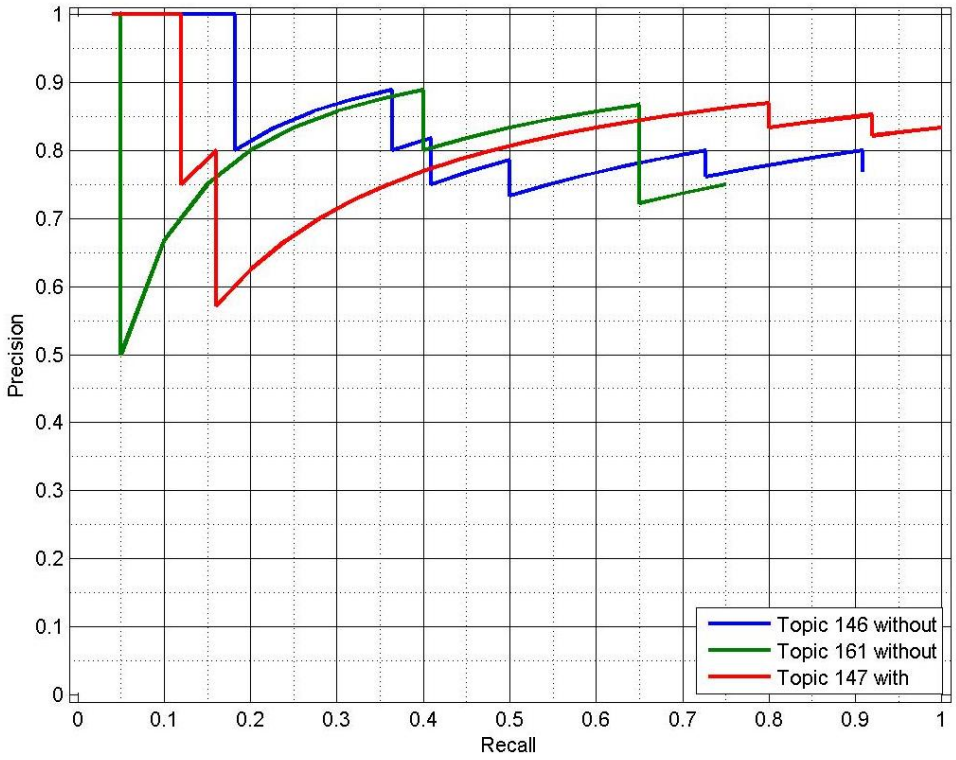
Here it is the first time that LDA seems to perform poorer than PLSI (because LSI does not work in the documents scenario), because with PLSI there are up to 6 topics selected, meanwhile here, there are only 3. However, they seem to describe better the *multimedia* product feature, although almost the same number of initial terms is included in the 30 first terms.

Place	*Topic 146* without *k=200*	*Topic 161* without *k=200*	*Topic 147* with *k=150*
1	0.0736 art (***tp***)	0.0917 music (***tp***)	0.0872 music (***tp***)
2	0.0552 phone (***tp***)	~~0.0521 phone~~ (*fn*)	0.0464 phone (***tp***)
3	0.0368 album (***tp***)	0.0373 button (*fp*)	0.0429 earphone (***tp***)
4	0.0337 version (***tp***)	0.0362 headphone (***tp***)	0.0304 k750i (*fp*)
5	0.0270 serie (*fp*)	0.0331 earphone (***tp***)	0.0269 player (***tp***)
6	0.0262 music (***tp***)	0.0301 player (***tp***)	0.0206 button (*fp*)
7	0.0246 player (***tp***)	0.0228 sound (***tp***)	0.0175 difference (*fp*)
8	0.0207 display (***tp***)	**0.0213 music player** (***tp***)	0.0174 function (***tp***)
9	**0.0195 video** (***tp***)	0.0170 bluetooth (***tp***)	0.0172 play (***tp***)
10	0.0193 unit (*fp*)	~~0.0133 picture~~ (*fn*)	0.0162 pause (***tp***)
11	0.0189 color (***tp***)	0.0130 song (***tp***)	**0.0146 mp3** (***tp***)
12	0.0172 level (*fp*)	0.0121 navigation (*fp*)	0.0132 option (***tp***)
13	0.0172 artist (***tp***)	0.0110 pause (***tp***)	0.0116 track (***tp***)
14	0.0171 resolution (***tp***)	0.0101 volume (***tp***)	0.0113 joystick (***tp***)
15	0.0167 front (*fp*)	**0.0090 stereo** (***tp***)	0.0112 play pause (***tp***)
16	~~0.0165 interface~~ (*fn*)	0.0090 speaker (***tp***)	0.0109 quality (***tp***)
17	0.0164 album art (***tp***)	~~0.0089 camera~~ (*fn*)	0.0104 song (***tp***)
18	0.0146 cover (***tp***)	0.0086 type (***tp***)	**0.0098 music player** (***tp***)
19	0.0143 image (***tp***)	0.0084 direction (*fp*)	0.0097 telephone (***tp***)
20	0.0135 genre (***tp***)	~~0.0084 one~~ (***tn***)	0.0096 sound (***tp***)
21	0.0126 light (***tp***)	0.0082 hand (*fp*)	0.0096 handset (***tp***)
22	0.0124 design (*fp*)	0.0075 conversation (*fp*)	0.0089 mp3 player (***tp***)
23	0.0121 camera interface (***tp***)	~~0.0074 quality~~ (*fn*)	0.0082 type (***tp***)
24	0.0116 media (***tp***)	0.0073 play (***tp***)	0.0082 level (*fp*)
25	0.0111 playlist (***tp***)	~~0.0073 side~~ (***tn***)	0.0081 memory (***tp***)
26	~~0.0111 side~~ (***tn***)	0.0065 jack (***tp***)	0.0081 headphone (***tp***)
27	**0.0109 music player** (***tp***)	~~0.0065 color~~ (*fn*)	0.0077 camera (***tp***)
28	~~0.0103 feature~~ (***tn***)	~~0.0063 day~~ (***tn***)	0.0075 design (*fp*)
29	0.0102 feedback (*fp*)	~~0.0061 time~~ (***tn***)	0.0073 device (***tp***)
30	~~0.0091 quality~~ (*fn*)	~~0.0061 home~~ (***tn***)	0.0073 colour (***tp***)

Table 5-39. Judgement applied over the top 30 terms of *topic 146* (without), *topic 161* (without) and *topic 147* (with) after discarding method.

Next step consists in the application of the discarding method, which do not fail at all, but decreases the performance of the two first topics, meanwhile it does not discard any term in *topic 147*. Results of this application are found in **Annex F: Discarding method applied to LDA topics**. Then, the human judgement is applied and showed in **Table 5-39**, where results co-exist with the discarding method ones to see how topics have been retrieved and evaluated.

Finally, to summarize the obtained results, the precision vs. recall graphic is presented in **Graphic 5-22**, where it is seen how *topic 146* has the best performance, but it still remain as poor as PLSI's performances. Moreover, it is noticed that there is no improvement by adding the *extra reviews*, due to the graphic shows poorer performance of the only topic selected in this section. Grouping results, it is concluded that from ***topic 146* without** *extra reviews*, terms such as *art*, *phone*, *album*, *version*, *music*, *player*, *display* and *color* are extracted with 81.81% of precision and accuracy, and a 40.91% of recall. From ***topic 161* without** *extra reviews*, only term *music* is extracted with 100% of precision and accuracy, but a 5% of recall. From ***topic 147* with** *extra reviews*, terms such as *music*, *phone* and *earphone* are extracted with 100% of precision and accuracy, and a 12% of recall.



Graphic 5-22. Precision and recall graph of selected topics of *multimedia* on documents.

5.4.3.2. Sections

The same process is carried out on this section, but this time looking for the *multimedia* topic on sections scenario. LDA returns many topics with many initial terms within their top 30 terms, some up to 11 initial terms of the list of 42 (with *extra reviews*), then, it is selected the best ones from **Table 11-8**, following the steps enounced in **Table 4-1**.

Once topics are selected, discarding method is applied getting resultant graphics located in **Annex F: Discarding method applied to LDA topics**, where many correct potential new terms are discarded, but somehow terms with a weak relation with the rest of the topic should be discarded, and discarding method removes those terms whose scores in other topics are higher than the current one.

Finally, terms are human judged in order to satisfy the *multimedia* context, being or not related with the *multimedia* initial product features. Then, they are judged as *true positive (tp),* a correct result*, false positive (fp),* an unexpected result*, true negative (tn),* a correct absence of result*,* and *false negative (fn)*, a missing result. Results of both actions over the terms are found in **Table 5-40** and **Table 5-41**, where also the meaning of the terms can be discussed.

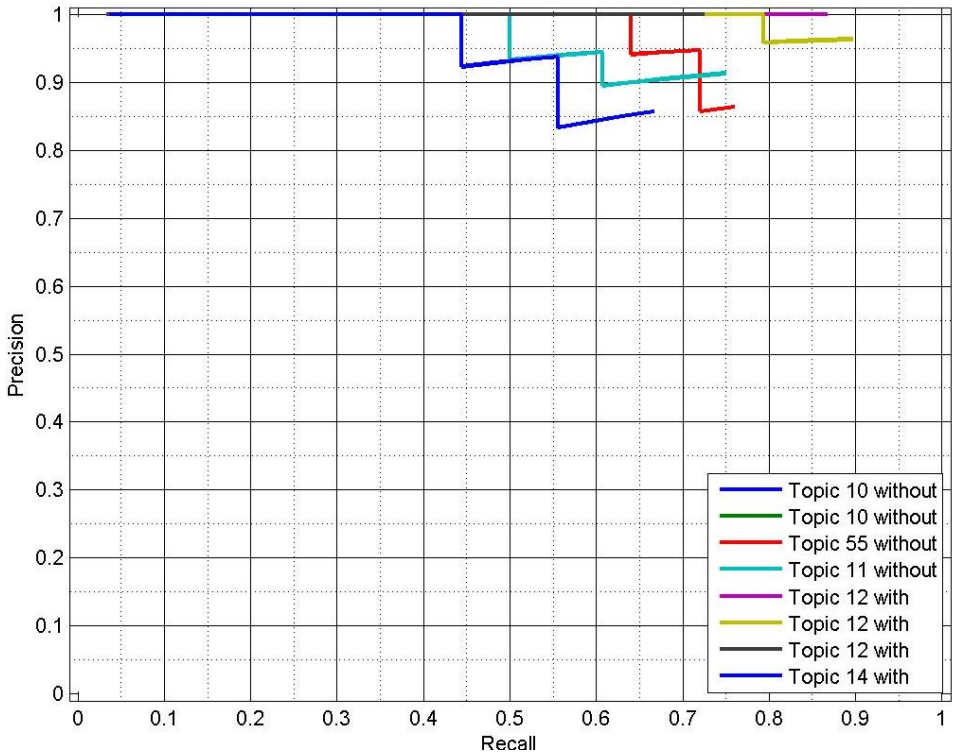
To summarize these results, a precision vs. recall graphic is showed in **Graphic 5-23**, where LDA's performances appear as good as it was thought, because all of them remain above the 80% of precision. However, their recall is poor due to the discarding method, which sometimes hit and removes wrong retrieved terms, but in the other hand, it sometimes fails. The rest of single graphics relatives to their accuracy, precision and recall are located in **Annex I: Precision, recall and accuracy graphs of LDA**.

Place	*Topic 10* without *k=50*	*Topic 10* without *k=100*	*Topic 55* without *k=150*	*Topic 11* without *k=200*
1	0.0847 music (***tp***)	0.0568 music (***tp***)	**0.1736 radio** (***tp***)	**0.0736 video** (***tp***)
2	0.0552 player (***tp***)	0.0519 player (***tp***)	**0.1453 fm** (***tp***)	0.0566 player (***tp***)
3	**0.0334 music player** (***tp***)	**0.0401 video** (***tp***)	0.1012 fm radio (***tp***)	0.0337 file (***tp***)
4	**0.0273 video** (***tp***)	0.0226 file (***tp***)	0.0387 player (***tp***)	0.0315 media (***tp***)
5	~~0.0259 phone~~ (*fn*)	0.0217 album (***tp***)	0.0351 music (***tp***)	**0.0280 multimedia** (***tp***)
6	**0.0256 radio** (***tp***)	~~0.0201 phone~~ (*fn*)	**0.0338 rds** (***tp***)	0.0267 playback (***tp***)
7	**0.0210 fm** (***tp***)	**0.0189 music player** (***tp***)	0.0252 station (***tp***)	0.0240 media player (***tp***)
8	0.0186 album (***tp***)	0.0154 song (***tp***)	~~0.0210 function~~ (*fn*)	**0.0221 divx** (***tp***)
9	0.0185 sound (***tp***)	**0.0153 multimedia** (***tp***)	0.0205 sound (***tp***)	0.0216 resolution (***tp***)
10	0.0166 track (***tp***)	0.0141 media (***tp***)	~~0.0204 phone~~ (*fn*)	**0.0200 xvid** (***tp***)
11	0.0153 fm radio (***tp***)	0.0140 sound (***tp***)	0.0189 headphone (***tp***)	~~0.0191 quality~~ (*fn*)
12	0.0142 option (***tp***)	0.0133 artist (***tp***)	**0.0181 audio** (***tp***)	**0.0147 video playback** (***tp***)
13	~~0.0135 interface~~ (*fn*)	0.0130 track (***tp***)	0.0180 antenna (***tp***)	**0.0135 audio** (***tp***)
14	0.0131 headphone (***tp***)	~~0.0130 option~~ (*fn*)	0.0169 frequency (***tp***)	~~0.0133 screen~~ (*fn*)
15	0.0128 song (***tp***)	0.0126 headphone (***tp***)	0.0150 transmitter (***tp***)	0.0129 format (***tp***)
16	0.0124 quality (***tp***)	~~0.0123 interface~~ (*fn*)	**0.0128 music player** (***tp***)	0.0129 pixel (***tp***)
17	**0.0112 multimedia** (***tp***)	~~0.0116 quality~~ (*fn*)	~~0.0122 quality~~ (*fn*)	~~0.0123 phone~~ (*fn*)
18	0.0108 file (***tp***)	**0.0111 radio** (***tp***)	0.0120 loudspeaker (***tp***)	0.0117 windows (*fp*)
19	0.0102 artist (***tp***)	0.0105 format (***tp***)	~~0.0113 device~~ (*fn*)	0.0112 windows media (***tp***)
20	0.0092 equalizer (***tp***)	0.0104 playlist (***tp***)	**0.0106 fm transmitter** (***tp***)	0.0100 windows media player (***tp***)
21	0.0079 setting (***tp***)	0.0099 media player (***tp***)	0.0097 case (*fp*)	~~0.0087 sound~~ (*fn*)
22	~~0.0078 function~~ (*fn*)	**0.0087 fm** (***tp***)	**0.0090 video player** (***tp***)	~~0.0084 music~~ (*fn*)
23	**0.0073 video player** (***tp***)	0.0086 genre (***tp***)	0.0087 radio station (***tp***)	~~0.0078 interface~~ (*fn*)
24	0.0072 playlist (***tp***)	~~0.0068 function~~ (*fn*)	~~0.0075 car~~ (***tn***)	**0.0078 mpeg4** (***tp***)
25	0.0069 speaker (***tp***)	0.0067 cover (***tp***)	0.0070 need (*fp*)	0.0078 problem (*fp*)
26	~~0.0067 one~~ (***tn***)	0.0067 equalizer (***tp***)	~~0.0064 test~~ (***tn***)	0.0071 codec (***tp***)
27	0.0066 cover (***tp***)	~~0.0066 resolution~~ (*fn*)	0.0059 something (*fp*)	0.0071 video file (***tp***)
28	0.0064 genre (***tp***)	~~0.0064 screen~~ (*fn*)	~~0.0052 version~~ (*fn*)	~~0.0071 headset~~ (*fn*)
29	~~0.0063 resolution~~ (*fn*)	**0.0064 video player** (***tp***)	~~0.0052 capable~~ (***tn***)	0.0068 content (***tp***)
30	0.0062 loudspeaker (***tp***)	~~0.0063 setting~~ (*fn*)	0.0049 noise (***tp***)	**0.0066 h264** (***tp***)

Table 5-40. Judgement applied over the top 30 terms of *topic 10* (without), *topic 10* (without), *topic 55* (without) and *topic 11* (without) after discarding method.

Place	*Topic 12* with *k=50*	*Topic 12* with *k=100*	*Topic 12* with *k=200*	*Topic 14* with *k=200*
1	0.0684 music (***tp***)	**0.0823 video** (***tp***)	0.0447 music (***tp***)	**0.0762 video** (***tp***)
2	0.0557 player (***tp***)	0.0441 player (***tp***)	0.0376 player (***tp***)	0.0378 player (***tp***)
3	**0.0379 video** (***tp***)	0.0304 playback (***tp***)	**0.0335 video** (***tp***)	0.0358 playback (***tp***)
4	**0.0225 music player** (***tp***)	0.0268 music (***tp***)	0.0299 album (***tp***)	0.0271 resolution (***tp***)
5	~~0.0195 phone~~ (*fn*)	0.0237 file (***tp***)	~~0.0259 phone~~ (*fn*)	**0.0244 multimedia** (***tp***)
6	0.0190 file (***tp***)	0.0235 quality (***tp***)	**0.0204 radio** (***tp***)	**0.0238 divx** (***tp***)
7	0.0178 sound (***tp***)	**0.0228 multimedia** (***tp***)	~~0.0186 option~~ (*fn*)	~~0.0226 quality~~ (*fn*)
8	0.0175 album (***tp***)	0.0226 resolution (***tp***)	0.0177 artist (***tp***)	**0.0212 xvid** (***tp***)
9	**0.0155 radio** (***tp***)	**0.0206 divx** (***tp***)	0.0174 track (***tp***)	~~0.0187 file~~ (*fn*)
10	**0.0147 multimedia** (***tp***)	**0.0190 audio** (***tp***)	~~0.0163 interface~~ (*fn*)	**0.0170 audio** (***tp***)
11	0.0140 song (***tp***)	**0.0151 xvid** (***tp***)	**0.0161 fm** (***tp***)	0.0166 pixel (***tp***)
12	0.0134 quality (***tp***)	**0.0145 video playback** (***tp***)	~~0.0152 file~~ (*fn*)	**0.0160 video playback** (***tp***)
13	**0.0134 fm** (***tp***)	0.0143 sound (***tp***)	**0.0144 music player** (***tp***)	~~0.0146 phone~~ (*fn*)
14	0.0132 headphone (***tp***)	~~0.0138 phone~~ (*fn*)	0.0139 playlist (***tp***)	~~0.0114 sound~~ (*fn*)
15	0.0127 track (***tp***)	0.0118 pixel (***tp***)	0.0138 headphone (***tp***)	**0.0113 fm** (***tp***)
16	~~0.0118 option~~ (*fn*)	0.0114 support (***tp***)	~~0.0129 sound~~ (*fn*)	~~0.0104 music~~ (*fn*)
17	0.0111 media (***tp***)	~~0.0111 screen~~ (*fn*)	0.0122 song (***tp***)	**0.0099 radio** (***tp***)
18	~~0.0105 interface~~ (*fn*)	**0.0103 radio** (***tp***)	**0.0119 multimedia** (***tp***)	0.0094 capable (*fp*)
19	0.0104 artist (***tp***)	**0.0101 video player** (***tp***)	0.0116 equalizer (***tp***)	0.0088 earphone (***tp***)
20	0.0092 fm radio (***tp***)	**0.0098 fm** (***tp***)	0.0115 fm radio (***tp***)	0.0088 content (***tp***)
21	0.0082 playback (***tp***)	**0.0087 music player** (***tp***)	0.0114 genre (***tp***)	~~0.0086 headset~~ (*fn*)
22	0.0080 format (***tp***)	0.0084 album (***tp***)	0.0095 cover (***tp***)	~~0.0086 screen~~ (*fn*)
23	0.0079 playlist (***tp***)	0.0079 codec (***tp***)	**0.0090 h263** (***tp***)	**0.0085 video player** (***tp***)
24	0.0075 media player (***tp***)	0.0077 clip (***tp***)	~~0.0083 quality~~ (*fn*)	~~0.0082 support~~ (*fn*)
25	0.0073 resolution (***tp***)	~~0.0070 option~~ (*fn*)	~~0.0080 setting~~ (*fn*)	0.0077 pair (*fp*)
26	0.0070 equalizer (***tp***)	0.0070 jack (***tp***)	**0.0080 video player** (***tp***)	0.0075 advertisment (*fp*)
27	0.0069 genre (***tp***)	0.0069 advertisment (*fp*)	**0.0080 mpeg4** (***tp***)	0.0073 album (***tp***)
28	**0.0065 video player** (***tp***)	**0.0067 mpeg4** (***tp***)	~~0.0076 function~~ (*fn*)	0.0073 loudspeaker (***tp***)
29	~~0.0065 function~~ (*fn*)	0.0067 audio player (***tp***)	0.0076 station (***tp***)	0.0073 clip (***tp***)
30	0.0063 setting (***tp***)	0.0066 headphone (***tp***)	~~0.0072 one~~ (***tn***)	~~0.0073 image~~ (*fn*)

Table 5-41. Judgement applied over the top 30 terms of *topic 12* (with), *topic 12* (with), *topic 14* (with) and *topic 12* (with) after discarding method.



Graphic 5-23. Precision and recall graph of selected topics of *multimedia* on sections.

From ***topic 10*** **without** *extra reviews* with **k=50**, terms such as *music*, *player*, *album*, *sound*, *track*, *fm radio*, *option*, *headphone*, *song*, *quality*, *file*, *artist*, *equalizer*, *setting*, *playlist*, *speaker*, *cover*, *genre* and *loudspeaker* are extracted with 100% of precision, 86.67% accuracy, and a 86.26% of recall. From ***topic 10*** **without** *extra reviews* with **k=100**, terms such as *music*, *player*, *file*, *audio*, *song*, *media*, *sound*, *artist*, *track*, *headphone*, *format*, *playlist*, *media player*, *genre*, *cover* and *equalizer* are extracted with 100% of precision, 75.86% accuracy, and a 73.33% of recall. From ***topic 55*** **without** *extra reviews*, terms such as *fm radio*, *player*, *music*, *station*, *sound*, *headphone*, *antenna*, *frequency*, *transmitter*, *loudspeaker* and *noise* are extracted with 86.36% of precision, 70% accuracy, and a 76% of recall. From ***topic 11*** **without** *extra reviews*, terms such as *player*, *file*, *media*, *playback*, *media player*, *resolution*, *format*, *pixel*, *windows media*, *windows media player*, *codec*, *video file* and *content* are extracted with 91.30% of precision, 70% accuracy, and a 75% of recall. From ***topic 12*** **with** *extra reviews* with **k=50**, terms such as *music*, *player*, *file*, *sound*, *album*, *song*, *quality*, *headphone*, *track*, *media*, *artist*, *fm radio*, *playback*, *format*, *playlist*, *media player*, *resolution*, *equalizer*, *genre* and *setting* are extracted with 100% of precision, 86.67% accuracy, and a 86.67% of recall. From ***topic 12*** **with** *extra reviews* with **k=100**, terms such as *player*, *playback*, *music*, *file*, *quality*, *resolution*, *sound*, *pixel*, *support*, *album*, *codec*, *clip*, *jack*, *audio player* and *headphone* are extracted with 96.30% of precision, 86.67% accuracy, and a 89.66% of recall. From ***topic 12*** **with** *extra reviews* with **k=200**, terms such as *music*, *player*, *album*, *artist*, *track*, *playlist*, *headphone*, *song*, *equalizer*, *fm radio*, *genre*, *cover* and *station* are extracted with 100% of precision, 72.41% accuracy, and a 72.41% of recall. From ***topic 14*** **with** *extra reviews*, terms such as *player*, *playback*, *resolution*, *pixel*, *earphone*, *content*, *album*, *loudspeaker* and *clip* are extracted with 85.71% of precision, 62.07% accuracy, and a 66.67% of recall.

5.4.3.3. Paragraphs

Finally, LDA's performances are analyzed on paragraphs scenario. LSI performs very well on this section, but PLSI does not achieve what it was supposed, then, following the same steps described in **Table 4-1**, LDA is run on paragraphs with the expectation of its improvement over, at least PLSI, but also LSI.

In **Table 11-9**, a wide selection of the best topics where *multimedia* initial product features are better scored and placed within the top 30 terms of the sorted topics is found. From it, a little selection is done considering the whole 30 terms, and starting the process named in the steps. After that, discarding method is applied getting many wrong discarded terms in topics without *extra reviews*. Single results of this application for each topic are found in **Annex F: Discarding method applied to LDA topics**. Then, the human judgement is applied on the results given by the discarding method, and both applications are shown in **Table 5-42** and **Table 5-43**.

Summarizing those judgements graphically, it is possible to see how in GRAPHIC where precision and recall are presented. Moreover, single graphics of accuracy, precision and recall are located in **Annex I: Precision, recall and accuracy graphs of LDA**. *Topic 35* seems to be the best performance, because its recall reaches the highest value. However, all of them remain

high over the 80% of precision during the first 30 terms, where *topic 56* has the poorest performance, although it remains high. As it is done before, results are summarized below.

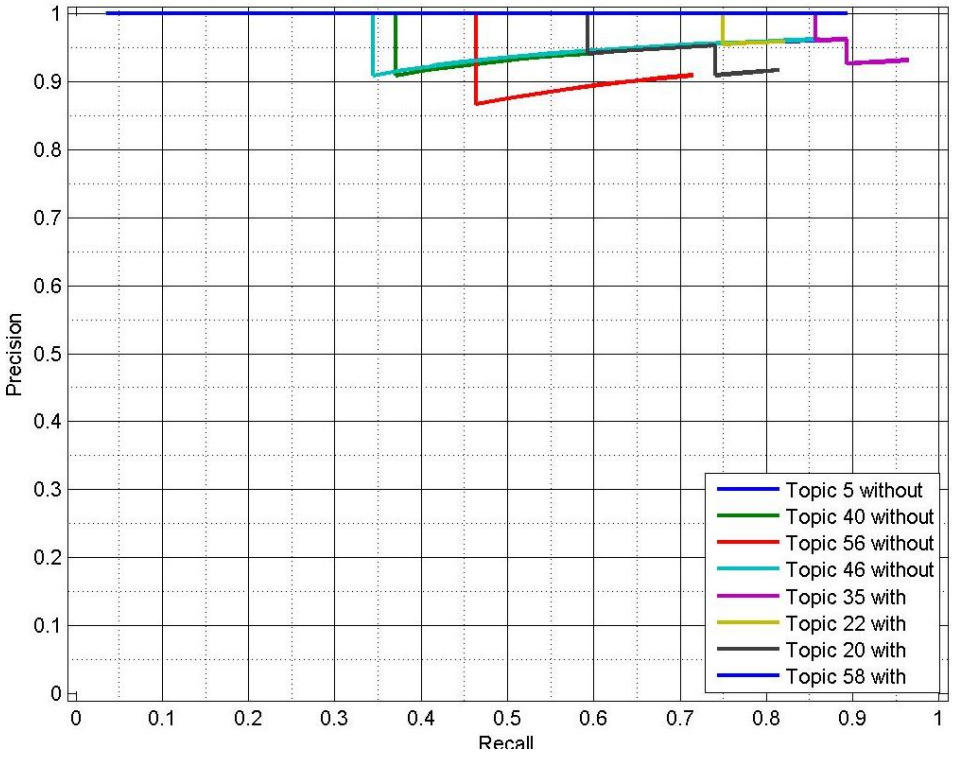
From ***topic 5*** **without** *extra reviews*, terms such *music*, *player*, *album*, *song*, *track*, *media*, *option*, *interface*, *artist*, *file*, *media player*, *art*, *equalizer*, *playlist*, *playback*, *genre*, *cover*, *walkman* and *play* are extracted with 96% of precision, 83.33% accuracy, and a 85.71% of recall. From ***topic 40*** **without** *extra reviews*, terms such as *player*, *media*, *media player*, *music*, *format*, *window*, *playback*, *windows media*, *picture*, *windows media player*, *photo*, *movie* and *media menu* are extracted with 95.24% of precision, 84% accuracy, and a 74.07% of recall. From ***topic 56*** **without** *extra reviews*, terms such *file*, *resolution*, *player*, *format*, *clip*, *pixel*, *video file*, *movie*, *support*, *playback*, *frame*, *video format*, *video clip*, *view* and *qcif* are extracted with 90.91% of precision, 66.67% accuracy, and a 71.43% of recall. From ***topic 46*** **without** *extra reviews*, terms such as *player*, *file*, *media*, *format*, *media player*, *album*, *windows media*, *windows media player*, *song*, *playlist*, *artist*, *genre*, *multimedia player*, *video file*, *playback* and *library* are extracted with 96.15% of precision, 83.33% accuracy, and a 86.21% of recall. From ***topic 35*** **with** *extra reviews*, terms such as *resolution*, *pixel*, *file*, *playback*, *sample*, *quality*, *clip*, *pixel resolution*, *frame*, *player*, *format*, *capture*, *qvga*, *vga*, *content*, *qvga resolution*, *support*, *video file*, *camcorder*, *movie* and *video clip* are extracted with 93.10% of precision, 90% accuracy, and a 96.43% of recall.

Place	*Topic 5* without *k=50*	*Topic 40* without *k=100*	*Topic 56* without *k=150*	*Topic 46* without *k=200*
1	0.0945 music (***tp***)	**0.2234 video** (***tp***)	**0.2338 video** (***tp***)	0.0863 player (***tp***)
2	0.0895 player (***tp***)	0.1427 player (***tp***)	0.0802 file (***tp***)	**0.0632 video** (***tp***)
3	**0.0350 music player** (***tp***)	0.0891 media (***tp***)	0.0606 resolution (***tp***)	0.0552 file (***tp***)
4	0.0257 album (***tp***)	0.0489 media player (***tp***)	0.0576 player (***tp***)	~~0.0468 music~~ (*fn*)
5	0.0225 song (***tp***)	0.0457 music (***tp***)	0.0467 format (***tp***)	0.0403 media (***tp***)
6	0.0225 track (***tp***)	**0.0398 video player** (***tp***)	**0.0384 mpeg4** (***tp***)	0.0346 format (***tp***)
7	0.0223 media (***tp***)	0.0352 format (***tp***)	**0.0355 video player** (***tp***)	0.0310 media player (***tp***)
8	**0.0203 video** (***tp***)	0.0249 window (***tp***)	0.0287 clip (***tp***)	0.0264 album (***tp***)
9	0.0195 option (***tp***)	0.0248 playback (***tp***)	**0.0282 h263** (***tp***)	0.0243 windows media (***tp***)
10	~~0.0163 phone~~ (*fn*)	**0.0248 multimedia** (***tp***)	~~0.0261 quality~~ (*fn*)	**0.0219 mp3** (***tp***)
11	0.0150 interface (***tp***)	0.0210 content (*fp*)	~~0.0257 screen~~ (*fn*)	0.0216 windows media player (***tp***)
12	0.0147 artist (***tp***)	0.0210 windows media (***tp***)	~~0.0202 display~~ (*fn*)	0.0215 windows (*fp*)
13	0.0145 file (***tp***)	0.0209 picture (***tp***)	0.0197 pixel (***tp***)	**0.0206 aac** (***tp***)
14	**0.0130 audio** (***tp***)	~~0.0203 screen~~ (*fn*)	0.0195 video file (***tp***)	0.0199 song (***tp***)
15	0.0121 media player (***tp***)	0.0192 window media player (***tp***)	~~0.0177 phone~~ (*fn*)	**0.0185 wma** (***tp***)
16	0.0113 art (***tp***)	**0.0142 audio** (***tp***)	0.0158 movie (***tp***)	0.0181 playlist (***tp***)
17	0.0111 equalizer (***tp***)	~~0.0112 problem~~ (***tn***)	0.0156 support (***tp***)	**0.0178 multimedia** (***tp***)
18	0.0111 playlist (***tp***)	0.0105 photo (***tp***)	~~0.0148 image~~ (*fn*)	**0.0174 wav** (***tp***)
19	**0.0105 multimedia** (***tp***)	~~0.0094 quality~~ (*fn*)	0.0139 problem (*fp*)	0.0170 artist (***tp***)
20	~~0.0092 function~~ (*fn*)	**0.0088 mpeg4** (***tp***)	~~0.0128 sound~~ (*fn*)	0.0161 genre (***tp***)
21	0.0091 playback (***tp***)	**0.0087 mp3** (***tp***)	0.0127 kbps (*fp*)	0.0133 multimedia player (***tp***)
22	0.0091 genre (***tp***)	**0.0083 wmv** (***tp***)	0.0120 playback (***tp***)	0.0131 video file (***tp***)
23	0.0069 cover (***tp***)	~~0.0081 version~~ (*fn*)	**0.0114 wmv** (***tp***)	**0.0129 audio** (***tp***)
24	0.0066 information (*fp*)	0.0080 movie (***tp***)	~~0.0113 music~~ (*fn*)	0.0117 playback (***tp***)
25	0.0064 walkman (***tp***)	0.0069 media menu (***tp***)	~~0.0110 picture~~ (*fn*)	~~0.0116 type~~ (*fn*)
26	**0.0060 mp3** (***tp***)	~~0.0064 image~~ (*fn*)	0.0096 frame (***tp***)	~~0.0114 function~~ (*fn*)
27	~~0.0058 setting~~ (*fn*)	~~0.0054 place~~ (***tn***)	0.0094 video format (***tp***)	0.0111 library (***tp***)
28	~~0.0058 feature~~ (***tn***)	~~0.0053 function~~ (*fn*)	0.0090 video clip (***tp***)	~~0.0110 headphone~~ (*fn*)
29	~~0.0057 menu~~ (*fn*)	~~0.0047 sound~~ (*fn*)	0.0076 view (***tp***)	**0.0106 video player** (***tp***)
30	0.0056 play (***tp***)	~~0.0043 handset~~ (*fn*)	0.0069 qcif (***tp***)	**0.0103 aac+** (***tp***)

Table 5-42. Judgement applied over the top 30 terms of *topic 5* (without), *topic 40* (without), *topic 56* (without) and *topic 46* (without) after discarding method.

Place	*Topic 35* with *k=50*	*Topic 22* with *k=100*	*Topic 20* with *k=150*	*Topic 58* with *k=200*
1	**0.3476 video** (***tp***)	**0.1561 video** (***tp***)	**0.1671 video** (***tp***)	0.1000 player (***tp***)
2	0.0955 resolution (***tp***)	0.0527 file (***tp***)	0.0527 resolution (***tp***)	**0.0761 video** (***tp***)
3	0.0590 pixel (***tp***)	0.0475 player (***tp***)	0.0471 file (***tp***)	0.0741 file (***tp***)
4	0.0403 file (***tp***)	0.0383 resolution (***tp***)	**0.0430 divx** (***tp***)	0.0664 media (***tp***)
5	0.0290 playback (***tp***)	0.0257 media (***tp***)	0.0399 player (***tp***)	0.0573 music (***tp***)
6	0.0270 sample (***tp***)	0.0253 playback (***tp***)	**0.0353 xvid** (***tp***)	0.0536 media player (***tp***)
7	0.0257 quality (***tp***)	**0.0241 divx** (***tp***)	0.0316 playback (***tp***)	0.0434 format (***tp***)
8	0.0244 clip (***tp***)	0.0237 format (***tp***)	~~0.0272 quality~~ (*fn*)	0.0410 windows (*fp*)
9	0.0233 pixel resolution (***tp***)	0.0208 pixel (***tp***)	**0.0243 mpeg4** (***tp***)	0.0360 windows media (***tp***)
10	0.0222 frame (***tp***)	~~0.0187 quality~~ (*fn*)	0.0238 pixel (***tp***)	**0.0334 mp3** (***tp***)
11	**0.0207 video player** (***tp***)	**0.0187 xvid** (***tp***)	**0.0232 video playback** (***tp***)	0.0326 window media player (***tp***)
12	0.0194 player (***tp***)	**0.0165 video player** (***tp***)	0.0231 support (***tp***)	**0.0257 wav** (***tp***)
13	**0.0186 video playback** (***tp***)	0.0165 media player (***tp***)	0.0218 format (***tp***)	0.0227 song (***tp***)
14	**0.0167 mpeg4** (***tp***)	0.0158 support (***tp***)	**0.0218 video player** (***tp***)	0.0215 playlist (***tp***)
15	0.0164 format (***tp***)	**0.0155 mpeg4** (***tp***)	~~0.0184 screen~~ (*fn*)	0.0215 artist (***tp***)
16	0.0159 capture (***tp***)	~~0.0148 screen~~ (*fn*)	**0.0180 h263** (***tp***)	**0.0188 aac** (***tp***)
17	0.0150 qvga (***tp***)	**0.0135 video playback** (***tp***)	**0.0175 h264** (***tp***)	**0.0184 multimedia** (***tp***)
18	0.0139 vga (***tp***)	0.0128 clip (***tp***)	~~0.0161 phone~~ (*fn*)	0.0179 playback (***tp***)
19	**0.0113 h263** (***tp***)	0.0121 codec (***tp***)	~~0.0153 display~~ (*fn*)	0.0163 headphone (***tp***)
20	0.0102 content (***tp***)	**0.0108 multimedia** (***tp***)	0.0141 codec (***tp***)	0.0159 type (***tp***)
21	~~0.0102 screen~~ (*fn*)	**0.0103 h263** (***tp***)	0.0137 content (*fp*)	**0.0156 h263** (***tp***)
22	0.0088 qvga resolution (***tp***)	**0.0102 h264** (***tp***)	**0.0130 multimedia** (***tp***)	0.0129 information (*fp*)
23	0.0087 support (***tp***)	0.0100 video file (***tp***)	~~0.0129 image~~ (*fn*)	**0.0125 video player** (***tp***)
24	0.0085 video file (***tp***)	0.0098 content (*fp*)	**0.0113 mp4** (***tp***)	~~0.0121 sound~~ (*fn*)
25	0.0083 camcorder (***tp***)	~~0.0092 image~~ (*fn*)	0.0103 clip (***tp***)	~~0.0114 quality~~ (*fn*)
26	0.0077 problem (*fp*)	0.0084 windows media (***tp***)	0.0095 video file (***tp***)	**0.0112 mpeg4** (***tp***)
27	0.0076 movie (***tp***)	~~0.0083 phone~~ (*fn*)	0.0091 capable (*fp*)	~~0.0091 resolution~~ (*fn*)
28	0.0070 capable (*fp*)	~~0.0082 problem~~ (***tn***)	~~0.0087 problem~~ (***tn***)	0.0087 shuffle (***tp***)
29	**0.0068 mp4** (***tp***)	0.0078 window media player (***tp***)	0.0081 qvga (***tp***)	~~0.0065 one~~ (***tn***)
30	0.0068 video clip (***tp***)	~~0.0075 display~~ (*fn*)	0.0078 movie (***tp***)	~~0.0058 way~~ (***tn***)

Table 5-43. Judgement applied over the top 30 terms of *topic 35* (with), *topic 22* (with), *topic 20* (with) and *topic 58* (with) after discarding method.



Graphic 5-24. Precision and recall graph of selected topics of *multimedia* on sections.

From ***topic 22*** **with** *extra reviews*, terms such as *file*, *player*, *resolution*, *media*, *playback*, *format*, *pixel*, *media player*, *support*, *clip*, *codec*, *video file*, *windows media* and *windows media player* are extracted with 95.83% of precision, 82.76% accuracy, and a 82.14% of recall. From ***topic 20*** **with** *extra reviews*, terms such as *resolution*, *file*, *player*, *playback*, *pixel*, *support*, *format*, *codec*, *clip*, *video file*, *qvga* and *movie* are extracted with 91.67% of precision, 76.67% accuracy, and a 81.48% of recall. From ***topic 58*** **with** *extra reviews*, terms such as *player*, *file*, *media*, *music*, *media player*, *format*, *windows media*, *windows media player*, *song*, *playlist*, *artist*, *playback*, *headphone*, *type* and *shuffle* are extracted with 100% of precision, 90% accuracy, and a 89.29% of recall.

6. Results

In this chapter it is summarized the whole results from all the analysis covered separated by each main product feature and each scenario. Although results from manual analysis are shown above only graphically, it is possible to establish the same pattern like in the rest of analysis. If it is considered that the terms restrieved are the new terms extracted (not those terms similar to initial product features, such as *hd video playback*), they suppose the complete recall, and due to they are all correctly extracted, all measurements of precision, accuracy and recall achieve the 100%. In the following sections, a comparision between all results is covered.

6.1. Summary and performance comparision on *battery* product feature

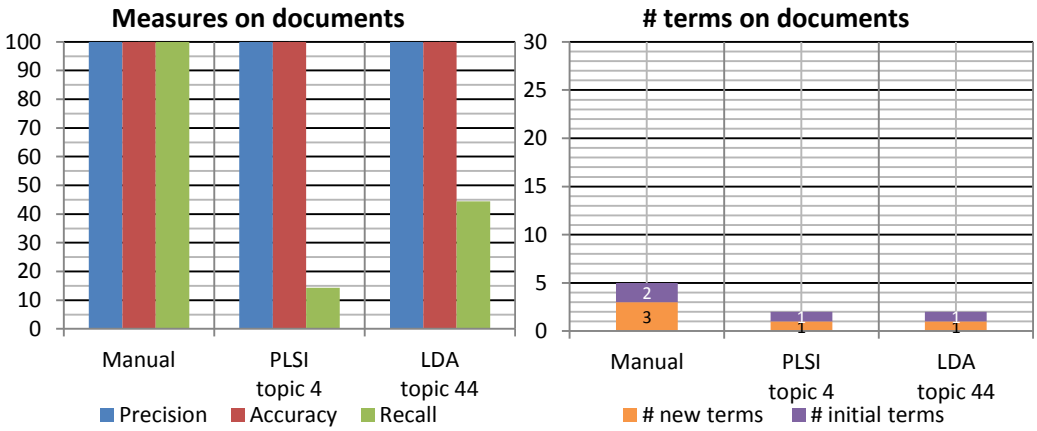
It is grouped here the entire collection of results around the *battery* product feature obtained in all the analysis carried out before. It is summarized all the data retrieved from the analysis and it is shown in graphics to understand it better and extract right conclusions.

In **Table 6-1** it is talk about the manual analysis, where any scenario was regarded, and because of that, it is not at all comparable with the rest performances on each scenario, then, it is compared with all to see the differences. Only *battery* and *talk time* are the initial temrs that appear in the mini corpus analyzed manually, then, they are considered as if they appear in the top 30 terms, like with automatic methods. It is also seen the single results obtained in the manual analysis, where the highest values of precision, accuracy and recall are reached, but with only 3 new terms. If it is taken a look, it can be seen how LSI performances looking for the *battery* dimension/topic are reduced to an only one. *Dimension 17* is the only on the paragraphs scenario. In the same line, it is shown the PLSI where appears only a single topic, *topic 4*, but this time in documents scenario. However, LDA achieves the best performances regarding the *battery* product feature, improving a little on documents scenario, but increasing considerable the number of terms extracted with high values of precision, accuracy and recall.

After that, results from analysis are compared on each scenario, where manual, LSI, PLSI and LDA results are compared. At first, although it is not comparable at all due to the lack of results on documents, in **Graphic 6-1** it is shown the only two performances found with significant results compared with the manual ones. It is seen how PLSI and LDA performances only differ in recall values, where LDA improves considerably PLSI values, but it still remain lower than 50%. Anyway, LDA performance is the best performance regarding the *battery* product feature on documents scenario.

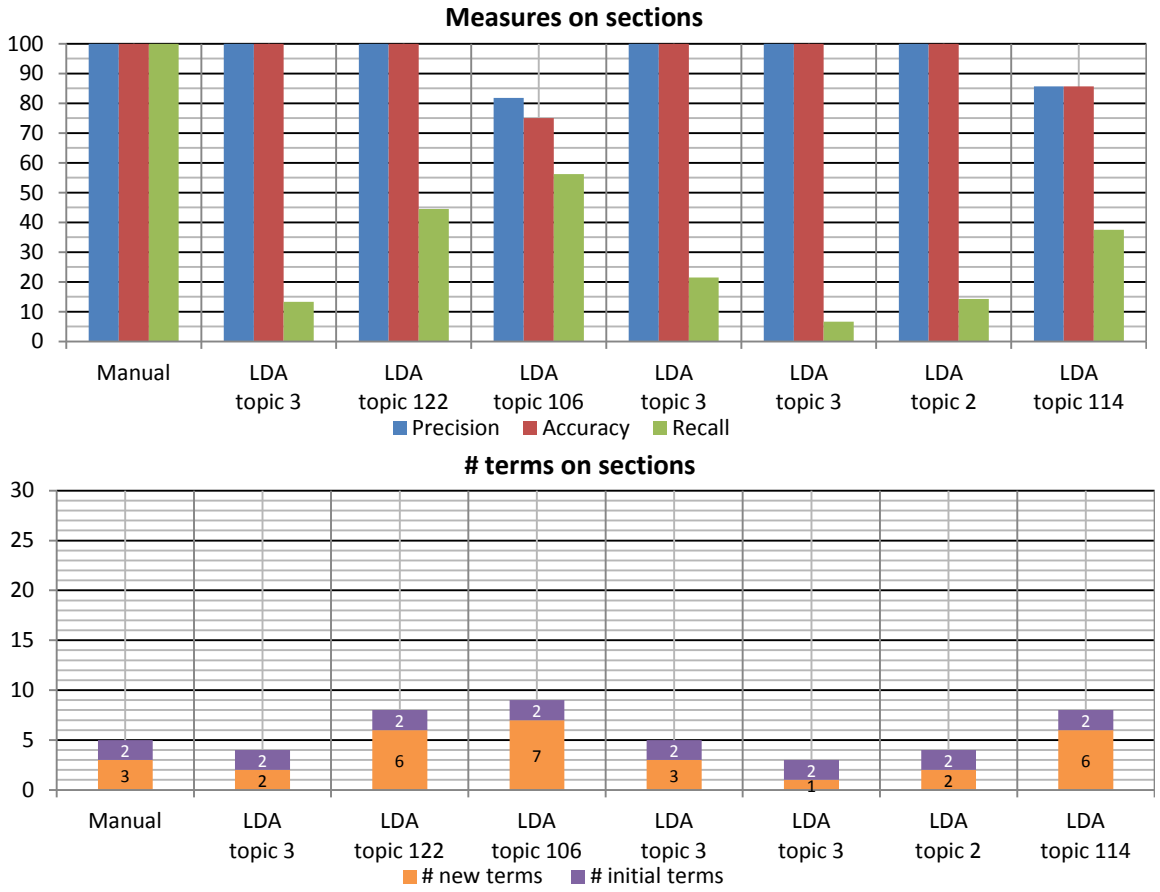
battery									
Method	Scenario	Topic	Mode	k parameter	Precision (%)	Accuracy (%)	Recall (%)	# new terms	# initial terms
Manual	context	-	without	-	100	100	100	3	2
LSI	documents	No dimensions	-	-	-	-	-	-	-
	sections	No dimensions	-	-	-	-	-	-	-
	paragraphs	*dimension 17*	without	100	100	100	41.18	5	2
PLSI	documents	*topic 4*	without	200	100	100	14.29	1	1
	sections	No topics	-	-	-	-	-	-	-
	paragraphs	No topics	-	-	-	-	-	-	-
LDA	documents	*topic 44*	with	100	100	100	20	1	1
	sections	*topic 3*	without	100	100	100	13.33	2	2
		topic 122	without	150	100	100	44.44	6	2
		topic 106	without	200	81.82	75	56.25	7	2
		topic 3	with	50	100	100	21.43	3	2
		topic 3	with	100	100	100	6.67	1	2
		topic 2	with	150	100	100	14.29	2	2
		topic 114	with	200	85.71	85.71	37.50	6	2
	paragraphs	*topic 26*	without	50	87.50	77.78	63.64	12	2
		topic 43	without	100	100	83.33	58.82	8	2
		topic 62	without	100	100	100	14.29	1	1
		topic 86	without	150	86.67	76.19	72.22	11	2
		topic 27	with	50	82.35	73.68	70	12	2
		topic 48	with	100	86.36	82.14	86.36	16	3
		topic 35	with	150	89.47	76.92	77.27	15	2

Table 6-1. Summary of the obtained results regarding on *battery*.



Graphic 6-1. Summarization of calculated meausres and terms of *battery* on documents.

In **Graphic 6-2** it is shown the graphics which correspond to the sections scenario. It is noticed that, as it is seen in the analysis, there are more performances to compare, but there are all LDA performances, then it is simply concluded that LDA performance is the best one regarding the *battery* product feature on sections. However, taking a look at the values obtained, if it is taken the number of new terms extracted to be the priority to measure, it is possible to decide that *topic 122*, *topic 106* and *topic 114* in this order, are the best representations of LDA, getting up to 7, 6 and 6 new terms respectively.

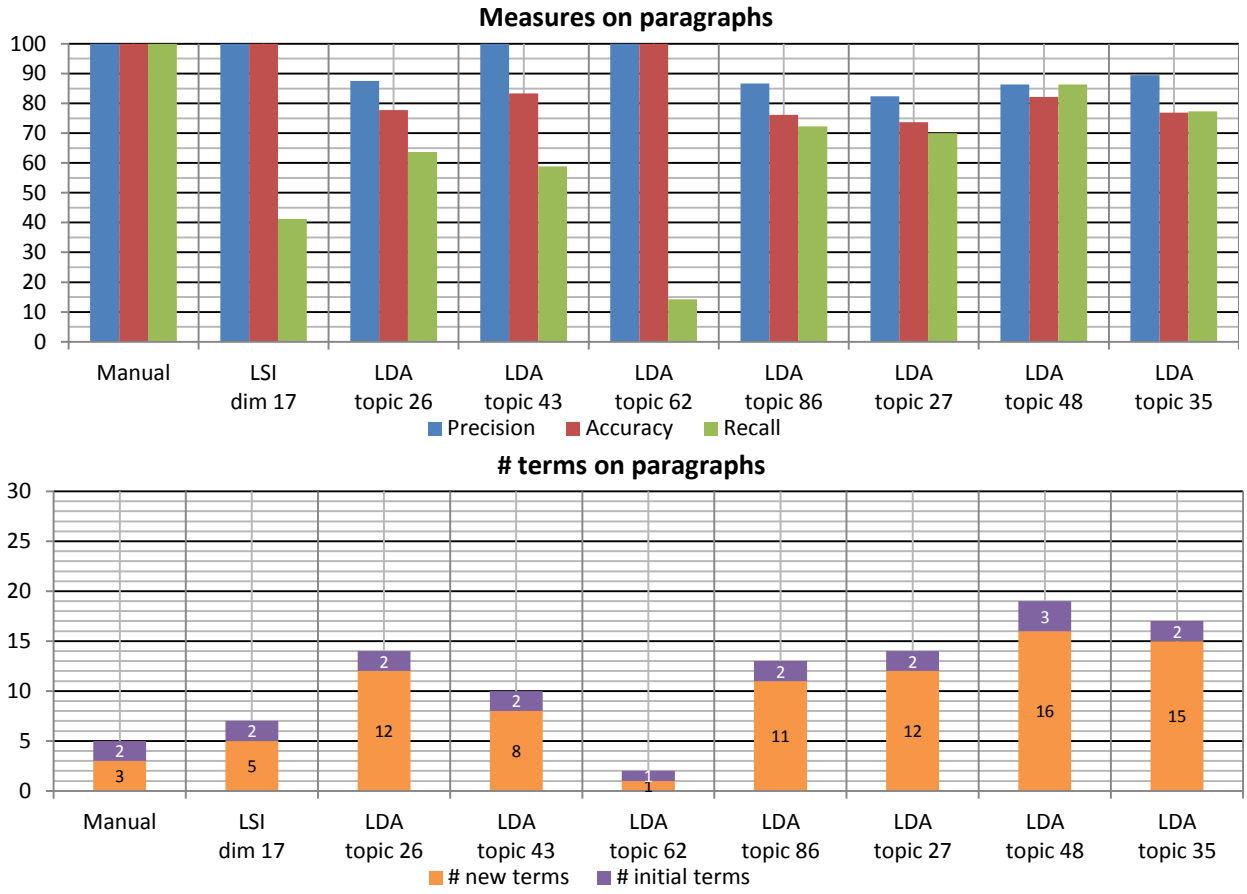


Graphic 6-2. Summarization of measures and extracted terms of *battery* on sections.

Finally, *battery* product feature analysis results are compared on paragraphs scenario. **Graphic 6-3** shows performance comparision and the number of terms respectively, from those LSI and LDA dimension and topics where there are results susceptible to be considered.

There, it is seen that LDA *topic 48* seems to be the best performance, because ommtting the fact that all of the satisfy the threshold of achieving a precision higher than 80%, it also gets more than 80% of recall, and this leads it to have the highest number of new terms extracted with the highest number of initial terms within its top 30 terms. *Topic 35* is also considerably better than the rest in performance and number of terms.

LDA remains as the best information retrieval technique to extract new *battery* terms. It also increases considerably manual efforts and at least triples LSI results, getting up to 16 new terms semantically related on its best performance. Moreover, it doubles the semantic initial group, on paragraphs scenario. Finally, LDA leads the semantic relationship extraction in reviews from initial product features of technical specifications related with the *battery* product feature on documents, sections and paragraphs scenarios.

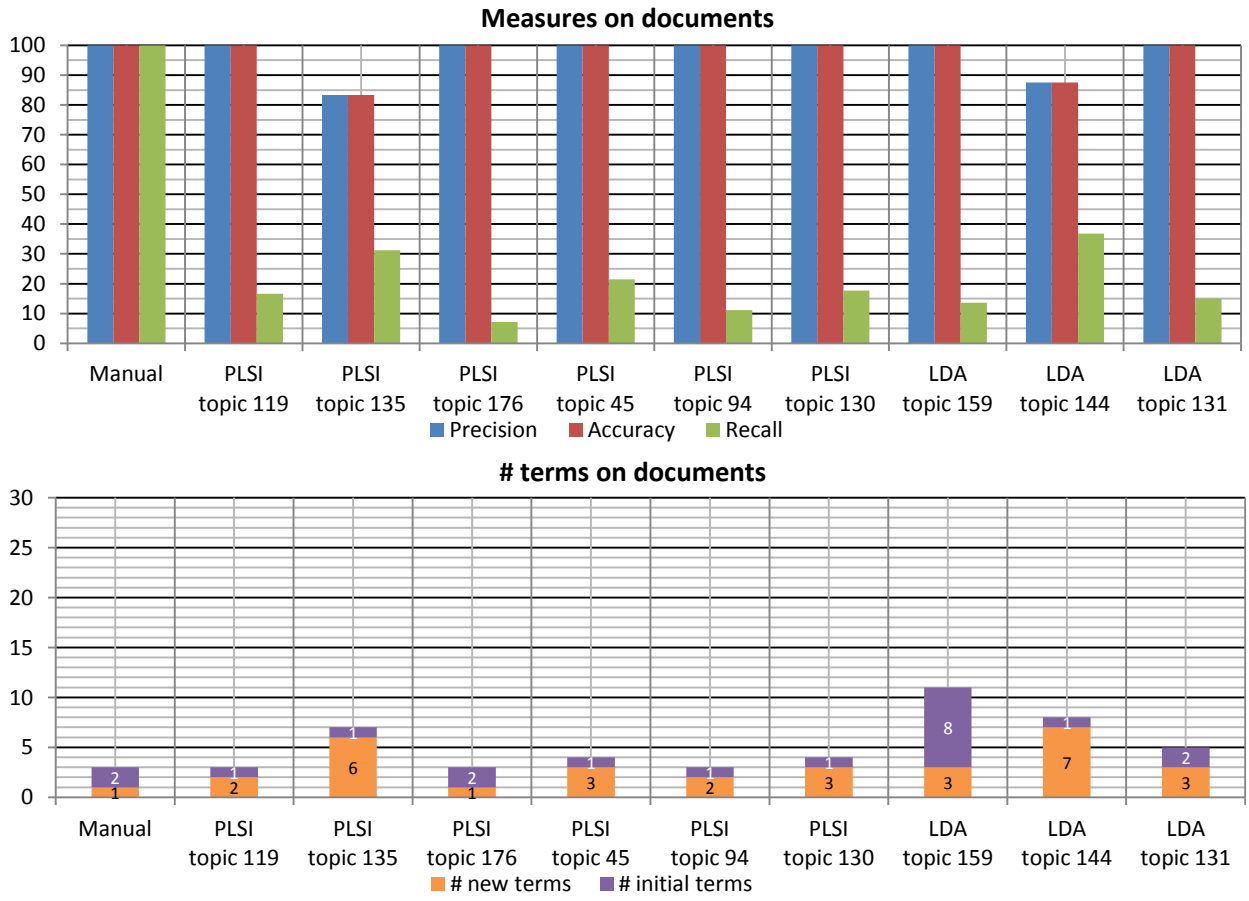


Graphic 6-3. Summarization of measures and extracted terms of *battery* on paragraphs.

6.2. Summary and performance comparision on *organizer* product feature

Results taken from all the analysis of the *organizer* product feature are summarized and compared here to extract better conclusions. It is grouped all the data retrieved by the whole process of the automatic extraction of the new semantically related product features.

In **Table 6-2** it is seen the data resulting from the manual, LSI, PLSI and LDA analysis. Manual analysis is considered here to compare how difficult is to get some significant results after hours of review comprobation and how with automatic techniques results are improved considerably. Manual results seem to be not scalable, because the more the number of reviews analysed is, the less the percentage of being in both way semantically related a term pair is. It is also seen the single results obtained in the manual analysis, where the highest values of precision, accuracy and recall are reached, but with only 1 new term is extracted. Unlike it happens with the *battery* product feature analysis, here it is checked that the only pair method-scenario where any dimension or topic is extracted occurs with LSI on documents, whose values were as bad as not to be extracted any new term. Although LDA fits best for *battery*, here LSI also gets incredible performances despite of being the less rated to do it. It was supposed to see an evolution on the three methods, but in the end LSI and LDA have performed almost pefect with some differences.



Graphic 6-4. Summarization of measures and extracted terms of *organizer* on documents.

organizer									
Method	Scenario	Topic	Mode	k parameter	Precision (%)	Accuracy (%)	Recall (%)	# new terms	# initial terms
Manual	context	-	without	-	100	100	100	1	6
LSI	documents	No dimensions	-	-	-	-	-	-	-
	sections	*dimension 4*	without	50	96.55	96.55	100	21	7
			with	50,150	96.67	96.67	100	21	8
		dimension 8	without	200	91.67	89.66	95.65	16	6
			with	150	82.35	80.95	66.67	9	6
		dimension 34	with	100,150	88.89	88.89	40	4	3
	paragraphs	*dimension 2*	without	50,100	85.71	80	92.31	17	6
			with	50	86.67	86.67	100	19	6
		dimension 9	without	150	100	90	86,96	11	8
			with	150	95,24	90	90,91	11	9
		dimension 17	without	50,100,200	100	100	10	2	6
			with	50,100,150,200	100	100	11.11	3	6
		dimension 37	with	100	100	100	6.25	1	6
PLSI	documents	*topic 119*	without	150	100	100	16.67	2	1
		topic 135	without	150	83.33	83.33	31.25	6	1
		topic 176	without	200	100	100	7.14	1	2
		topic 45	with	50	100	100	21.43	3	1
		topic 94	with	150	100	100	11.11	2	1
		topic 130	with	200	100	100	17.65	3	1
	sections	*topic 28*	without	50	100	100	5.88	1	1
		topic 6	without	100	100	100	6.67	1	1
		topic 49	without	100	100	100	12.50	2	2
		topic 60	without	150	87.50	87.50	50	4	3
		topic 154	without	200	87.50	87.50	43.75	6	2
		topic 11	with	100	81.82	81.82	50	8	1
		topic 164	with	200	88.89	88.89	47.06	8	2
	paragraphs	*topic 2*	without	100	100	100	15.38	2	1
		topic 107	with	150	100	100	11.11	1	1
LDA	documents	*topic 159*	without	200	100	100	13.64	3	8
		topic 144	with	150	87.50	87.50	36.84	7	1
		topic 131	with	200	100	100	15	3	2
	sections	*topic 43*	without	50	93.10	90	96.43	18	9
		topic 35	without	100	100	93.33	93.10	16	10
		topic 36	without	100	96.15	86.67	89.29	19	6
		topic 31	without	200	96	80	82.76	14	10
		topic 54	without	200	86.36	73.33	79.17	10	9
		topic 8	with	50	93.10	90	96.43	17	10
		topic 33	with	100	96.43	90	93.10	17	10
		topic 31	with	200	84.62	73.33	84.62	16	5
	paragraphs	*topic 4*	without	50	92.86	86.67	92.86	17	9
		topic 3	without	100	85.71	80	92.31	20	4
		topic 60	without	150	96.43	90	93.10	22	4
		topic 3	with	50	93.10	90	96.43	18	9
		topic 3	with	100	84.62	81.48	95.65	18	4
		topic 65	with	100	92.31	86.21	88.89	13	11
		topic 123	with	150	100	85.71	54.55	6	8
		topic 99	with	200	84.62	76.67	88	17	5

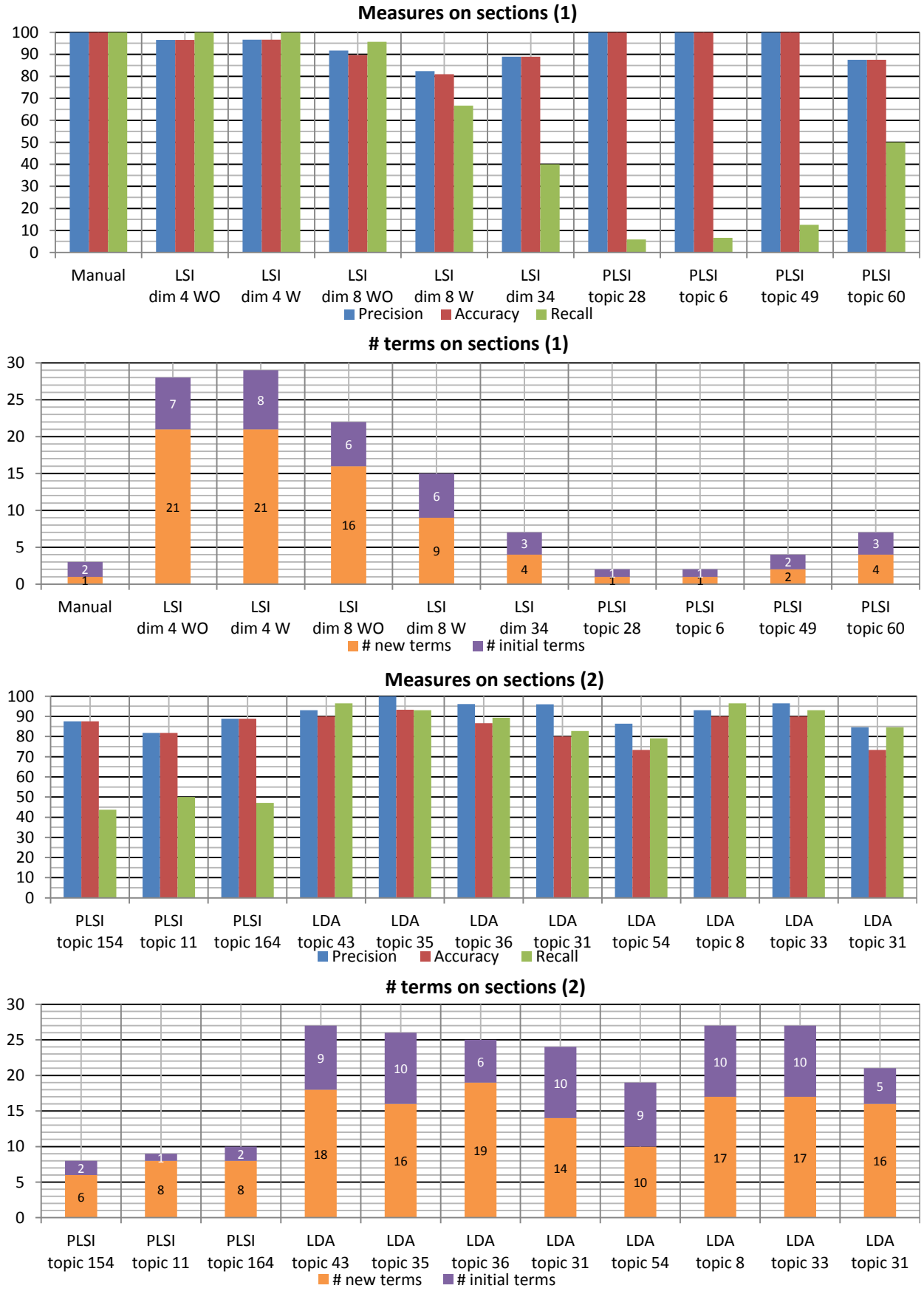
Table 6-2. Summary of the obtained results regarding on *organizer*.

Once results are summarized, they are better compared graphically, where all methods are rated on each single scenario. In **Graphic 6-4**, it is seen how difficult has been to extract some relevant results from documents. Otherwise, it is presented PLSI and LDA results where

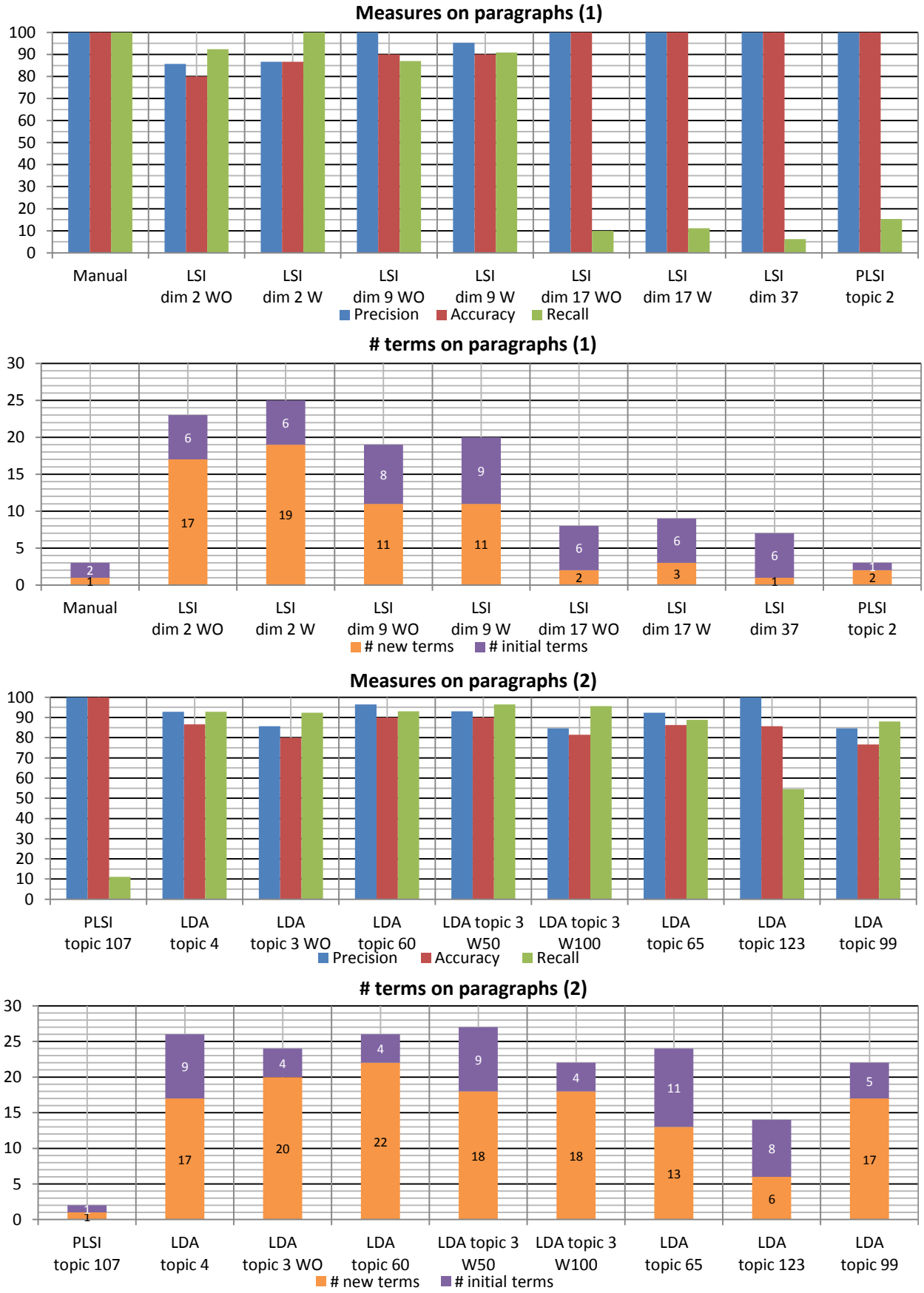
their best performances do not differ too much. *Topic 144* of LDA is the best performance, but with only a single new term lower, PLSI on *topic 135* is almost the best. This is the best rank of a topic of PLSI performances compared with the rest of IR methods. It is found in *topic 159* of LDA how it reaches a higher total value of number of terms, but many of it are initial terms, which shows that it defines very well the *organizer* product feature, but it does not achieve good results in the extraction process. Otherwise, the rest of topics reach all of the highest values of precision and accuracy, but they do not achieve a significant recall, which leads a poor number of new semantically related terms extracted regarding the *organizer* initial terms on documents.

Looking at performances on sections scenario it is found some of the best performance of the whole research, where almost all the top 30 retrieved terms by methods on some topics represent precisely the *organizer* product feature. In **Graphic 6-5** it is shown how LSI on *dimension 4* with (better) and without *extra reviews* gets the maximum recall remaining precision and accuracy over the 95%, and extracts up to 21 new terms, which in addition with the initial terms only cause 1 and 2 errors, with and without, respectively. *Dimension 8* without *extra reviews* performance gets up to 16 new terms, whose recall reaches more than 95% too, but with less initial terms and less new terms extracted, it remains as the second option of the LSI performances. Surprisingly, PLSI does not achieve good results, even though they are quite poor in comparision with the LSI ones that it was supposed to be worse in this kind of performance. Finally, LDA shows very regular performances extracting from 10 to 19 new terms, but any of those topics overcomes the LSI ones. Adding the initial terms, they remain in a high position with 27 terms of the top 30 semantically related, which is not a low percentage. *Topic 43*, *topic 8* and *topic 33* respresent best LDA performances. Their precision, accuracy and recall are as well as the LSI ones, without reaching the 100%, but still high. Although LDA is the most regular, LSI results give it the best rank of being the best IR method which better describe the *organizer* product feature by its initial terms.

Finally, in **Graphic 6-6** it is summarize graphically performances relatives to the paragraphs scenario, where as in sections scenario, all IR methods can be compared due to they get relevant results on them. Also in the same line like in sections, PLSI performs worse that it was thought, because once again LSI and LDA lead the best results. *Dimension 2* in performances with (better too) and without *extra reviews* are the best LSI performances with 19 and 17 new terms extracted respectively, which in addition with the initial terms get 25 and 23, respectively, from the top 30 terms. Moreover, in performance with it achieves the highest recall and more than 90% without them, keeping precision and recall high too. Although these values are considerably high, LDA outperforms them getting *topic 3* without *extra reviews* and *topic 60* 20 and 22 new terms extracted. Particular cases occur with *topic 4*, *topic 3* with and k=50 and *topic 65*, whose new terms extracted do not reach the top imposed by the first two, but they double the number of initial terms. This fact means that those topics which have more initial terms within its top 30 terms, have fewer probabilities to find new terms, then, it is considered both number of terms, new and initials to define better the *organizer* product feature. As a result, it is concluded that LDA outperforms LSI and PLSI performances on the seeking of new terms related with the *organizer* initial product features, but remaining LSI on the second place.



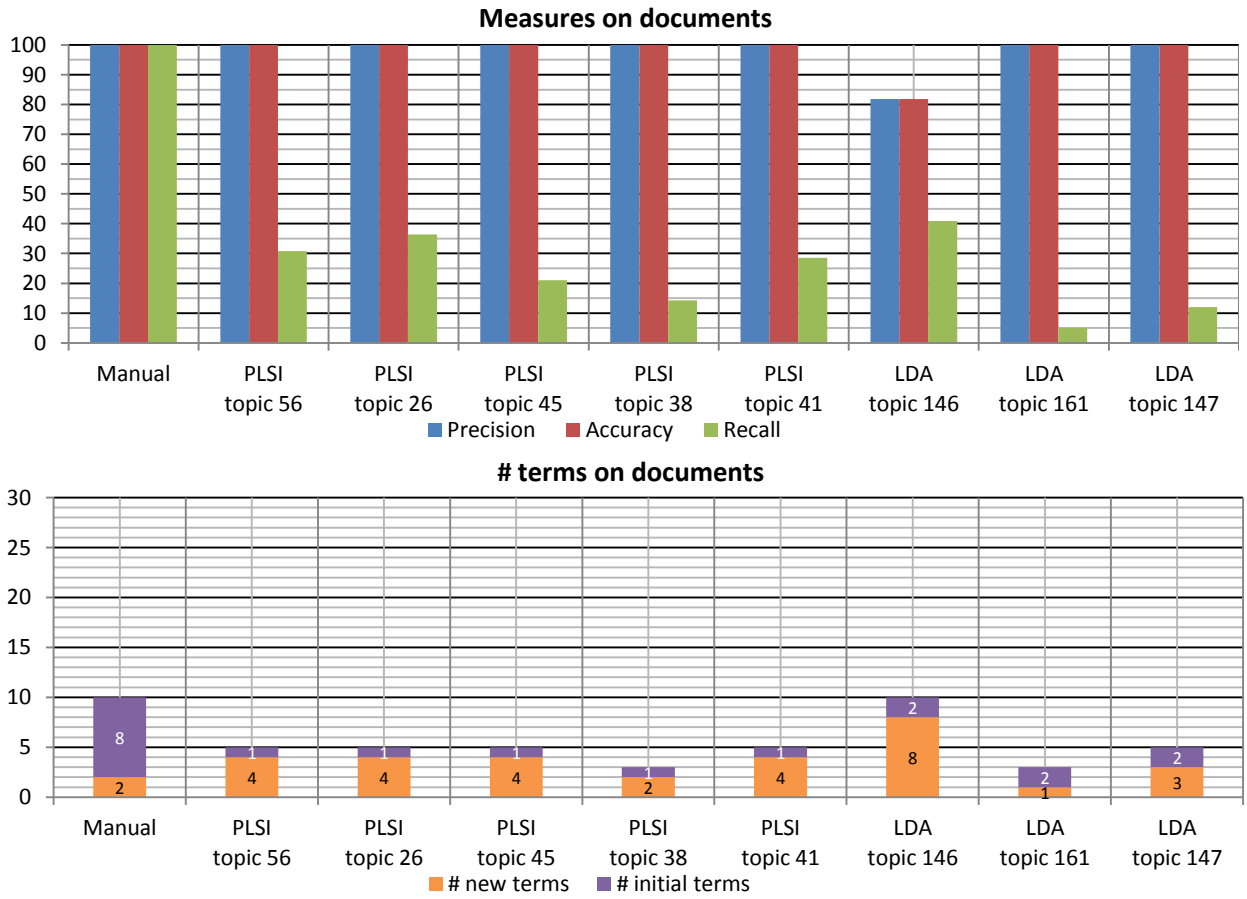
Graphic 6-5. Summarization of measures and extracted terms of *organizer* on sections.



Graphic 6-6. Summarization of measures and extracted terms of *organizer* on paragraphs.

6.3. Summary and performance comparision on *multimedia* product feature

Multimedia results from analysis on each method, manual and IR techniques, are collected in this section to see at one sight how they performs alone and in comparision with the rest of methods. In **Table 6-3** all results retrieved from the different analysis are summarized. Manual analysis is included here to see how different are results got in a tedious and not scalable task from the performances of automatic methods where some parameters are set and the data is prepared and it is done many times in some differents scenarios. It is seen that single results obtained in the manual analysis, where the highest values of precision, accuracy and recall are reached, but with only 2 new terms are extracted. As it happened with the *organizer* product feature, in the *multimedia* analysis there is only one pair method-scenario where there is any result relevant to be shown here. This exception occurs with LSI on documents scenario, which demonstrates that LSI does not run correctly in a wide context scenario, where the whole review is taken as context.



Graphic 6-7. Summarization of measures and extracted terms of *multimedia* on documents.

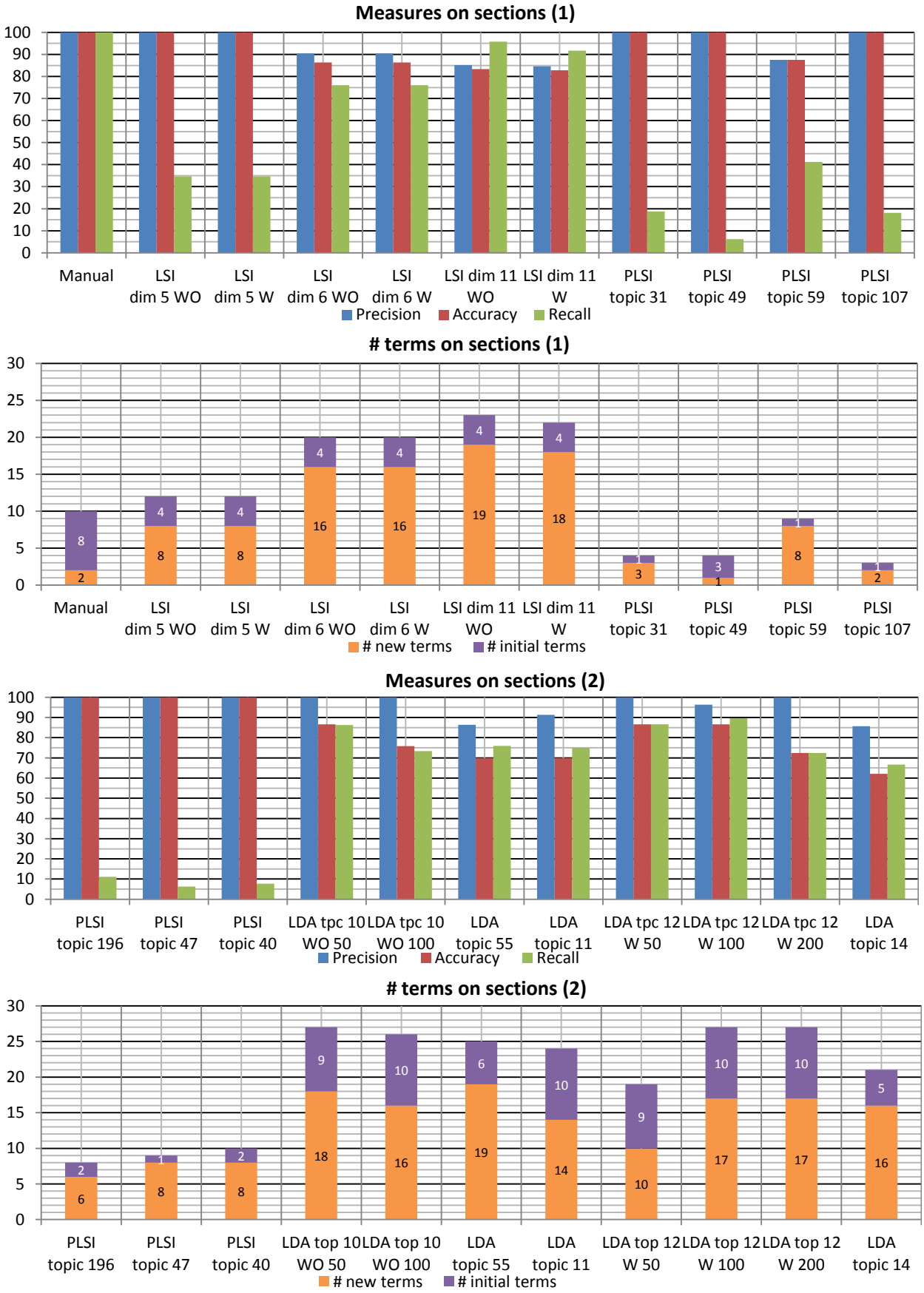
multimedia									
Method	Scenario	Topic	Mode	k parameter	Precision (%)	Accuracy (%)	Recall (%)	# new terms	# initial terms
Manual	context	-	without	-	100	100	100	2	8
LSI	documents	No dimensions	-	-	-	-	-	-	-
	sections	*dimension 5*	without	50,100,200	100	100	34,62	8	4
			with	50,100,150,200	100	100	34,62	8	4
		dimension 6	without	50,150,200	90,48	86,36	76	16	4
			with	50,200	90,48	86,36	76	16	4
		dimension 11	without	50,100,150	85.19	83.33	95.83	19	4
			with	50,150	84.62	82.76	91.67	18	4
	paragraphs	*dimension 3*	without	100	95	89.66	86.36	15	4
			with	50	86.36	76.67	82.61	14	4
		dimension 5	without	50,100,200	100	66.67	65.52	15	4
			with	50,100	100	63.33	62.07	15	3
		dimension 15	without	100,200	100	100	31.25	4	3
			with	50,100,200	100	100	16.67	2	3
		dimension 39	without	100,150	100	100	11.76	1	4
		dimension 13	with	100,150	100	100	15.38	1	3
PLSI	documents	*topic 56*	without	150	100	100	30.77	4	1
		topic 25	without	200	100	100	36.36	4	1
		topic 45	without	200	100	100	21.05	4	1
		topic 38	with	50	100	100	14.29	2	1
		topic 41	with	100	100	100	28.57	4	1
	sections	*topic 31*	without	50	100	100	18.75	3	1
		topic 49	without	100	100	100	6.25	1	3
		topic 59	without	100	87.50	87.50	41.18	8	1
		topic 107	without	200	100	100	18.18	2	1
		topic 196	without	200	100	100	11.11	1	1
		topic 47	with	150	100	100	6.25	1	1
		topic 40	with	200	100	100	7.69	1	1
	paragraphs	*topic 153*	without	200	100	100	12.50	1	1
		topic 21	with	100	100	100	18.18	2	2
		topic 107	with	150	100	100	9.09	1	1
LDA	documents	*topic 146*	without	200	81.82	81.82	40.91	8	2
		topic 161	without	200	100	100	5	1	2
		topic 147	with	150	100	100	12	3	2
	sections	*topic 10*	without	50	100	86.67	86.26	19	6
		topic 10	without	100	100	75.86	73.33	16	6
		topic 55	without	150	86.36	70	76	11	7
		topic 11	without	200	91.30	70	75	12	8
		topic 12	with	50	100	86.67	86.67	20	6
		topic 12	with	100	96.30	86.67	89.66	15	11
		topic 12	with	200	100	72.41	72.41	13	8
		topic 14	with	200	85.71	62.07	66.67	9	9
	paragraphs	*topic 5*	without	50	96	83.33	85.71	19	5
		topic 40	without	100	95.24	84	74.07	13	7
		topic 56	without	150	90.91	66.67	71.43	15	5
		topic 46	without	200	96.15	83.33	86.21	16	9
		topic 35	with	50	93.10	90	96.43	21	6
		topic 22	with	100	95.83	82.76	82.14	14	9
		topic 20	with	150	91.67	76.67	81.48	12	10
		topic 58	with	200	100	90	89.29	14	8

Table 6-3. Summary of the obtained results regarding on *multimedia*.

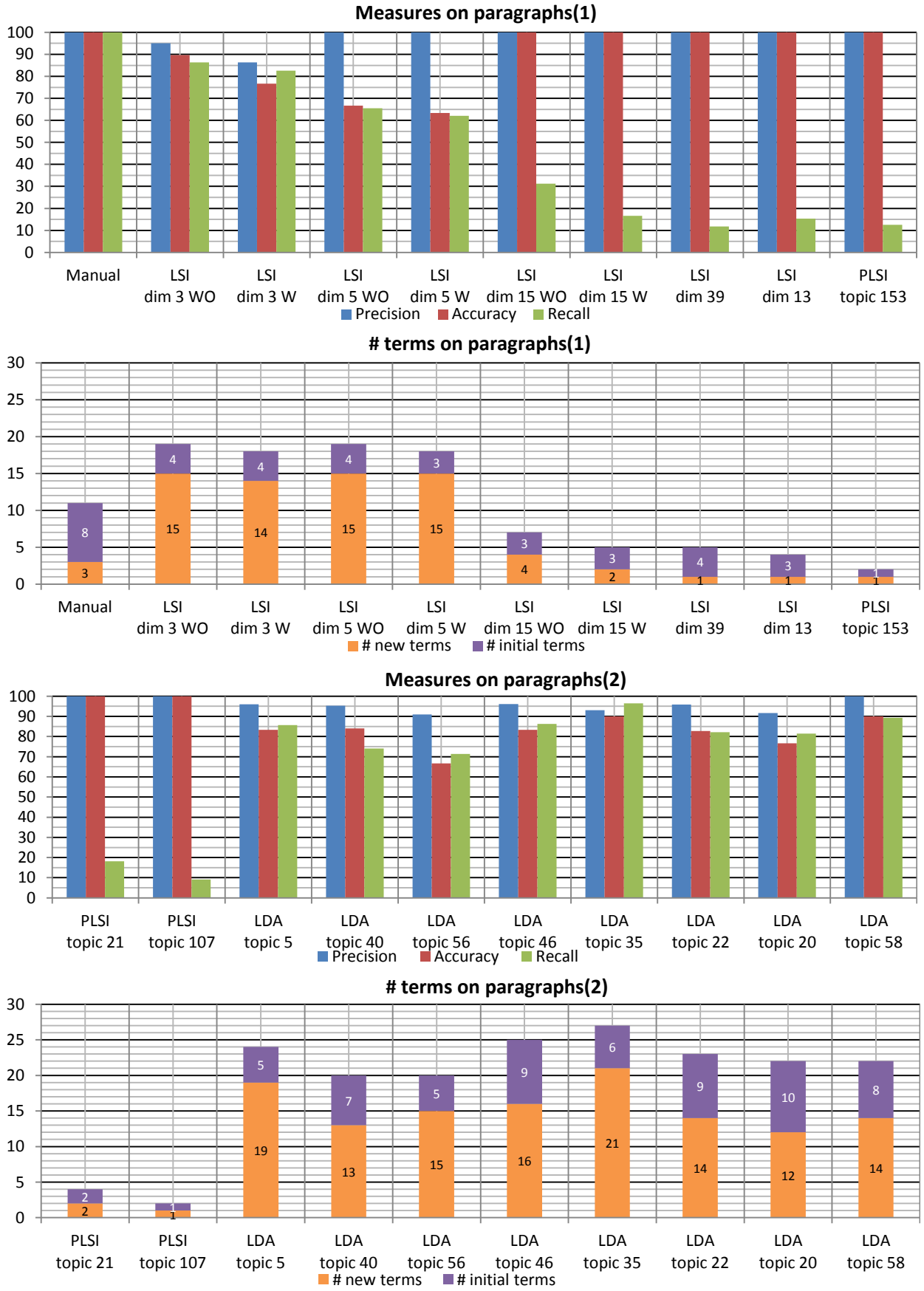
In **Graphic 6-7** it is seen the summarization of performances of PSLI and LDA performances in comparision with the manual analysis on documents scenario. Although they do not achieve as good results as it was supposed, like on sections or paragraphs, it is also noticed that it is obtained the best performance of the entire collection of performances on documents. *Topic 146* on LDA gets the highest number of new terms extracted on this kind of scenario, keeping the rest of measures high, but with a recall still lower than 50%. It extracts up to 8 new semantically related terms, having only 2 initial terms, the opposite of the manual results. PLSI performances are more regular, but they only reach 4 new terms on their extractions, with the same number of initial terms and the highest values of precision and accuracy, but lower recall than in *topic 146* of LDA. Then LDA is the best IR technique to extract new product features semantically related with the *multimedia* intial ones in documents scenario.

Results collected from the analysis on section scenario are showed in **Graphic 6-8**, where up to 22 extractions are compared. It is the scenario where most methods extract relevant results on their performances. It is due to the correct context reduction and also helped by the high number of *multimedia* initial terms, which increase the possibilities of finding some others related. LSI achieves its best results in *dimension 11*, with and without *extra reviews*, where they extract 18 and 19 new terms respectively. *Dimension 6*, also with and without, gets high values of precision, accuracy and recall, but it only extracts 16 new terms in both performances. PLSI remains in regular but low number of terms extracted, whose best performances extract up to 8, less than the half of new terms extracted in best LSI performance. This fact does not reflect what it was supposed about PLSI. However, LDA satisfy considerably the expectations generated around its performances, getting from 16 to 19 new terms on its best performances. The difference with the LSI performances is that in LDA these best topics havefrom 6 to 10 initial terms within their top 30 terms, while LSI dimensions only have 4. This is the main difference between those methods, because their measures are similar, having best recall LSI and an abnormal precision on LDA performances, quite higher than accuracy and recall. It is concluded that in with similar number of new terms extracted, LDA is better, because their performances have more initial terms, which include more sense and coherence to the results obtained.

Finally, in **Graphic 6-9** it is summarized the results obtained on paragraphs scenario, while LSI and PLSI performances reduce their number of new term extracted, LDA performs even better than on sections. Best LSI results are found in *dimension 3*, with and without *extra reviews*, which measures are high, unlike *dimension 5*, with and without too, whose precision is the highest, but its accuracy and recall are atypically much lower. PLSI gets poorer results than on sections, whose best performance obtains up to 4 new terms. Nevertheless, LDA obtains its second best performances with *topic 35* with up to 21 new terms semantically related with the *multimedia* intial one, but only with 6 initial terms, although with accuracy, precision and recall more or equal than 90%. *Topic 5* and *topic 46* also obtain good results with 19 and 16 new term extracted. As a result, LDA remains better than LSI and PLSI, and of course better than manual results, then, LDA is the best IR technique to extract semantically related terms with the *multimedia* initial product features on paragraphs scenario.



Graphic 6-8. Summarization of measures and extracted terms of *multimedia* on sections.



Graphic 6-9. Summarization of measures and extracted terms of *multimedia* on paragraphs.

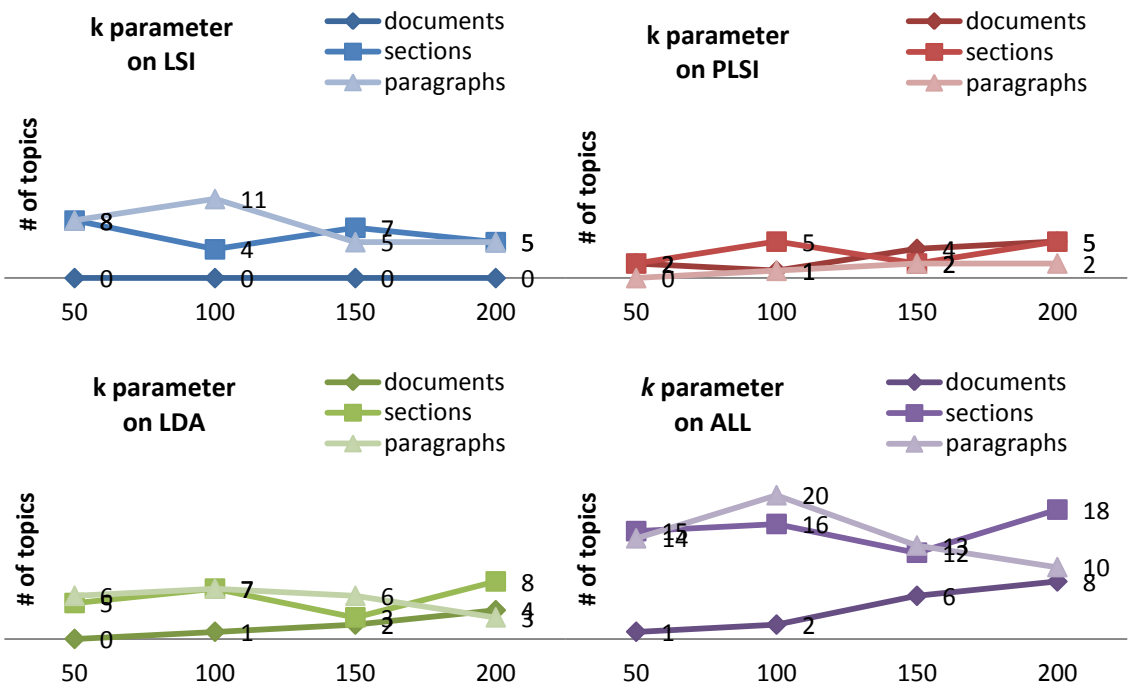
6.4. Regarding the *k* parameter and *extra reviews*

Although it is talked about results obtained in the different performances, it is not yet presented any conclusion about the inference of the *k* parameter on results depending on used values and the changes observed with the inclusion of the *extra reviews*. In **Table 6-4** it is summarized information extracted from the tables presented before. In LSI performances it is counted the times that a dimension has obtained relevant results it all the performances analyzed. It means that if a dimension, although it has the same values and performance in more than one performance with different values of the *k* parameter, it is counted each single performance to extract the concluisions from its performance with the different values.

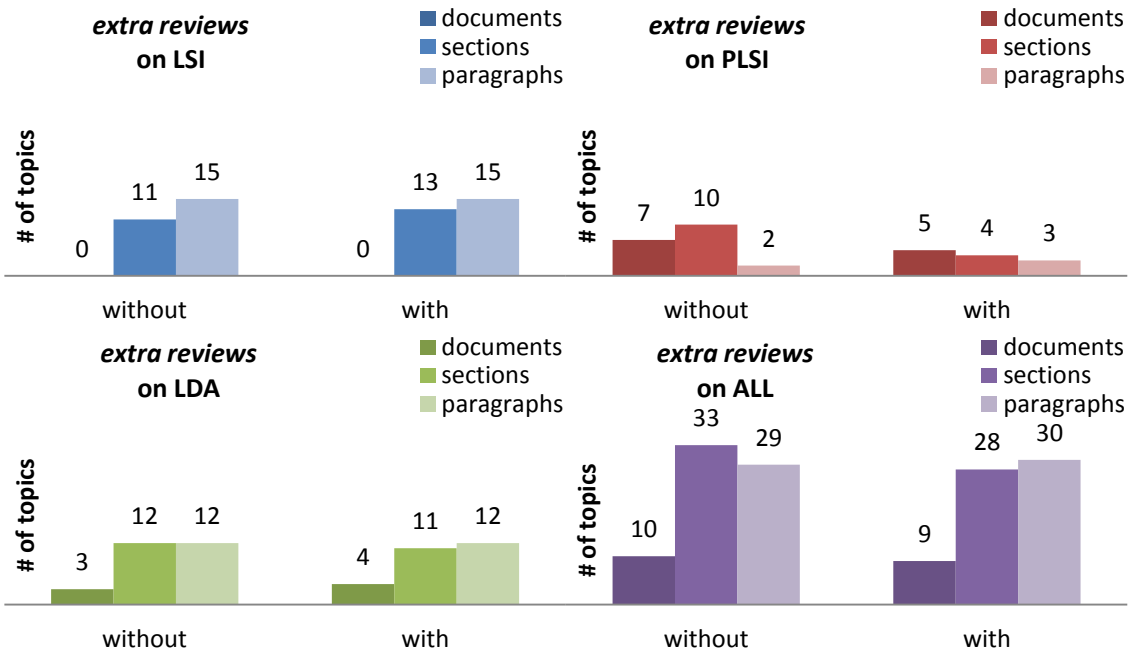
Product feature	Method	Scenario	50	100	150	200	without	with
battery	LSI	documents	-	-	-	-	-	-
		sections	-	-	-	-	-	-
		paragraphs	-	1	-	-	1	-
	PLSI	documents	-	-	-	1	1	-
		sections	-	-	-	-	-	-
		paragraphs	-	-	-	-	-	-
	LDA	documents	-	1	-	-	-	1
		sections	1	2	2	2	3	4
		paragraphs	2	2	2	-	5	3
organizer	LSI	documents	-	-	-	-	-	-
		sections	2	1	3	1	2	5
		paragraphs	4	4	3	2	6	7
	PLSI	documents	1	-	3	2	3	3
		sections	1	3	1	2	5	2
		paragraphs	-	1	1	-	1	1
	LDA	documents	-	-	1	2	1	2
		sections	2	3	-	3	5	3
		paragraphs	2	3	2	1	3	5
multimedia	LSI	documents	-	-	-	-	-	-
		sections	6	3	4	4	9	8
		paragraphs	4	7	2	3	8	8
	PLSI	documents	1	1	1	2	3	2
		sections	1	2	1	3	5	2
		paragraphs	-	-	1	2	1	2
	LDA	documents	-	-	1	2	2	1
		sections	2	2	1	3	4	4
		paragraphs	2	2	2	2	4	4

Table 6-4. Summary of the *k* parameter and *extra reviews*.

In **Graphic 6-10** it is shown graphically the tendency of the most relevant dimensions and topics analysed before separated by method and also grouped all together. On documents, although LSI does not has any relevant dimension, it is seen that PLSI and LDA tend to get more relevant topics as the *k* parameter is increased, where most of of topics are located around k=200. On sections LSI performs as the opposite of the others regarding the *k* parameter, getting more dimensions on k=50 and k=150, but PLSI and LDa in k=100 and k=200, which is shown on the general graphic as a quasi-regular line. Finally, on paragraphs it is found a regular behaviour in PLSI, but LSI and LDA get their most of topics on k=100. Global values correspond with the majority of those mentioned before.



Graphic 6-10. Overview of the *k* parameter values.



Graphic 6-11. Overview of the *extra reviews* inclusion.

Finally, **Graphic 6-11** shows graphically how the most relevant dimensions and topics are affected by the addition of the *extra reviews*. However, with the exception of the PLSI performances on documents and sections scenarios, which decrease its number of relevant analysed topics. LSI and LDA show a similar number of relevant dimension and topics respectively, where they are equal or increased a little, on each scenario analysed, as PLSI does on paragraphs too.

7. Conclusions and future work

In this research, it is covered the goal of discovering new semantic relationships between terms which appear in reviews and technical specifications, all together, of smartphones, through three classical information retrieval techniques, such as Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (PLSI) and Latent Dirichlet Allocation (LDA). New relationships are extracted following an automatic process where terms, nouns and compound nouns, have been treated from their natural language form using Natural Language Processing (NLP) techniques to the specific format required by IR techniques. In order to check that they satisfy the objective ennounced, an analysis has been carried out.

LSI, PLSI and LDA, in the named order, represent the evolution of the information retrieval methods, covering from the use of Term Frequency-Inverse Document Frequency weight (TF-IDF), to the topic modelling algorithms, passing through probabilistic approaches which exploit co-occurence. Having this progression line as a background, it has been analysed in determinated and coherent contexts, where it was supposed that the known evolution will be reflected on results, being LSI which poorest results obtains and going up with the rest. This analysis of the results shows the statements which support these conclusions.

In general, LDA has satisfied overall the assumptions of being the best method, obtaining the best results in most of all performances comparations. It only has been overcome by LSI in the extraction of semantically related terms with the *organizer* product feature on sections scenario. This fact also demonstrates the poor performance of PLSI, which has not satisfied the expectations of being better than LSI. Otherwise, it is known that, finding a PLSI implementation that fits to run on the entire collection of scenarios in the same conditions like the other two has been a hard task. Apart from this negative point, the rest of the analysis shows how efficiently IR techniques can be applied in the extraction of semantic relations between product features.

As it was expected, *battery* product featured accompanied by its initial product features has achieved the poorest results due to its small initial semantic group. Unlike it occurs in *organizer* and *multimedia* initial semantic groups, whose higher number of terms has brought a higher number of new semantically related terms.

Regarding the scenario environment and the classical assumption of getting reviews as a whole, it has produced the poorest results, where only LDA has obtained significant results. Sections and paragraphs has shown their little differences getting similar results from them, although sections seem to be more coherent and considering results, even PLSI has obtained its best outcomes in this context. The successfully results obtained on sections and paragraphs justify and satisfy the decision of dividing reviews on bigger and smaller parts. The closest the

context is, the better performances of IR techniques are, except for PLSI, which does not satisfy this thesis, but it does not satisfy most of the assumptions figured out around it.

Referring to the values used in the *k* parameter, which has been treated in the same way in all the methods, because of it determines similar concepts in all of them. It is noticed that k=100 is the value that more dimensions and topics has grouped, being the middle point between k=50, where it may exists a need of higher subdivision, and k=200, where many topics do not improve the obtained refinement. The point is that there is a huge number of *k* possible values that can be analysed like 50, 100, 150 and 200 have been analysed here, but the decision of starting in k=50 and taking step by step of 50 has shown how methods perform in a wide range of values. Anyway, better and worse performances can be found using similar values, but it is a representation of one possibility.

Moreover, it is proved that the altered context with the addition of *extra reviews*, which contain those product features extracted from the technical specifications and grouped by each main product feature, has shown more than punctual improvements or deficiencies in comparision with the natural contexts. Then, efforts which have consisted in double the entire analysis considering those additional ideal reviews, present any significant improvement.

Apart from the analysis conclusions, it has been attempted to develop some own methods, at least, some new concepts about the same objective, but the technologies and implementation used seem not to fit the proposals. It is thought that if it is developed a parallel implementation of the co-occurence and context similarity method and it is run on a group of machines which can be dedicated to the proposed goal in this research, an actual performance may be carried out and could be tested properly. Moreover, in order to run successfully the proposed Cascade LDA method (CLDA), which conceptual objective consists on applying LDA twice, once normally and once again on results, a dedicated implementation for this prupose could be developed, and then, results may be tested properly too.

Finally, it is noted that most of work developed, used, tested and compared in this research depend on the trainng data, whose environment can be modified just a little, and a new analysis will show surely big differences with results obtained here. However, it has been tested the capacity of classical IR techniques, which has satisfied most of the thesis proposed here. Moreover, it is mentioned once again that processes of extraction of sematically related terms like the described in this document, could be included in semantic and classical search engines, query expansion processes, machine learning tasks and opinion mining. The YAGO-NAGA project, which leads one to think that the arrival of the semantic search is closer day by day, deserves a particular mention. Efforts like this research can help or be helped by the ontology concepts to reach bigger and wider objectives. In the end, it is known that research is alive, which means that people like us around the world are working in researches similar to this one, when sometimes obtained results become ephermeral, but sometimes not.

8. zBibliography

8.1. References

1. Ohler, J., *The semantic web in education.* EDUCAUSE Quarterly, 2008. **31**(4): p. 7-9.

2. Hu, M. and B. Liu, *Mining and summarizing customer reviews*, in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, ACM: Seattle, WA, USA. p. 168-177.

3. Hu, M. and B. Liu, *Mining opinion features in customer reviews*, in *Proceedings of the 19th national conference on Artifical intelligence*. 2004, AAAI Press: San Jose, California. p. 755-760.

4. Liu, B., M. Hu, and J. Cheng, *Opinion observer: analyzing and comparing opinions on the Web*, in *Proceedings of the 14th international conference on World Wide Web*. 2005, ACM: Chiba, Japan. p. 342-351.

5. Popescu, A.-M. and O. Etzioni, *Extracting product features and opinions from reviews*, in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 2005, Association for Computational Linguistics: Vancouver, British Columbia, Canada. p. 339-346.

6. Scaffidi, C., *Application of a probability-based algorithm to extraction of product features from online reviews*. 2006.

7. Scaffidi, C., et al., *Red Opal: product-feature scoring from reviews*, in *Proceedings of the 8th ACM conference on Electronic commerce*. 2007, ACM: San Diego, California, USA. p. 182-191.

8. Miao, Q., Q. Li, and R. Dai, *An integration strategy for mining product features and opinions*, in *Proceeding of the 17th ACM conference on Information and knowledge management*. 2008, ACM: Napa Valley, California, USA. p. 1369-1370.

9. Somprasertsri, G. and P. Lalitrojwong, *Automatic product feature extraction from online product reviews using maximum entropy with lexical and syntactic features*, in *Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference*. 2008. p. 250 -255.

10. Guo, H., et al., *Product feature categorization with multilevel latent semantic association*, in *Proceeding of the 18th ACM conference on Information and knowledge management*. 2009, ACM: Hong Kong, China. p. 1087-1096.

11. Kim, W.Y., et al., *A method for opinion mining of product reviews using association rules*, in *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*. 2009, ACM: Seoul, Korea. p. 270-274.

12. Lopez-Fernandez, A., T. Veale, and P. Majumder, *Feature Extraction from Product Reviews using Feature Similarity and Polarity*. 2009.

13. Khoo, C.S.G. and J.-C. Na, *Semantic relations in information science.* Annual Review of Information Science and Technology, 2006. **40**(1): p. 157-228.

14. Saussure, F.d., *Course in general linguistics*. 1959, New York: Philosophical Library.

15. Asher, R.E. and J.M.Y. Simpson, *The Encyclopedia of language and linguistics*. 1994: Pergamon Press.

16. Calzolari, N., *The dictionary and the thesaurus can be combined*, in *Relational models of the lexicon*. 1988, Cambridge University Press. p. 75-96.

17. Cruse, D.A., *Hyponymy and its varieties.* In R. Green, C. Bean, & S. H. Myaeng, The Semantics of Relationships: An interdisplinary perspective, 2002: p. 3-22.

18. Tversky, B., *Where partonomies and taxonomies meet.* In S. L. Tsohatzidis (Ed.), Meanings and prototypes: Studies in linguistic categorization, 1990: p. 334-344.

19. Cruse, D.A., *Lexical semantics*. 1986, Cambridge: Cambridge University Press.

20. Lyons, J., *Linguistic semantics: An introduction*. 1995, Cambridge: Cambridge University Press.

21. Jones, S., *Antonymy: A corpus-based perspective*. 2002, London: Routledge.

22. Justeson, J.S. and S.M. Katz, *Co-occurrence o antonymous adjectives and their contexts.* Computational Linguistics, 1991. **17**: p. 1-19.

23. Devi, S.L. and M. S., *Semantic Representation of Causality.* In Kommaluri Vijayanand, and L. Ramamoorthy, (ed), Language in India, 2011. **11**.

24. Turney, P.D., *Measuring semantic similarity by latent relational analysis*, in *Proceedings of the 19th international joint conference on Artificial intelligence*. 2005, Morgan Kaufmann Publishers Inc.: Edinburgh, Scotland. p. 1136-1141.

25. Hofmann, T., *Probabilistic latent semantic indexing*, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999, ACM: Berkeley, California, United States. p. 50-57.

26. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation.* J. Mach. Learn. Res., 2003. **3**: p. 993-1022.

27. Titov, I. and R. McDonald, *Modeling online reviews with multi-grain topic models*, in *Proceeding of the 17th international conference on World Wide Web*. 2008, ACM: Beijing, China. p. 111-120.

28. Zhai, Z., et al., *Grouping product features using semi-supervised learning with soft-constraints*, in *Proceedings of the 23rd International Conference on Computational Linguistics*. 2010, Association for Computational Linguistics: Beijing, China. p. 1272-1280.

29. Zhai, Z., et al., *Clustering product features for opinion mining*, in *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011, ACM: Hong Kong, China. p. 347-354.

30. Pangos, A., et al., *Combining Statistical Similarity Measures for Automatic Induction of Semantic Classes*. 2005.

31. Sahami, M. and T.D. Heilman, *A web-based kernel function for measuring the similarity of short text snippets*, in *Proceedings of the 15th international conference on World Wide Web*. 2006, ACM: Edinburgh, Scotland. p. 377-386.

32. Bollegala, D., Y. Matsuo, and M. Ishizuka, *Measuring semantic similarity between words using web search engines*, in *Proceedings of the 16th international conference on World Wide Web*. 2007, ACM: Banff, Alberta, Canada. p. 757-766.

33. Takale, S.A. and S.S. Nandgaonkar, *Measuring semantic similarity between words using web search engines*, in *Proceedings of the 16th international conference on World Wide Web*. 2007, ACM: Banff, Alberta, Canada. p. 757-766.

34. Iosif, E. and A. Potamianos, *Unsupervised Semantic Similarity Computation using Web Search Engines*, in *Web Intelligence, IEEE/WIC/ACM International Conference*. 2007. p. 381 -387.

35. Church, K.W. and P. Hanks, *Word association norms, mutual information, and lexicography.* Comput. Linguist., 1990. **16**(1): p. 22-29.

36. Cilibrasi, R.L. and P.M.B. Vitanyi, *The Google Similarity Distance.* IEEE Trans. on Knowl. and Data Eng., 2007. **19**(3): p. 370-383.

37. Lin, D., *An Information-Theoretic Definition of Similarity*, in *Proceedings of the Fifteenth International Conference on Machine Learning*. 1998, Morgan Kaufmann Publishers Inc. p. 296-304.

38. Hindle, D., *Noun classification from predicate-argument structures*, in *Proceedings of the 28th annual meeting on Association for Computational Linguistics*. 1990, Association for Computational Linguistics: Pittsburgh, Pennsylvania. p. 268-275.

39. Han, L., et al., *ADSS: an approach to determining semantic similarity.* Adv. Eng. Softw., 2006. **37**(2): p. 129-132.

40. Budanitsky, A. and G. Hirst, *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*, in *IN WORKSHOP ON WORDNET AND OTHER LEXICAL RESOURCES, SECOND MEETING OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*. 2001.

41. Li, Y., Z.A. Bandar, and D. McLean, *An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources.* IEEE Trans. on Knowl. and Data Eng., 2003. **15**(4): p. 871-882.

42. King, J. and V. Satuluri, *Extracting Semantic Relations Using Dependency Paths.* Unpublished.

43. Stoutenburg, S., J. Kalita, and S. Hawthorne, *Extracting semantic relationships between Wikipedia articles*, in *In Proc. 35th International Conference on Current Trends in Theory and Practice of Computer Science*. 2009: Spindelruv Mlyn, Czech Republic.

44. Spagnola, S. and C. Lagoze, *Edge dependent pathway scoring for calculating semantic similarity in ConceptNet*, in *Proceedings of the Ninth International Conference on Computational Semantics*. 2011, Association for Computational Linguistics: Oxford, United Kingdom. p. 385-389.

45. Tous, R. and J. Delgado, *A vector space model for semantic similarity calculation and OWL ontology alignment*, in *IN DEXA 2006*. 2006.

46. Akbik, A. and J. Broß, *Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns*, in *WWW Workshop*. 2009.

47. Nagano, S., M. Inaba, and T. Kawamura, *Extracting Semantic Relations for Mining of Social Data*, in *SDoW2010 Workshop at the 9th International Semantic Web Conference (ISWC2010)*. 2010: Shanghai (China).

48. Lin, D., *Automatic retrieval and clustering of similar words*, in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*. 1998, Association for Computational Linguistics: Montreal, Quebec, Canada. p. 768-774.

49. Kozima, H. and T. Furugori, *Similarity between words computed by spreading activation on an English dictionary*, in *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*. 1993, Association for Computational Linguistics: Utrecht, The Netherlands. p. 232-239.

50. Kasneci, G., et al., *The YAGO-NAGA approach to knowledge discovery.* SIGMOD Rec., 2009. **37**(4): p. 41-47.

51. Suchanek, F.M., G. Kasneci, and G. Weikum, *YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia*, in *Proceedings of the International World Wide Web Conference*. 2007.

52. Kasneci, G., et al., *NAGA: Searching and Ranking Knowledge*, in *International Conference on Data Engineering*. 2008. p. 953-962.

53. Hoffart, J., et al., *YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages*, in *Proceedings of the 20th international conference companion on World Wide Web (WWW 2011)*. 2011, ACM: Hyderabad, India. p. 229-232.

54. Mahinovs, A. and A. Tiwari, *Text classification method review*, R. Roy and D. Baxter, Editors. 2007, Cranfield University.

55. Hull, D.A., *Stemming Algorithms - A Case Study for Detailed Evaluation.* Journal of the American Society for Information Science, 1996. **47**: p. 70-84.

56. Lovins, J.B., *Development of a Stemming Algorithm.* Mechanical Translation and Computational Linguistics, 1968. **11**: p. 22-31.

57. Porter, M.F., *An algorithm for suffix stripping*, in *Readings in information retrieval*, J. Karen Sparck and W. Peter, Editors. 1997, Morgan Kaufmann Publishers Inc. p. 313-316.

58. Deerwester, S.C., et al., *Indexing by Latent Semantic Analysis.* Journal of the American Society of Information Science, 1990. **41**(6): p. 391-407.

59. Hasan, M.M. and Y. Matsumoto, *Document Clustering: Before and After the Singular Value Decomposition.* Joho Shori Gakkai Kenkyu Hokoku, 1999. **99**(95(NL-134)): p. 47-54.

60. Berry, M.W., S.T. Dumais, and G.W. O'Brien, *Using Linear Algebra for Intelligent Information Retrieval.* SIAM Review, 1995. **37**: p. 573-595.

61. Dumais, S.T., *Improving the retrieval of information from external sources.* Behavior Research Methods, Instruments and Computers, 1991. **23**(2): p. 229-236.

62. Tu, N.C., *Hidden Toopic Discovery Toward Classification and Clustering in Vietnamese Web Documents*. 2008, Vietnam National University: Hanoi.

63. Kakkonen, T., et al. *Comparison of Dimension Reduction Methods for Automated Essay Grading*. 2008.

64. Hofmann, T., *Probabilistic Latent Semantic Analysis*, in *Proceedings of Uncertainty in Artificial Intelligence, UAI'99*. 1999: Stockholm.

65. Kontostathis, A. and W. Pottenger, *Detecting patterns in the LSI term-term matrix*. 2002.

9. Annex A: LSI's running tables

9.1. *Battery*

9.1.1. Documents

	Without	With
K=50	× **Dim 1** · ____ 0.6014 – top dimension · 191st with 0.2470 – battery · 379th with 0.1855 – capacity · 444th with 0.1738 – video playback · 600th with 0.1439 – talk time · 1220th ... → No negative values × **Dim 5** · ____ 0.7153 – top dimension · 142nd 0.1043 – battery · 205th ... → 3 negative values	× **Dim 1** · ____ 0.6026 – top dimension · 191st with 0.2477 – battery · 385th with 0.1874 – capacity · 455th with 0.1723 – video playback · 597th with 0.1447 – talk time · 1225th ... → No negative values × **Dim 5** · ____ 0.7150 – top dimension · 142nd 0.1049 – battery · 203rd ... → 3 negative values × **Dim 10** · ____ 0.3931 – top dimension · 71th with 0.1115 – capacity · 189th ... → 3 negative values
K=100	× **Dim 1** (same as k=50) × **Dim 10** · ____ 0.3933 – top dimension · 68th with 0.1120 – capacity · 175th ... → 3 negative values × **Dim 5** (negative values)	× **Dim 1** (same as k=50) × **Dim 10** (same as k=50) × **Dim 46** · ____ 0.9598 – top dimension · 22nd with 0.1724 – li-ion · 221st ... → 3 negative values × **Dim 5** (negative values)
K=150	× **Dim 1** (same as k=50, k=100) × **Dim 5** (same as k=50) × **Dim 10** (same as k=100)	× **Dim 1** (same as k=50, k=100) × **Dim 5** (same as k=50) × **Dim 10** (same as k=100) × **Dim 46** (negative values)
K=200	× **Dim 1** (same as k=50, k=100, k=150) × **Dim 10** (same as k=100, k=150) × **Dim 5** (negative values)	× **Dim 1** (same as k=50, k=100, k=150) × **Dim 5** (same as k=50, k=100) × **Dim 10** (same as k=100) × **Dim 46** (same as k=100)

Table 9-1. Best scored dimensions of LSI considering the *battery* product feature on documents.

9.1.2. Sections

	Without	With
K=50	× **Dim 1** · ____ 0.9478 – top dimension · 23rd with 0.5445 – battery · 347th with 0.1922 – talk time · 371st ... → No negative values × **Dim 12** · ____ 0.6280 – top dimension · 13th with 0.2916 – battery · 26th with 0.2091 – talk time · 548th ... → 3 negative values × **Dim 25** · ____ 0.3671 – top dimension · 17th with 0.2056 – battery · 89th with 0.2091 – talk time · 173rd ... → 2 negative values	× **Dim 1** · ____ 0.9482 – top dimension · 23rd with 0.5442 – battery · 345th with 0.1926 – talk time · 371st ... → No negative values × **Dim 12** · ____ 0.6281 – top dimension · 13th with 0.2928 – battery · 25th with 0.2112 – talk time · 517th ... → 3 negative values × **Dim 24** · ____ 0.4718 – top dimension · 3rd with 0.3622 – battery · 43th with 0.1649 – talk time · 348th ... → No negative values ✓ **Dim 26** · ____ 0.48022 – top dimension · 6th with 0.2895 – battery · 16th with 0.2140 – talk time · 169th ... → 1 negative value .
K=100	× **Dim 1** (same as k=50) × **Dim 12** (same as k=50) × **Dim 24** · ____ 0.4710 – top dimension · 3rd with 0.3615 – battery · 44th with 0.1622 – talk time · 369th ... → No negative values ✓ **Dim 56** · ____ 0.3375 – top dimension · 2nd with 0.2914 – battery · 23rd with 0.1549 – talk time · 247th ... → No negative values × **Dim 25** (negative values)	× **Dim 1** (same as k=50) × **Dim 12** (same as k=50) × **Dim 24** (same as k=50) ✓ **Dim 56** · ____ 0.3255 – top dimension · 2nd with 0.2636 – battery · 22nd with 0.1608 – talk time · 207th ... → No negative value × **Dim 26** (negative values)
K=150	× **Dim 1** (same as k=50, k=100) × **Dim 12** (same as k=50) × **Dim 24** (same as k=100) × **Dim 25** (same as k=50) × **Dim 56** (negative values)	× **Dim 1** (same as k=50, k=100) ✓ **Dim 56** (same as k=100) × **Dim 12, 24, 26** (negative values)
K=200	× **Dim 1** (same as k=50, k=100, k=150) × **Dim 24** (same as k100, k=150) × **Dim 175** · ____ 0.2548 – top dimension · 3rd with 0.2423 – battery · 358th ... → 1 negative value × **Dim 12, 25, 26, 56** (negative values)	× **Dim 1** (same as k=50, k=100, k=150) × **Dim 24** (same as k=50, k=100) ✓ **Dim 26** (same as k=50) ✓ **Dim 56** (same as k=100, k=150) × **Dim 12** (negative values)

Table 9-2. Best scored dimensions of LSI considering the *battery* product feature on sections.

9.1.3. Paragraphs

	Without	With
K=50	× **Dim 3** · ____ 0.9263 – top dimension · 6th with 0.6237 – battery · 33rd ... → 2 negative values × **Dim 7** · ____ 0.5030 – top dimension · 8th with 0.3286 – battery · 18th with 0.2597 – talk time · 462nd ... → 2negative values × **Dim 16** · ____ 0.5922 – top dimension · 3rd with 0.4499 – battery · 34th ... →1 negative value	× **Dim 14** · ____ 0.6435 – top dimension · 4th with 0.4770 – battery · 36th ... → 1 negative value
K=100	× **Dim 16** (same as k=50) ✓ **Dim 18** · 1st with 0.6716 – battery · 12th with 0.3148 – talk time · 95th ... → 1 negative value · × **Dim 58** · ____ 0.7981 – top dimension · 4th with 0.3457 – battery · 42nd with 0.1493 – talk time · 504th ... → 1 negative value × **Dim 3, 7, 14** (negative values)	× **Dim 3** · ____ 0.9319 – top dimension · 6th with 0.6206 – battery · 34th ... → 2 negative values × **Dim 16** · ____ 0.5950 – top dimension · 3rd with 0.4652 – battery · 29th with 0.2018 – talk time · 39th ... →1 negative value ✓ **Dim 18** · 1st with 0.6550 – battery · 12th with 0.3143 – talk time · 93th ... → 1 negative value × **Dim 14** (negative values)
K=150	× **Dim 3** (same as k=50) × **Dim 7** (same as k=50) × **Dim 14** (same as k=50) ✓ **Dim 17** · ____ 0.9383 – top dimension · 2nd with 0.7465 – battery · 4th with 0.4850 – talk time · 141st ... → 1 negative value ✓ **Dim 18** (same as k=100) × **Dim 136** · ____ 0.3775 – top dimension · 3rd with 0.2686 – video playback · 53rd with 0.1165 – talk time · 92nd ... → 2 negative values × **Dim 16, 58** (negative values)	× **Dim 3** (same as k=100) × **Dim 4** · ____ 0.8234 – top dimension · 7th with 0.4249 – battery · 22nd with 0.2835 – talk time · 71st ... → No negative values × **Dim 14** (same as k=50) ✓ **Dim 18** (same as k=100) × **Dim 137** · ____ 0.3325 – top dimension · 8th with 0.2401 – video playback · 18th with 0.1695 – talk time · 223rd ... → 1 negative value × **Dim 16** (negative values)
K=200	× **Dim 3** (same as k=50, k=150) × **Dim 4** · ____ 0.7613 – top dimension · 8th with 0.4149 – battery · 22nd with 0.2742 – talk time · 74th ... → 2 negative values × **Dim 7** (same as k=50, k=150) × **Dim 16** (same as k=50, k=100) ✓ **Dim 18** (same as k=100, k=150) × **Dim 58** (same as k=100) × **Dim 136** (same as k=150) × **Dim 14, 17**(negative values)	× **Dim 3** (same as k=100, k=150) × **Dim 7** · ____ 0.4989 – top dimension · 8th with 0.3264 – battery · 18th with 0.2605 – talk time · 427th ... → 2 negative values × **Dim 16** (same as k=100) ✓ **Dim 18** (same as k=100, k=150) × **Dim 137** (same as k=150) × **Dim 4, 14** (negative values)

Table 9-3. Best scored dimensions of LSI considering the *battery* product feature on paragraphs.

9.2. *Organizer*

9.2.1. Documents

	Without	With
K=50	× **Dim 1** · ____ 0.6014 – top dimension · 40th with 0.3688 – alarm · 42nd with 0.3605 – note · 58th with 0.3363 – task · 69th with 0.3246 – calendar · 129th ... → No negative values × **Dim 3** · ____ 0.5057 – top dimension · 36th with 0.1756 – memo · 61st with 0.1460 – alarm · 65th with 0.1445 – to-do · 67th with 0.1440 – stopwatch · 145th ... → 4 negative values × **Dim 6** · ____ 0.5290 – top dimension · 25th with 0.2458 – to-do · 72nd with 0.1746 – office · 106th ... → 10 negative values × **Dim 10** · ____ 0.7543 – top dimension · 35th with 0.1773 – pdf · 68th with 0.1423 – note · 79th with 0.1334 – office · 153rd ... → 3 negative values	× **Dim 1** · ____ 0.6026 – top dimension · 40th with 0.3681 – alarm · 42nd with 0.3599 – note · 61st with 0.3356 – task · 69th with 0.3245 – calendar · 130th ... → No negative values × **Dim 3** · ____ 0.5068 – top dimension · 36th with 0.1749 – memo · 62nd with 0.1458 – alarm · 68th with 0.1438 – stopwatch · 69th with 0.1435 – to-do · 143rd ... → 4 negative values × **Dim 6** · ____ 0.5301 – top dimension · 25th with 0.2437 – to-do · 73rd with 0.1742 – office · 106th ... → 10 negative values
K=100	× **Dim 1** (same as k=50) × **Dim 3** (same as k=50) × **Dim 6** (same as k=50) × **Dim 10** (negative values)	× **Dim 1** (same as k=50) × **Dim 3** (same as k=50) × **Dim 6** (same as k=50) × **Dim 87** · ____ 0.4175 – top dimension · 24th with 0.1273 – task · 39th with 0.1088 – memo · 369th ... →6 negative values
K=150	× **Dim 1** (same as k=50, k=100) × **Dim 6** (same as k=50, k=100) × **Dim 3, 10** (negative values)	× **Dim 1** (same as k=50, k=100) × **Dim 3** (same as k=50, k=100) × **Dim 6** (same as k=50, k=100) × **Dim 10, 87** (negative values)
K=200	× **Dim 1** (same as k=50, k=100, k=150) × **Dim 87** · ____ 0.4170 – top dimension · 24th with 0.1773 – task · 38th with 0.1334 – memo · 369th ... → 4 negative values × **Dim 3, 6, 10** (negative values)	× **Dim 1** (same as k=50, k=100, k=150) × **Dim 6** (same as k=50, k=100, k=150) × **Dim 3, 10, 87** (negative values)

Table 9-4. Best scored dimensions of LSI considering the *organizer* product feature on documents.

9.2.2. Sections

	Without	With
K=50	× **Dim 3** · ____ 0.3583 – top dimension · 9th with 0.2485 – alarm · 21st with 0.2061 – note · 23rd with 0.2034 – organizer · 66th ... → 1 negative value ✓ **Dim 4** · ____ 0.7573 – top dimension · 2nd with 0.6480 – calendar · 3rd with 0.5613 – alarm · 4th with 0.5102 – note · 6th with 0.4634 – task · 19th with 0.3941 – organizer · 20th with 0.3768 – calculator · 29th with 0.2853 – office · 31st ... → No negative values × **Dim 16** · ____ 0.2029 – top dimension · 11th with 0.1364 – note · 16th with 0.1578 – alarm · 21st with 0.1232 – to-do · 31st ... → 2 negative values × **Dim 23** · ____ 0.3748 – top dimension · 4th with 0.3141 – pdf · 7th with 0.3023 – office · 42nd ... → 1 negative value × **Dim 25** · ____ 0.3671 – top dimension · 5th with 0.2492 – office · 8th with 0.2347 – pdf · 50th ... → 5 negative values × **Dim 29** · ____ 0.3583 – top dimension · 9th with 0.2485 – alarm · 21st with 0.2061 – office · 23rd with 0.2034 – pdf · 66th ... → 1 negative value ✓ **Dim 35** · 1st with 0.4625 – alarm · 4th with 0.3482 – calculator · 5th with 0.3456 – organizer · 10th with 0.2573 – memo · 12th with 0.2227 – calendar · 15th with 0.1984 – stopwatch · 18th with 0.1963 – world clock · 19th with 0.1947 – note · 20th with 0.1716 – converter · 28th with 0.1695 – countdown timer · 124th ... → 2 negative values	× **Dim 3** · ____ 0.3583 – top dimension · 9th with 0.2485 – alarm · 21st with 0.2061 – note · 22nd with 0.2034 – organizer · 65th ... → 1 negative value ✓ **Dim 4** · ____ 0.7558 – top dimension · 2nd with 0.6490 – calendar · 3rd with 0.5624 – alarm · 4th with 0.5111 – note · 6th with 0.4649 – task · 17th with 0.3954 – organizer · 20th with 0.3784 – calculator · 29th with 0.2888 – office · 30th with 0.2827 – world clock · 36th with 0.2740 – converter · 77th ... → No negative values · × **Dim 16** · ____ 0.2023 – top dimension · 11th with 0.1357 – note · 13th with 0.1287 – alarm · 21st with 0.1244 – to-do · 30th with 0.1170 – memo · 42nd ... → 2 negative values × **Dim 22** · ____ 0.4216 – top dimension · 6th with 0.3057 – office · 10th with 0.2657 – pdf · 44th ... → 3 negative values × **Dim 25** · ____ 0.3633 – top dimension · 5th with 0.2489 – office · 9th with 0.2334 – pdf · 50th ... → 5 negative values × **Dim 29** · ____ 0.4262 – top dimension · 8th with 0.2223 – alarm · 17th with 0.1997 – office · 25th with 0.1822 – pdf · 89th ... → 1 negative value × **Dim 47** · ____ 0.4190 – top dimension · 3rd with 0.3165 – alarm · 13th with 0.2341 – organizer · 14th with 0.2324 – memo · 41st ... → 5 negative values
K=100	× **Dim 16** (same as k=50) × **Dim 23** (same as k=50)	× **Dim 3** (same as k=50) × **Dim 16** (same as k=50) × **Dim 29** (same as k=50)

	× **Dim 34 (particular case)** · ____ 0.3754 – top dimension · 4th with 0.3234 – alarm · 10th with 0.2577 – calendar · 14th with 0.2444 – organizer · 31st ... → 3 negative values ✓ **Dim 35** (same as k=50, k=100) × **Dim 52** · ____ 0.3041 – top dimension · 3rd with 0.1834 – organizer · 16th with 0.1568 – stopwatch · 20th with 0.1473 – alarm · 29th with 0.1286 – calculator · 53rd ... → 4 negative values × **Dim 3, 4, 25, 29, 47** (negative values)	✓ **Dim 34** · 1st with 0.4511 – alarm · 2nd with 0.3466 – organizer · 3rd with 0.3180 – calendar · 9th with 0.2382 – calculator · 12th with 0.2156 – note · 15th with 0.2078 – stopwatch · 16th with 0.1985 – memo · 19th with 0.1938 – world clock · 25th with 0.1785 – converter · 27th with 0.1755 – countdown timer · 52nd ... → 3 negative values ✓ **Dim 35** · ____ 0.5405 – top dimension · 3rd with 0.3439 – alarm · 6th with 0.2975 – calculator · 8th with 0.2632 – organizer · 12th with 0.2176 – memo · 28th with 0.1616 – converter · 34th ... → 2 negative values × **Dim 47** (same as k=50) × **Dim 4, 22, 25** (negative values)
K=150	× **Dim 3** (same as k=50) × **Dim 16** (same as k=50, k=100) × **Dim 22** · ____ 0.4145 – top dimension · 7th with 0.2932 – office · 12th with 0.2542 – pdf · 45th ... → 3 negative values × **Dim 23** (same as k=50, k=100) × **Dim 25**(same as k=50) × **Dim 29** (same as k=50) ✓ **Dim 35** (same as k=50, k=100) × **Dim 47** (same as k=50) × **Dim 57** · ____ 0.3218 – top dimension · 6th with 0.1977 – alarm · 16th with 0.1544 – calculator · 28th with 0.1374 – organizer · 84th ... → 1 negative value × **Dim 4, 34, 52** (negative values)	× **Dim 3** (same as k=50) ✓ **Dim 4** (same as k=100) ✓ **Dim 8** · 1st with 0.4787 – alarm · 3rd with 0.4325 – note · 10th with 0.3320 – calendar · 20th with 0.2844 – task · 25th with 0.2634 – organizer · 27th with 0.2508 – calculator · 37th ... → 3 negative values × **Dim 22** (same as k=50) × **Dim 29** (same as k=50, k=100) × **Dim 34** (same as k=100) × **Dim 35** (same as k=100) × **Dim 47** (same as k=50) × **Dim 16, 25** (negative values)
K=200	✓ **Dim 8** · ____ 0.4792 – top dimension · 2nd with 0.4785 – alarm · 3rd with 0.4325 – note · 10th with 0.3328 – calendar · 20th with 0.2848 – task · 25th with 0.2632 – organizer · 27th with 0.2503 – calculator · 38th ... → 3 negative values × **Dim 23** (same as k=50, k=100, k=150) × **Dim 35** (same as k=50, k=100, k=150) × **Dim 47** (same as k=50, k=150) × **Dim 52** (same as k=100) × **Dim 3, 4, 16, 22, 25, 29, 34, 57** (negative values)	× **Dim 25** (same as k=50) × **Dim 29** (same as k=50, k=100, k=150) × **Dim 3, 4, 8, 16, 22, 34, 35, 47** (negative values)

Table 9-5. Best scored dimensions of LSI considering the *organizer* product feature on sections.

9.2.3. Paragraphs

	Without	With
K=50	✓ **Dim 2** · ____ 1.1088 – top dimension · 2nd with 0.7872 – top dimension6th with 0.5376 – calendar · 8th with 0.5083 – alarm · 10th with 0.4843 – note · 20th with 0.2943 – task · 24th with 0.2677 – calculator · 28th with 0.2609 – organizer · 44th ... → No negative values × **Dim 8** · ____ 0.6447 – top dimension · 7th with 0.4354 – calendar · 10th with 0.4270 – alarm · 27th with 0.2550 – task · 38th ... → 3 negative values × **Dim 10** · ____ 0.7111 – top dimension · 9th with 0.3055 – calendar · 14th with 0.2837 – note · 20th with 0.2411 – office · 21st with 0.2342 – calculator · 30th with 0.2070 – pdf · 35th ... → 1 negative value ✓ **Dim 17** · ____ 0.7416 – top dimension · 11th with 0.2920 – office · 16th with 0.2757 – note · 27th with 0.2633 – alarm · 19th with 0.2567 – calculator · 25th with 0.2088 – memo · 26th with 0.2085 – pdf · 39th ... → No negative values × **Dim 18** · ____ 0.5267 – top dimension · 6th with 0.3125 – calendar · 8th with 0.3080 – note · 10th with 0.3023 – task · 223rd ... → 7 negative values ✓ **Dim 37** · 1st with 0.5605 – alarm · 8th with 0.2883 – calculator · 11th with 0.2441 – world clock · 22nd with 0.1781 – converter · 24th with 0.1755 – memo · 48th ... → 3 negative values × **Dim 49** · ____ 0.6058 – top dimension · 7th with 0.3039 – office · 13th with 0.2595 – pdf · 15th with 0.2388 – task · 41st ... → 7 negative values	✓ **Dim 2** · ____ 1.1045 – top dimension · 2nd with 0.7856 – top dimension · 6th with 0.5404 – calendar · 7th with 0.5124 – alarm · 10th with 0.4870 – note · 19th with 0.2981 – task · 23rd with 0.2723 – calculator · 26th with 0.2642 – organizer · 42nd ... → No negative values × **Dim 10** · ____ 0.7232 – top dimension · 10th with 0.2896 – calendar · 13th with 0.2792 – note · 18th with 0.2512 – office · 21st with 0.2369 – calculator · 24th with 0.2146 – pdf · 38th ... → 1 negative value ✓ **Dim 17** · ____ 0.7348 – top dimension · 9th with 0.3060 – office · 17th with 0.2695 – note · 18th with 0.2559 – calculator · 20th with 0.2550 – alarm · 24th with 0.2209 – pdf · 26th with 0.2122 – memo · 41st ... → No negative values × **Dim 18** · ____ 0.5219 – top dimension · 7th with 0.3104 – calendar · 9th with 0.3015 – note · 10th with 0.2988 – task · 243rd ... → 9 negative values × **Dim 35** · ____ 0.4592 – top dimension · 7th with 0.2884 – alarm · 10th with 0.2586 – organizer · 15th with 0.2372 – calculator · 26th with 0.1729 – converter · 32th ... → 1 negative value × **Dim 49** · ____ 0.6380 – top dimension · 9th with 0.3014 – office · 12th with 0.2584 – pdf · 16th with 0.2328 – task · 39th ... → 7 negative values
K=100	✓ **Dim 2** (same as k=50) ✓ **Dim 17** (same as k=50)	× **Dim 10** (same as k=50) ✓ **Dim 17** (same as k=50)

	✓ **Dim 37** (same as k=50) × **Dim 47** · 1st with 0.3975 – task · 15th with 0.1987 – calculator · 21st with 0.1791 – alarm · 30th with 0.1576 – organizer · 84th ... → 2 negative values × **Dim 49** (same as k=50) × **Dim 58** · ____ 0.5859 – top dimension · 3rd with 0.4545 – task · 5th with 0.3158 – note · 83rd ... → 8 negative values × **Dim 81** · ____ 0.4256 – top dimension · 3rd with 0.3378 – note · 12th with 0.2511 – calendar · 32nd ... → 2 negative values × **Dim 8, 10, 18** (negative values)	× **Dim 35** (same as k=50) ✓ **Dim 37** · ____ 0.5454 – top dimension · 2nd with 0.5331 – alarm · 6th with 0.3589 – calculator · 8th with 0.3248 – world clock · 14th with 0.2406 – memo · 18th with 0.2325 – converter · 30th with 0.1688 – note · 42nd ... → 3 negative values × **Dim 57** · ____ 0.4633 – top dimension · 6th with 0.2600 – calendar · 21st with 0.2089 – alarm · 29th with 0.1842 – calculator · 54th ... → 4 negative values × **Dim 2, 18, 49** (negative values)
K=150	× **Dim 8** (same as k=50) ✓ **Dim 9** · ____ 0.7830 – top dimension · 2nd with 0.7422 – alarm · 3rd with 0.7283 – calendar · 5th with 0.5353 – task · 8th with 0.4337 – note · 10th with 0.4056 – calculator · 15th with 0.3247 – organizer · 25th with 0.2517 – world clock · 27th with 0.2448 – converter · 33rd ... → No negative values ✓ **Dim 37** (same as k=50, k=100) × **Dim 47** (same as k=100) × **Dim 57** · ____ 0.4586 – top dimension · 5th with 0.2624 – calendar · 13th with 0.2373 – alarm · 26th with 0.1970 – calculator · 49th ... → 4 negative values × **Dim 81** (same as k=100) × **Dim 2, 10, 17, 18, 49, 58** (negative values)	× **Dim 8** · ____ 0.6430 – top dimension · 7th with 0.4518 – calendar · 9th with 0.4445 – alarm · 24th with 0.2682 – task · 37th ... → 3 negative values ✓ **Dim 9** · ____ 0.7613 – top dimension · 2nd with 0.7443 – alarm · 3rd with 0.7267 – calendar · 5th with 0.5399 – task · 7th with 0.4428 – note · 9th with 0.4200 – calculator · 15th with 0.3327 – organizer · 25th with 0.2689 – converter · 26th with 0.2630 – world clock · 29th with 0.2425 – memo · 37th ... → No negative values ✓ **Dim 17** (same as k=50, k=100) × **Dim 35** (same as k=50, k=100) × **Dim 2, 10, 18, 37, 49, 57** (negative values)
K=200	× **Dim 10** (same as k=50) ✓ **Dim 17** (same as k=50, k=100) ✓ **Dim 37** (same as k=50, k=100, k=150) × **Dim 47** (same as k=100, k=150) × **Dim 58** (same as k=100) × **Dim 2, 8, 9, 18, 49, 57, 81** (negative values)	× **Dim 10** (same as k=50) ✓ **Dim 17** (same as k=50, k=100, k=150) × **Dim 47** · ____ 0.4220 – top dimension · 2nd with 0.3551 – task · 16th with 0.1836 – calculator · 38th ... → No negative values × **Dim 57** (same as k=100) × **Dim 2, 8, 9, 18, 35, 37, 49** (negative values)

Table 9-6. Best scored dimensions of LSI considering the *organizer* product feature on paragraphs.

9.3. *Multimedia*

9.3.1. Documents

	Without	With
K=50	× **Dim 1** · ____ 0.6014 – top dimension · 30th with 0.3919 – video · 76th with 0.3150 – multimedia · 146th ... → No negative values	× **Dim 1** · ____ 0.6026 – top dimension · 31st with 0.3913 – video · 76th with 0.3155 – multimedia · 147th ... → No negative values
K=100	× **Dim 1** (same as k=50) × **Dim 20** · ____ 0.4214 – top dimension · 41st with 0.1253 – aac · 65th with 0.1102 – h264 · 72nd with 0.1095 – eaac+ · 149th ... → 3 negative values × **Dim 93** · ____ 0.2363 – top dimension · 31st with 0.1236 – album art cover · 130th ... → 12 negative values	× **Dim 1** (same as k=50) × **Dim 20** · ____ 0.4193 – top dimension · 38th with 0.1272 – aac · 64th with 0.1121 – h264 · 72nd with 0.1102 – radio · 138th ... → 3 negative values × **Dim 93** · ____ 0.2344 – top dimension · 28th with 0.1249 – album art cover · 134th ... → 11 negative values
K=150	× **Dim 1** (same as k=50, k=100) × **Dim 2** · ____ 0.6011 – top dimension · 32nd with 0.2671 – xvid · 37th with 0.2524 – divx · 195th ... → 17 negative values × **Dim 20** (same as k=100) × **Dim 65** · ____ 0.2232 – top dimension · 26th with 0.1188 – mp3 · 51st with 0.0947 – mp4 · 112th ... → 7 negative values × **Dim 93** (same as k=100)	× **Dim 1** (same as k=50, k=100) × **Dim 20** (same as k=100) × **Dim 65** · ____ 0.2232 – top dimension · 26th with 0.1188 – mp3 · 47th with 0.0947 – mp4 · 107th ... → 7 negative values × **Dim 93** (negative values)
K=200	× **Dim 1** (same as k=50, k=100, k=150) × **Dim 2** (same as k=150) × **Dim 65** (same as k=150) × **Dim 93** (same as k=50, k=100) × **Dim 20** (negative values)	× **Dim 1** (same as k=50, k=100, k=150) × **Dim 20, 65, 93** (negative values)

Table 9-7. Best scored dimensions of LSI considering the *multimedia* product feature on documents.

9.3.2. Sections

	Without	With
K=50	✓ **Dim 5** · 1st with 0.7405 – video · 18th with 0.3024 – radio · 22nd with 0.2826 – audio · 30th with 0.2637 – multimedia · 32nd ... → No negative values ✓ **Dim 6** · ____ 0.5263 – top dimension · 4th with 0.3446 – music player · 10th with 0.2879 – radio · 17th with 0.2656 – video · 28th with 0.2143 – fm · 35th with 0.2003 – mp3 · 44th ... → 3 negative values ✓ **Dim 11** · ____ 0.5212 – top dimension · 5th with 0.3482 – music player · 7th with 0.3137 – radio · 12th with 0.2691 – fm · 27th with 0.1929 – multimedia · 32nd with 0.1844 – mp3 · 37th with 0.1654 – video player · 40th with 0.1539 – stereo · 62nd ... → 3 negative values	✓ **Dim 5** · 1st with 0.7433 – video · 17th with 0.3065 – radio · 21st with 0.2860 – audio · 30th with 0.2664 – multimedia · 31st ... → No negative values ✓ **Dim 6** · ____ 0.5287 – top dimension · 4th with 0.3466 – music player · 9th with 0.2903 – radio · 17th with 0.2651 – video · 28th with 0.2186 – fm · 34th with 0.2042 – mp3 · 42nd ... → 2 negative values × **Dim 9** · ____ 0.5683 – top dimension · 11th with 0.3246 – fm · 12th with 0.3208 – radio · 21st with 0.2836 – music player · 101st ... → 1 negative value ✓ **Dim 11** · ____ 0.5130 – top dimension · 5th with 0.3456 – music player · 8th with 0.3101 – radio · 12th with 0.2685 – fm · 27th with 0.1923 – multimedia · 31st with 0.1882 – mp3 · 36th with 0.1674 – video player · 42nd with 0.1552 – stereo · 59th ... → 2 negative values
K=100	✓ **Dim 5** (same as k=50) × **Dim 9** · ____ 0.5622 – top dimension · 11th with 0.3165 – fm · 12th with 0.3153 – radio · 25th with 0.2785 – music player · 106nd ... → 1 negative value ✓ **Dim 11** (same as k=50) × **Dim 74** · ____ 0.2721 – top dimension · 3rd with 0.2293 – divx · 9th with 0.1944 – xvid · 28th with 0.1373 – audio · 40rd with 0.1219 – video playback · 161st ... → 16 negative values ✓ **Dim 91** · ____ 0.2721 – top dimension · 7th with 0.1924 – fm · 9th with 0.1815 – divx · 12th with 0.1718 – xvid · 19th with 0.1522 – radio · 34th ... → 7 negative values × **Dim 93** · ____ 0.2628 – top dimension · 4th with 0.2028 – divx	✓ **Dim 5** (same as k=50) × **Dim 9** (same as k=50) × **Dim 11** (same as k=50) × **Dim 74** · ____ 0.2733 – top dimension · 3rd with 0.2310 – divx · 8th with 0.1974 – xvid · 25th with 0.1415 – audio · 39rd with 0.1235 – video playback · 162nd ... → 15 negative values × **Dim 84** · ____ 0.2734 – top dimension · 4th with 0.2227 – divx · 12th with 0.1890 – video · 13th with 0.1878 – xvid · 54th with 0.1113 – video playback · 81st ... → 18 negative values ✓ **Dim 91** · ____ 0.3272 – top dimension · 7th with 0.1996 – divx · 8th with 0.1914 – xvid · 9th with 0.1913 – fm · 12th with 0.1678 – radio · 17th with 0.1548 – video player · 32nd ... → 3 negative values

	· 7th with 0.1902 – xvid · 18th with 0.1528 – video · 23rd with 0.1401 – radio · 48th ... → 8 negative values × **Dim 95** · ____ 0.3355 – top dimension · 6th with 0.2129 – fm · 15th with 0.1569 – radio · 41st ... → 14 negative values × **Dim 6** (negative values)	× **Dim 93** · ____ 0.2630 – top dimension · 10th with 0.1743 – divx · 12th with 0.1699 – xvid · 25th with 0.1419 – video · 29th with 0.1381 – radio · 46th ... → 10 negative values × **Dim 6** (negative values)
K=150	✓ **Dim 6** (same as k=50) ✓ **Dim 9** (same as k=100) ✓ **Dim 11** (same as k=50, k=100) × **Dim 63** · ____ 0.3254 – top dimension · 6th with 0.2048 – divx · 10th with 0.1849 – xvid · 33rd ... → 20 negative values ✓ **Dim 91** (same as k=100) × **Dim 93** (same as k=100) × **Dim 137** · 1st with 0.2525 – fm · 2nd with 0.2383 – radio · 35th ... → 19 negative values × **Dim 140** · ____ 0.2434 – top dimension · 6th with 0.1765 – radio · 18th with 0.1400 – fm · 22nd with 0.1323 – multimedia · 26th with 0.1257 – video · 49th ... → 11 negative values × **Dim 5, 74, 95** (negative values)	✓ **Dim 5** (same as k=50, k=100) × **Dim 63** · ____ 0.3195 – top dimension · 6th with 0.2061 – divx · 9th with 0.1858 – xvid · 26th with 0.1456 – music player · 34th ... → 18 negative values × **Dim 74** (same as k=100) × **Dim 84** (same as k=100) × **Dim 137** · 1st with 0.2594 – fm · 2nd with 0.2397 – radio · 32nd ... → 18 negative values × **Dim 140** · ____ 0.2449 – top dimension · 5th with 0.1773 – radio · 17th with 0.1395 – fm · 30th with 0.1228 – multimedia · 47th ... → 6 negative values × **Dim 6, 9, 11, 91, 93** (negative values)
K=200	✓ **Dim 5** (same as k=50, k=100) ✓ **Dim 6** (same as k=50, k=150) × **Dim 74** (same as k=100) ✓ **Dim 91** (same as k=100, k=150) × **Dim 95** (same as k=100) × **Dim 137** (same as k=150) × **Dim 186** · ____ 0.2407 – top dimension · 8th with 0.1708 – stereo 17th with 0.1425 – divx · 27th with 0.1170 – xvid · 30th with 0.1123 – multimedia · 202nd ... → 16 negative values × **Dim 9, 11, 63, 93, 140** (negative values)	✓ **Dim 5** (same as k=50, k=100, k=150) ✓ **Dim 6** (same as k=50) × **Dim 63** (same as k=150) × **Dim 93** (same as k=100) × **Dim 140** (same as k=150) × **Dim 9, 11, 63, 74, 84, 91, 137** (negative values)

Table 9-8. Best scored dimensions of LSI considering the *multimedia* product feature on sections.

9.3.3. Paragraphs

	Without	With
K=50	× **Dim 5** · ____ 0.8356 – top dimension · 6th with 0.4779 – video · 8th with 0.4424 – music player · 24th with 0.2587 – mp3 · 29th with 0.2291 – radio · 40th ... → 1 negative value × **Dim 26** · 1st with 0.4118 – video · 9th with 0.2195 – divx · 16th with 0.1870 – xvid · 51st ... → 16 negative values × **Dim 37** · ____ 0.5605 – top dimension · 2nd with 0.5159 – radio · 4th with 0.4100 – fm · 13th with 0.2350 – video · 34th ... → 6 negative values × **Dim 44** · ____ 0.4861 – top dimension · 5th with 0.3783 – radio · 9th with 0.2938 – fm · 31st ... → 1 6 negative values	× **Dim 3** · 1st with 0.9619 – video · 19th with 0.2678 – multimedia · 21st with 0.2521 – music player · 23rd with 0.2462 – divx · 34th ... → 1 negative value × **Dim 5** · ____ 0.7795 – top dimension · 5th with 0.4671 – video · 8th with 0.4240 – music player · 24th with 0.2501 – mp3 · 33th ... → No negative value × **Dim 15** · 1st with 0.8543 – video · 17th with 0.2851 – divx · 21st with 0.2482 – xvid · 33rd ... → 10 negative value × **Dim 44** · ____ 0.4861 – top dimension · 5th with 0.3783 – radio · 9th with 0.2938 – fm · 26th with 0.1406 – multimedia · 38th ... → 1 2 negative values
K=100	× **Dim 3** · 1st with 0.9544 – video · 19th with 0.2607 – multimedia · 28th with 0.2398 – divx · 29th with 0.2355 – music player · 34th ... → 1 negative value × **Dim 5** (same as k=50) × **Dim 15** · 1st with 0.8736 – video · 17th with 0.2898 – divx · 20th with 0.2517 – xvid · 34th ... → 10 negative value × **Dim 26** (same as k=50) × **Dim 37** (same as k=50) × **Dim 39** · 1st with 0.6561 – radio · 2nd with 0.5462 – fm · 16th with 0.2292 – rds · 26th with 0.1856 – video · 57th ... → 12 negative value × **Dim 44** (same as k=50) × **Dim 51** · ____ 0.4007 – top dimension · 8th with 0.3011 – radio · 10th with 0.2790 – music player · 11th with 0.2786 – fm · 57th ... → 24 negative values	× **Dim 13** · ____ 0.4986 – top dimension · 2nd with 0.4704 – radio · 19th with 0.2268 – divx · 29th with 0.1989 – xvid · 78th ... → 5 negative values × **Dim 15** (same as k=50) × **Dim 26** · 1st with 0.4127 – video · 9th with 0.2191 – divx · 16th with 0.1866 – xvid · 51st ... → 22 negative values × **Dim 37** · ____ 0.5454 – top dimension · 9th with 0.2979 – radio · 15th with 0.2373 – fm · 41st ... → 4 negative values × **Dim 38** · ____ 0.4309 – radio · 4th with 0.3472 – fm · 14th with 0.1212 – rds · 37th with 0.1210 – video · 357th ... → 14 negative values × **Dim 51** · ____ 0.4011 – top dimension · 8th with 0.3018 – radio · 10th with 0.2797 – fm · 11th with 0.2779 – music player · 57th ... → 24 negative values × **Dim 65** · ____ 0.8820 – top dimension · 7th with 0.2830 – multimedia · 11th with 0.1849 – fm · 18th with 0.2324 – radio

		· 52nd ... → 8 negative values × **Dim 67** · ____ 0.4165 – top dimension · 9th with 0.2330 – fm · 10th with 0.2319 – radio · 42nd ... → 8 negative values × **Dim 3, 5, 44** (negative values)
K=150	× **Dim 26** (same as k=50, k=100) × **Dim 37** (same as k=50, k=100) × **Dim 39** (same as k=100) × **Dim 3, 5, 15, 44, 51** (negative values)	× **Dim 13** (same as k=100) × **Dim 38** (same as k=100) × **Dim 44** (same as k=50) × **Dim 51** (same as k=100) × **Dim 65** (same as k=100) × **Dim 67** (same as k=100) × **Dim 105** · ____ 0.3330 – top dimension · 3rd with 0.3093 – divx · 7th with 0.2580 – xvid · 64th ... → 10 negative values × **Dim 115** · ____ 0.5488 – top dimension · 3rd with 0.2569 – multimedia · 10th with 0.1857 – divx · 18th with 0.1643 – xvid · 147th ... → 17 negative values × **Dim 3, 5, 15, 26, 37** (negative values)
K=200	× **Dim 5** (same as k=50, k=100) × **Dim 15** (same as k=100) × **Dim 26** (same as k=50, k=100, k=150) × **Dim 37** (same as k=50, k=100, k=150) × **Dim 51** (same as k=100) × **Dim 3, 39, 44** (negative values)	× **Dim 5** (same as k=50) × **Dim 15** (same as k=50, k=100) × **Dim 38** (same as k=100, k=150) × **Dim 44** (same as k=50, k=150) × **Dim 51** (same as k=100, k=150) × **Dim 105** (same as k=150) × **Dim 3, 13, 26, 37, 65, 67, 115** (negative values)

Table 9-9. Best scored dimensions of LSI considering the *multimedia* product feature on paragraphs.

10. Annex B: PLSI's running tables

10.1. *Battery*

10.1.1. Documents

	Without	With
K=50	× **Topic 41** · 56th with 0.4573 – stand-by time · 188th with 0.0048 – talk time · 210th ... → 5 values set to zero	× **Topic 11** · 30th with 0.9636 – video playback · 136th with 0.0221 – battery · 505th ... → 6 values set to zero × **Topic 19** · 27th with 0.8675 – li-ion · 178th ... → 8 values set to zero
K=100	× **Topic 8** · 10th with 0.9957 – stand-by time · 192nd ... → 2 values set to zero × **Topic 26** · 26th with 0.4636 – battery · 204th ... → 2 values set to zero	× **Topic 83** · 36th with 0.0084 – stand-by time · 93rd ... → 5 values set to zero × **Topic 98** · 13th with 0.9801 – video playback · 92nd ... → 2 values set to zero
K=150	× **Topic 60** · 2nd with 0.0203 – li-ion · 863rd ... → 2 values set to zero ✓ **Topic 89** · 5th with 0.3105 – battery · 180th with 0.0439 – talk time · 477th ... → 4 values set to zero	✓ **Topic 18** · 14th with 0.0099 – stand-by time · 19th with 0.0036 – capacity · 32nd ... → NO values set to zero × **Topic 46** · 15th with 0.2037 – battery · 265th ... → 5 values set to zero × **Topic 59** · 1st with 0.0239 – talk time · 698th ... → 3 values set to zero
K=200	✓ **Topic 4** · 3rd with 0.9927 – stand-by time · X ... → ALL values set to zero ✓ **Topic 154** · 12th with 0.2949 – battery · 48th ... → 3 values set to zero	× **Topic 36** · 27th with 0.2305 – battery · 516th ... → 6 values set to zero × **Topic 65** · 2nd with 0.0179 – battery · 851st ... → 10 values set to zero

Table 10-1. Best scored topics of PLSI considering the *battery* product feature on documents.

10.1.2. Sections

	Without	With
K=50	× **Topic 43** · 33rd with 0.9594 – battery · 406th ... → 5 values set to zero	
K=100	× **Topic 7** · 12th with 0.0810 – battery · X ... → ALL values set to zero	× **Topic 99** · 13th with 0.9760 – battery · 164th ... → 8 value set to zero
K=150	× **Topic 14** · 18th with 0.9808 – battery · X ... → ALL values set to zero	✓ **Topic 95** · 20th with 0.9600 – battery · 257th ... → 9 values set to zero × **Topic 122** · 28th with 0.0228 – li-polymer · 1350th ... → 1 value set to zero
K=200	✓ **Topic 92** · 8th with 0.0180 – li-ion · 3073rd ... → 1 value set to zero	× **Topic 107** · 13th with 0.9593 – li-ion · X ... → ALL values set to zero

Table 10-2. Best scored topics of PLSI considering the *battery* product feature on sections.

10.1.3. Paragraphs

	Without	With
K=50		× **Topic 2** · 3rd with 1.0000 – li-ion · 292nd ... → 8 values set to zero × **Topic 42** · 3rd with 0.9978 – video call time · 224th ... → 9 values set to zero
K=100		× **Topic 42** · 27th with 1.0000 – video call time · X ... → ALL values set to zero
K=150	× **Topic 19** · 32nd with 0.4999 – li-polymer · 145th ... → 6 values set to zero	× **Topic 94** · 3rd with 1.0000 – li-ion · 406th ... → 9 values set to zero
K=200	× **Topic 141** · 27th with 0.0403 – battery · X ... → ALL values set to zero × **Topic 149** · 13th with 1.0000 – li-polymer · X ... → ALL values set to zero	× **Topic 57** · 20th with 1.0000 – battery · X ... → ALL values set to zero × **Topic 77** · 4th with 0.9943 – talk time 3g · X ... → ALL values set to zero × **Topic 156** · 14th with 0.9271 – video playback · 523rd ... → 3 values set to zero

Table 10-3. Best scored topics of PLSI considering the *battery* product feature on paragraphs.

10.2. Organizer

10.2.1. Documents

	Without	With
K=50	× **Topic 2** · 22nd with 0.0375 – organizer · 53rd ... → 14 values set to zero	✓ **Topic 45** · 18th with 0.6394 – note · 70th ... → 5 values set to zero
K=100	× **Topic 25** · 28th with 0.9872 – calendar · 87th ... → 6 values set to zero × **Topic 26** · 12th with 0.8447 – calculator · 187th ... → 2 values set to zero	× **Topic 43** · 8th with 0.9899 – document viewer · 153rd ... → 5 values set to zero × **Topic 56** · 26th with 0.8402 – world clock · 61st ... → 9 values set to zero
K=150	✓ **Topic 119** · 20th with 0.9762 – calendar · 105th ... → 5 values set to zero ✓ **Topic 135** · 2nd with 0.9999 – document viewer · 82nd ... → 9 values set to zero	× **Topic 6** · 28th with 0.1155 – note · 34th ... → 1 values set to zero ✓ **Topic 94** · 29th with 0.9993 – document viewer · 78th ... → 5 values set to zero
K=200	× **Topic 96** · 1st with 0.9962 – stopwatch · 129th ... → 4 values set to zero × **Topic 148** · 1st with 0.9999 – document viewer · 83rd ... → 4 values set to zero ✓ **Topic 176** · 13th with 0.3649 – alarm · 19th with 0.3087 – note · 199th ... → 9 values set to zero	× **Topic 48** · 14th with 0.8473 – task · 30th with 0.7946 – calendar · 112th ... → 2 values set to zero ✓ **Topic 130** · 9th with 0.9997 – document viewer · 78th ... → 5 values set to zero × **Topic 186** · 14th with 0.0119 – task · 28th with 0.0073 – note · 42nd ... → 12 values set to zero

Table 10-4. Best scored topics of PLSI considering the *organizer* product feature on documents.

10.2.2. Sections

	Without	With
K=50	× **Topic 19** · 16th with 0.9953 – stopwatch · 30th with 0.9922 – calendar · 108th ... → 3 values set to zero × **Topic 27** · 1st with 1.0000 – document viewer · 12th with 0.9587 – world clock · 52nd ... → 9 values set to zero ✓ **Topic 28** · 23rd with 0.9789 – memo · 187th ... → 14 values set to zero	× **Topic 32** · 7th with 0.9945 – stopwatch · 14th with 0.9237 – world clock · 98th ... → 11 values set to zero × **Topic 48** · 30th with 0.9770 – calculator · 34th ... → 10 values set to zero
K=100	✓ **Topic 6** · 8th with 0.9969 – document viewer · 61st ... → 13 values set to zero × **Topic 12** · 5th with 0.9997 – calendar · 25th with 0.0273 – world clock · 49th ... → 9 values set to zero ✓ **Topic 49** · 7th with 0.9958 – calculator · 17th with 0.9923 – stopwatch · X ... → ALL values set to zero	✓ **Topic 11** · 6th with 0.9916 – stopwatch · 115th ... → NO values set to zero × **Topic 20** · 6th with 0.9924 – document viewer · 9th with 0.9914 – calculator · 70th ... → 14 values set to zero
K=150	× **Topic 14** · 14th with 0.9891 – calendar · 15th with 0.9862 – calculator · 36th ... → 10 values set to zero ✓ **Topic 60** · 2nd with 0.0963 – memo · 3rd with 0.0913 – office · 6th with 0.0674 – task · 41st ... → 4 values set to zero	× **Topic 9** · 17th with 0.0082 – note · 24th with 0.0079 – alarm · 1174th ... → 6 values set to zero × **Topic 82** · 13th with 0.0109 – converter · 116th ... → 5 values set to zero × **Topic 96** · 2nd with 0.9980 – document viewer · 48th ... → 14 values set to zero × **Topic 143** · 12th with 0.9904 – calculator · 18th with 0.9822 – memo · 91st ... → 9 values set to zero
K=200	× **Topic 16** · 6th with 0.0083 – organizer · 27th with 0.0065 – alarm · 307th ... → 7 values set to zero × **Topic 123** · 3rd with 0.0105 – world clock · 27th with 0.0074 – organizer · 129th ... → 5 values set to zero ✓ **Topic 154** · 1st with 0.9969 – calendar · 10th with 0.9154 – world clock · 86th ... → 11 values set to zero × **Topic 179** · 14th with 0.0018 – converter · 18th with 0.0012 – task · 44th ... → 13 values set to zero	× **Topic 89** · 13th with 0.2943 – memo · 25th with 0.0081 – task · 64th ... → 11 values set to zero × **Topic 125** · 3rd with 0.1378 – memo · 6th with 0.0485 – task · 120th ... → 3 values set to zero × **Topic 143** · 5th with 0.9958 – document viewer · 29th with 0.0277 – world clock · 73rd ... → 13 values set to zero ✓ **Topic 164** · 16th with 0.1391 – task · 26th with 0.0085 – convert · 62nd ... → 11 values set to zero

Table 10-5. Best scored topics of PLSI considering the *organizer* product feature on sections.

10.2.3. Paragraphs

	Without	With
K=50		
K=100	✓ **Topic 26** · 11th with 1.0000 – currency converter · 101st ... → 10 values set to zero	× **Topic 99** · 30th with 0.9997 – currency converter · X ... → ALL values set to zero
K=150	× **Topic 39** · 4th with 0.9957 – calendar · 234th ... → 13 values set to zero × **Topic 121** · 4th with 1.0000 – currency converter · 141st ... → 14 values set to zero	× **Topic 71** · 19th with 1.0000 – currency converter · X ... → ALL values set to zero ✓ **Topic 107** · 30th with 0.9998 – pdf · X ... → ALL values set to zero
K=200	× **Topic 8** · 19th with 0.9998 – stopwatch · X ... → ALL values set to zero × **Topic 98** · 36th with 0.0010 – note · 38th with 0.0007 – task · 40th with 0.0004 – calculator · 49th with 0.0002 – converter · 59th ... → 10 values set to zero × **Topic 125** · 21st with 0.9981 – to-do · 70th ... → 10 values set to zero	× **Topic 2** · 3rd with 0.9680 – office · 47th ... → 11 values set to zero × **Topic 101** · 1st with 0.9995 – stopwatch · 110th ... → 12 values set to zero

Table 10-6. Best scored topics of PLSI considering the *organizer* product feature on paragraphs.

10.3. *Multimedia*

10.3.1. Documents

	Without	With
K=50	× **Topic 2** · 17th with 0.0453 – radio · 21st with 0.0385 – fm · 78th ... → 33 values set to zero	✓ **Topic 38** · 11th with 0.9872 – eaac · 874th ... → 9 values set to zero ✓ **Topic 50** · 13th with 0.0903 – streaming · 33rd ... → 2 values set to zero
K=100	× **Topic 39** · 20th with 0.7578 – wmv · 67th ... → 15 values set to zero × **Topic 54** · 22nd with 0.2397 – wmv · 64th ... → 10 values set to zero	× **Topic 20 (example misspelling)** · 5th with 0.0395 – multimedia · 10th with 0.0392 – midi · 11th with 0.0391 – 3gpp · 834th ... → 10 values set to zero ✓ **Topic 41** · 30th with 0.0512 – midi · 245th ... → 13 values set to zero × **Topic 71** · 15th with 0.0238 – m4a · 453rd ... → 10 values set to zero
K=150	× **Topic 32** · 27th with 0.2374 – music player · 34th ... → 16 values set to zero ✓ **Topic 56** · 27th with 0.0121 – fm · 41st with 0.0079 – music player · 46th ... → 24 values set to zero × **Topic 119** · 9th with 0.9914 – avi · 69th ... → 23 values set to zero	× **Topic 105** · 22nd with 0.2397 – wmv · 64th ... → 30 values set to zero × **Topic 114** · 22nd with 0.0126 – xvid · 25th with 0.0111 – video · 31st ... → 30 values set to zero
K=200	✓ **Topic 25** · 8th with 0.9852 – avi · 57th ... → 20 values set to zero ✓ **Topic 45** · 29th with 0.0794 – fm · 38th ... → 17 values set to zero	× **Topic 156** · 13th with 0.2397 – streaming · 24th with 0.2397 – avi · 26th with 0.2397 – youtube player · 27th with 0.2397 – fm transmitter · 37th ... → 4 values set to zero × **Topic 170** · 9th with 0.2397 – music player · 27th with 0.2397 – streaming · 84th ... → 31 values set to zero

Table 10-7. Best scored topics of PLSI considering the *multimedia* product feature on documents.

10.3.2. Sections

	Without	With
K=50	× **Topic 26** · 28th with 0.0038 – fm · 34th ... → 31 values set to zero ✓ **Topic 31** · 7th with 0.99997 – mpeg4 · 89th ... → 19 values set to zero × **Topic 39** · 18th with 0.0013 – video · 19th with 0.0012 – divx · 27th with 0.0006 – xvid · 54th ... → 29 values set to zero	× **Topic 9** · 12th with 0.9934 – mpeg4 · 25th with 0.9789 – music player · 119th ... → 27 values set to zero × **Topic 18** · 11th with 0.1049 – youtube player · 26th with 0.1019 – realaudio · 219th ... → 6 values set to zero
K=100	✓ **Topic 49** · 3rd with 0.9965 – mpeg4 · 16th with 0.9928 – music player · 27th with 0.9705 – multimedia · 63rd ... → 20 values set to zero ✓ **Topic 59** · 12th with 0.9949 – radio · 81st ... → 23 values set to zero × **Topic 79** · 10th with 0.0022 – fm · 13th with 0.0017 – video · 19th with 0.0007 – divx · 22nd with 0.0006 – xvid · 38th ... → 31 values set to zero	× **Topic 13** · 13th with 0.8138 – video · 17th with 0.7060 – fm · 98th ... → 9 values set to zero × **Topic 64** · 21st with 0.0369 – ogg · 32nd ... → 4 values set to zero × **Topic 81** · 4th with 0.1771 – mp3 · 41st ... → 5 values set to zero
K=150	× **Topic 122** · 8th with 0.0095 – divx · 26th with 0.0086 – aac · 36th ... → 3 values set to zero × **Topic 149** · 2nd with 1.0000 – mpeg4 · 93rd ... → 33 values set to zero	× **Topic 26** · 19th with 0.9746 – video player · 23rd with 0.9028 – multimedia · 78th ... → 27 values set to zero ✓ **Topic 47** · 3rd with 0.9871 – mpeg4 · X ... → ALL values set to zero × **Topic 70** · 21st with 0.0031 – video · 22nd with 0.0029 – fm · 52nd ... → 28 values set to zero
K=200	× **Topic 14** · 18th with 0.0068 – fm · 28th with 0.0076 – xvid · 30th with 0.0076 – mp3 · 291st ... → 6 values set to zero ✓ **Topic 107** · 4th with 0.9970 – mpeg4 · 89th ... → 33 values set to zero ✓ **Topic 196** · 6th with 0.9872 – music player · 98th ... → 32 values set to zero	× **Topic 13** · 3rd with 0.0215 – eaac · 9th with 0.0209 – amr · 29th with 0.0202 – background playback · 338th ... → 7 values set to zero ✓ **Topic 40** · 7th with 0.9689 – video player · 40th ... → 37 values set to zero × **Topic 145** · 3rd with 0.0173 – amr · 87th ... → 6 values set to zero

Table 10-8. Best scored topics of PLSI considering the *multimedia* product feature on sections.

10.3.3. Paragraphs

	Without	With
K=50	× **Topic 50** · 24th with 1.0000 – streaming · 216th ... → 26 values set to zero	× **Topic 13** · 11th with 1.0000 – avi · 272nd ... → 34 values set to zero × **Topic 37** · 9th with 1.0000 – wmv · 290th ... → 30 values set to zero
K=100	✓ **Topic 10** · 2nd with 1.0000 – avi · 81st ... → 33 values set to zero × **Topic 60** · 10th with 1.0000 – fm transmitter · 87th ... → 34 values set to zero	✓ **Topic 21** · 16th with 1.0000 – 3gpp · 18th with 1.0000 – m4a · 152nd ... → 39 values set to zero × **Topic 92** · 14th with 1.0000 – recording option · 131st ... → 39 values set to zero
K=150	× **Topic 35** · 2nd with 1.0000 – avi · 135th ... → 33 values set to zero × **Topic 100** · 1st with 1.0000 – midi · 62nd ... → 33 values set to zero	× **Topic 33** · 6th with 1.0000 – mp4 · 103rd ... → 39 values set to zero ✓ **Topic 107** · 18th with 1.0000 – realaudio · 33rd with 0.00007 – video · X ... → ALL values set to zero
K=200	✓ **Topic 153 (exception)** · 33rd with 0.9970 – video · 76th ... → 34 values set to zero ·	× **Topic 43** · 7th with 1.0000 – 3gpp · 8th with 1.0000 – m4a · X ... → ALL values set to zero × **Topic 57** · 28th with 0.9985 – radio · 29th with 0.9985 – youtube player · 41st ... → 39 values set to zero × **Topic 83** · 35th with 0.0018 – video player · 37th with 0.0011 – video · 46th with 0.0002 – divx · 49th with 0.0001 – multimedia · X ... → ALL values set to zero

Table 10-9. Best scored topics of PLSI considering the *multimedia* product feature on paragraphs.

11. Annex C: LDA's running tables

11.1. *Battery*

11.1.1. Documents

	Without	With
K=50	✓ **Topic 3** · 8th with 0.0123 – battery · 27th with 0.0054 – talk time · 883rd ... → 4 values set to zero	× **Topic 2** · 16th with 0.0081 – battery · 34th with 0.0048 – talk time · 621st ... → 8 values set to zero × **Topic 4** · 9th with 0.0113 – battery · 105th ... → 2 values set to zero
K=100	× **Topic 1** · 15th with 0.0453 – radio · 57th ... → 6 values set to zero × **Topic 56** · 9th with 0.0453 – radio · 80th ... → 6 values set to zero	✓ **Topic 1** · 18th with 0.0071 – battery · 28th with 0.0047 – talk time · 485th ... → 5 values set to zero ✓ **Topic 44** · 21st with 0.0081 – battery · 88th ... → 9 values set to zero
K=150	× **Topic 75** · 16th with 0.0164 – battery · X ... → ALL values set to zero × **Topic 127** · 9th with 0.0148 – battery · 187th ... → 6 values set to zero	× **Topic 4** · 4th with 0.0141 – battery · 155th ... → 6 values set to zero × **Topic 70** · 13th with 0.0086 – battery · 112th ... → 9 values set to zero ✓ **Topic 134** · 19th with 0.0107 – battery · X ... → ALL values set to zero
K=200	× **Topic 168** · 16th with 0.0081 – battery · X ... → ALL values set to zero × **Topic 170** · 14th with 0.0125 – battery · 66th ... → 6 values set to zero	× **Topic 87** · 16th with 0.0099 – battery · X ... → ALL values set to zero × **Topic 152** · 5th with 0.0198 – battery · X ... → ALLL values set to zero

Table 11-1. Best scored topics of LDA considering the *battery* product feature on documents.

11.1.2. Sections

	Without	With
K=50	× **Topic 3** · 8th with 0.0313 – battery · 15th with 0.0120 – talk time · 144th ... → 1 value set to zero	✓ **Topic 3** · 7th with 0.0293 – battery · 16th with 0.0114 – talk time · 140th ... → 1 values set to zero
K=100	✓ **Topic 3** · 8th with 0.0333 – battery · 15th with 0.0128 – talk time · 621st ... → 1 value set to zero × **Topic 77** · 16th with 0.0197 – battery · X ... → ALL values set to zero	✓ **Topic 3** · 8th with 0.0249 – battery · 15th with 0.0126 – talk time · 143rd ... → 1 value set to zero × **Topic 74** · 18th with 0.0081 – battery · X ... → ALL values set to zero × **Topic 88** · 2nd with 0.0678 – battery · X ... → ALL values set to zero
K=150	× **Topic 86** · 2nd with 0.0400 – battery · X ... → ALL values set to zero × **Topic 102** · 18th with 0.0137 – battery · X ... → ALL values set to zero ✓ **Topic 122** · 2nd with 0.0879 – battery · 8th with 0.0360 – talk time · 99th ... → 4 values set to zero	✓ **Topic 2** · 8th with 0.0288 – battery · 14th with 0.0140 – talk time · 151st ... → 2 values set to zero ✓ **Topic 36** · 1st with 0.0602 – battery · 41st with 0.0054 – talk time · X ... → ALL values set to zero .
K=200	× **Topic 91** · 2nd with 0.0683 – battery · 57th ... → 5 values set to zero ✓ **Topic 106** · 2nd with 0.0889 – battery · 9th with 0.0429 – talk time · 54th ... → 5 values set to zero × **Topic 111** · 23rd with 0.0133 – battery · 69th ... → 6 values set to zero × **Topic 113** · 17th with 0.0186 – battery · X ... → ALL values set to zero	× **Topic 59** · 24th with 0.0128 – battery · X ... → ALL values set to zero ✓ **Topic 114** · 4th with 0.0600 – battery · 10th with 0.0329 – talk time · X ... → ALL values set to zero × **Topic 172** · 2nd with 0.0854 – battery · X ... → ALL values set to zero

Table 11-2. Best scored topics of LDA considering the *battery* product feature on sections.

11.1.3. Paragraphs

	Without	With
K=50	✓ **Topic 26** · 2nd with 0.1138 – battery · 7th with 0.0332 – talk time · 36th ... → 2 values set to zero	✓ **Topic 27** · 2nd with 0.1149 – battery · 7th with 0.0332 – talk time · 33rd ... → 4 values set to zero
K=100	✓ **Topic 29** · 1st with 0.2064 – battery · 16th with 0.0108 – talk time · 34th with 0.0041 – li-ion · X ... → ALL values set to zero ✓ **Topic 43** · 4th with 0.0703 – battery · 5th with 0.0606 – talk time · 50th ... → 5 values set to zero ✓ **Topic 62** · 25th with 0.0114 – battery · 56th ... → 6 values set to zero	× **Topic 23** · 4th with 0.0081 – battery · X ... → ALL values set to zero ✓ **Topic 48** · 4th with 0.0859 – battery · 5th with 0.0607 – talk time · 22nd with 0.0055 – capacity · 36th ... → 6 values set to zero × **Topic 59** · 1st with 0.2920 – battery · X ... → ALL values set to zero × **Topic 61** · 24th with 0.0116 – battery · 64th ... → 9 values set to zero
K=150	✓ **Topic 86** · 3rd with 0.1143 – battery · 7th with 0.0537 – talk time · 48th ... → 5 values set to zero	✓ **Topic 35** · 2nd with 0.0883 – battery · 6th with 0.0399 – talk time · 38th ... → 6 values set to zero × **Topic 81** · 18th with 0.0177 – battery · 44th ... → 9 values set to zero
K=200	× **Topic 61** · 16th with 0.0081 – battery · 34th with 0.0048 – talk time · 621st ... → 8 values set to zero × **Topic 93** · 1st with 0.1390 – battery · 8th with 0.0603 – talk time · X ... → ALL values set to zero × **Topic 162** · 2nd with 0.0813 – battery · 8th with 0.0288 – talk time · 34th ... → 5 values set to zero	× **Topic 71** · 1st with 0.0662 – battery · 62nd ... → 8 values set to zero × **Topic 74** · 3rd with 0.0839 – battery · 5th with 0.0654 – talk time · X ... → ALL values set to zero

Table 11-3. Best scored topics of LDA considering the *battery* product feature on paragraphs.

11.2. Organizer

11.2.1. Documents

	Without	With
K=50	× **Topic 38** · 17th with 0.0076 – alarm · 60th ... → 2 values set to zero × **Topic 40** · 13th with 0.0070 – alarm · 14th with 0.0065 – note · 62nd ... → 8 values set to zero	× **Topic 28** · 18th with 0.0069 – calendar · 25th with 0.0051 – note · 60th ... → 8 values set to zero × **Topic 39** · 7th with 0.0099 – alarm · 21st with 0.0059 – note · 38th ... → NO values set to zero
K=100	× **Topic 51** · 14th with 0.0101 – note · 15th with 0.0096 – alarm · 37th ... → 8 values set to zero	× **Topic 38** · 9th with 0.0097 – alarm · 12th with 0.0089 – note · 60th ... → 9 values set to zero
K=150	× **Topic 84** · 22nd with 0.0054 – alarm · 24th with 0.0054 – note · 25th with 0.0052 – calculator · 51st ... → 7 values set to zero × **Topic 90** · 7th with 0.0110 – note · 11th with 0.0087 – organizer · 17th with 0.0079 – calendar · 24th with 0.0072 – calculator · 72nd ... → 7 values set to zero	× **Topic 115** · 6th with 0.0081 – office · 27th with 0.0048 – calendar · 37th ... → 6 values set to zero ✓ **Topic 144** · 18th with 0.0104 – alarm · 41st ... → 6 values set to zero
K=200	× **Topic 92** · 24th with 0.0055 – office · 25th with 0.0055 – task · 29th with 0.0055 – note · 36th ... → 6 values set to zero ✓ **Topic 159** · 8th with 0.0188 – calendar · 11th with 0.0172 – alarm · 12th with 0.0166 – world clock · 13th with 0.0165 – converter · 14th with 0.0161 – organizer · 18th with 0.0121 – task · 21st with 0.0115 – calculator · 25th with 0.0111 – note · X ... → ALL values set to zero	× **Topic 97** · 25th with 0.0055 – task · 28th with 0.0055 – note · 29th with 0.0055 – office · 37th ... → 7 values set to zero ✓ **Topic 131** · 16th with 0.0100 – note · 21st with 0.0085 – calendar · 45th ... → 11 values set to zero

Table 11-4. Best scored topics of LDA considering the *organizer* product feature on documents.

11.2.2. Sections

	Without	With
K=50	× **Topic 9** · 5th with 0.0267 – office · 13th with 0.0147 – pdf · 85th ... → 12 values set to zero × **Topic 42** · 3rd with 0.0273 – task · 5th with 0.0192 – note · 8th with 0.0170 – calendar · 24th with 0.0092 – calculator · 29th with 0.0073 – alarm · 85th ... → 7 values set to zero ✓ **Topic 43** · 1st with 0.0494 – alarm · 2nd with 0.0428 – calendar · 3rd with 0.0391 – organizer · 5th with 0.0322 – note · 9th with 0.0231 – calculator · 14th with 0.0161 – task · 17th with 0.0133 – world clock · 20th with 0.0120 – converter · 21st with 0.0116 – alarm · 34th ... → 2 values set to zero .	✓ **Topic 8** · 1st with 0.0407 – alarm · 2nd with 0.0385 – calendar · 3rd with 0.0331 – organizer · 5th with 0.0302 – note · 8th with 0.0209 – calculator · 13th with 0.0145 – task · 17th with 0.0113 – world clock · 19th with 0.0112 – memo · 21st with 0.0109 – converter · 30th with 0.0063 – stopwatch · 63rd ... → 2 values set to zero × **Topic 11** · 7th with 0.0227 – office · 13th with 0.0128 – pdf · 94th ... → 11 values set to zero × **Topic 44** · 2nd with 0.0254 – task · 3rd with 0.0184 – note · 7th with 0.0164 – calendar · 10th with 0.0096 – alarm · 22nd with 0.0084 – calculator · 27th with 0.0046 – organizer · 60th ... → 8 values set to zero
K=100	× **Topic 34** · 3rd with 0.0269 – task · 6th with 0.0192 – note · 11th with 0.0166 – calendar · 26th with 0.0085 – alarm · 27th with 0.0081 – calculator · 61st ... → 8 values set to zero ✓ **Topic 35** · 1st with 0.0508 – organizer · 2nd with 0.0500 – calendar · 3rd with 0.0489 – alarm · 5th with 0.0403 – note · 6th with 0.0299 – calculator · 10th with 0.0205 – task · 14th with 0.0159 – memo · 15th with 0.0155 – world clock · 18th with 0.0138 – converter · 25th with 0.0095 – stopwatch · 46th ... → 3 values set to zero ✓ **Topic 36** · 2nd with 0.0338 – alarm · 5th with 0.0246 – calendar · 13th with 0.0130 – note · 14th with 0.0120 – organizer · 17th with 0.0106 – task · 24th with 0.0082 – calculator · 53rd ... → 5 values set to zero × **Topic 87** · 2nd with 0.0788 – office · 4th with 0.0362 – pdf · 45th ... → 11 values set to zero .	✓ **Topic 33** · 1st with 0.0478 – alarm · 2nd with 0.0451 – calendar · 3rd with 0.0420 – organizer · 5th with 0.0350 – note · 8th with 0.0253 – calculator · 12th with 0.0177 – task · 17th with 0.0124 – memo · 18th with 0.0123 – converter · 19th with 0.0122 – world clock · 28th with 0.0075 – stopwatch · 58th ... → 3 values set to zero × **Topic 34** · 3rd with 0.0268 – battery · 6th with 0.0189 – talk time · 10th with 0.0168 – battery · 26th with 0.0087 – battery · 27th with 0.0084 – battery · 63rd ... → 8 values set to zero

K=150	× **Topic 30** · 1st with 0.0547 – alarm · 2nd with 0.0499 – organizer · 3rd with 0.0455 – calendar · 4th with 0.0331 – note · 8th with 0.0291 – calculator · 12th with 0.0183 – task · 15th with 0.0156 – world clock · 18th with 0.0139 – converter · 19th with 0.0130 – memo · 28th with 0.0079 – stopwatch · 55th ... → 2 values set to zero × **Topic 104** · 4th with 0.0300 – note · 5th with 0.0285 – alarm · 14th with 0.0149 – converter · 19th with 0.0125 – office · 28th with 0.0107 – calendar · 29th with 0.0103 – calculator · 31st ... → 6 values set to zero	× **Topic 29** · 1st with 0.0551 – alarm · 2nd with 0.0453 – calendar · 3rd with 0.0451 – organizer · 5th with 0.0343 – note · 8th with 0.0282 – calculator · 14th with 0.0179 – task · 16th with 0.0155 – world clock · 17th with 0.0148 – converter · 19th with 0.0127 – memo · 26th with 0.0084 – stopwatch · 52nd ... → 4 values set to zero × **Topic 93** · 2nd with 0.1124 – office · 4th with 0.0512 – pdf · 135th ... → 14 values set to zero × **Topic 117** · 3rd with 0.0533 – calendar · 5th with 0.0352 – task · 9th with 0.0244 – note · 28th with 0.0095 – organizer · 38th ... → 12 values set to zero
K=200	✓ **Topic 31** · 1st with 0.0557 – alarm · 2nd with 0.0533 – calendar · 3rd with 0.0424 – organizer · 6th with 0.0328 – note · 8th with 0.0285 – calculator · 10th with 0.0234 – task · 18th with 0.0144 – memo · 19th with 0.0124 – world clock · 22nd with 0.0109 – converter · 28th with 0.0094 – stopwatch · 50th ... → 4 values set to zero ✓ **Topic 54** · 5th with 0.0265 – note · 6th with 0.0260 – alarm · 14th with 0.0146 – office · 21st with 0.0105 – converter · 25th with 0.0091 – calendar · 27th with 0.0084 – organizer · 28th with 0.0080 – calculator · 29th with 0.0076 – to-do · 30th with 0.0075 – pdf · 131st ... → 6 values set to zero × **Topic 161** · 1st with 0.0081 – organizer · 5th with 0.0048 – alarm · 7th with 0.0081 – calculator · 8th with 0.0081 – calendar · 21st with 0.0081 – note · 25th with 0.0081 – currency converter · 26th with 0.0081 – converter · 27th with 0.0081 – stopwatch · 31st ... → 6 values set to zero	× **Topic 6** · 1st with 0.0457 – organizer · 2nd with 0.0409 – alarm · 4th with 0.0382 – calendar · 5th with 0.0273 – calculator · 6th with 0.0246 – note · 10th with 0.0188 – task · 15th with 0.0153 – converter · 16th with 0.0146 – memo · 19th with 0.0136 – world clock · 22nd with 0.0104 – stopwatch · 42nd ... → 2 values set to zero ✓ **Topic 31** · 3rd with 0.0482 – alarm · 6th with 0.0349 – calendar · 11th with 0.0177 – note · 23rd with 0.0103 – world clock · 25th with 0.0083 – task · 32nd ... → 10 values set to zero × **Topic 103** · 2nd with 0.0344 – alarm · 3rd with 0.0297 – note · 22nd with 0.0126 – converter · 26th with 0.0109 – calendar · 27th with 0.0108 – calculator · 30th with 0.0101 – alarm · 56th ... → 8 values set to zero × **Topic 150** · 2nd with 0.0749 – calendar · 3rd with 0.0414 – note · 8th with 0.0285 – task · 16th with 0.0154 – organizer · 23rd with 0.0095 – memo · 43rd ... →11 values set to zero

Table 11-5. Best scored topics of LDA considering the *organizer* product feature on sections.

11.2.3. Paragraphs

	Without	With
K=50	✓ **Topic 4** · 1st with 0.0481 – calendar · 2nd with 0.0422 – alarm · 4th with 0.0355 – note · 8th with 0.0246 – task · 10th with 0.0221 – calculator · 11th with 0.0183 – organizer · 18th with 0.0124 – world clock · 20th with 0.0110 – converter · 23rd with 0.0083 – memo · 34th ... → 2 values set to zero × **Topic 36** · 2nd with 0.0639 – task · 29th with 0.0060 – calendar · 123rd ... → 10 values set to zero	✓ **Topic 3** · 1st with 0.0471 – calendar · 2nd with 0.0429 – alarm · 4th with 0.0359 – note · 6th with 0.0291 – task · 10th with 0.0227 – calculator · 11th with 0.0178 – organizer · 18th with 0.0123 – world clock · 20th with 0.0115 – converter · 24th with 0.0076 – memo · 32nd ... → 2 values set to zero .
K=100	✓ **Topic 3** · 2nd with 0.0488 – calendar · 4th with 0.0366 – task · 11th with 0.0155 – note · 17th with 0.0104 – alarm · 50th ... → 8 values set to zero × **Topic 58** · 1st with 0.1030 – alarm · 4th with 0.0538 – calculator · 5th with 0.0525 – note · 7th with 0.0427 – organizer · 8th with 0.0380 – calendar · 9th with 0.0370 – world clock · 11th with 0.0326 – converter · 15th with 0.0170 – task · 16th with 0.0163 – stopwatch · 19th with 0.0139 – memo · 29th with 0.0077 – countdown timer · X ... → ALL values set to zero	✓ **Topic 3** · 3rd with 0.0466 – task · 7th with 0.0282 –calendar · 19th with 0.0095 – note · 20th with 0.0090 – alarm · X ... → ALL values set to zero × **Topic 4** · 2nd with 0.0506 – calendar · 6th with 0.0195 – note · 62nd ... → 11 values set to zero ✓ **Topic 65** · 1st with 0.1081 – alarm · 3rd with 0.0607 – note · 4th with 0.0537 – calculator · 5th with 0.0500 – calendar · 8th with 0.0422 – organizer · 9th with 0.0339 – world clock · 11th with 0.0308 – converter · 14th with 0.0265 – memo · 17th with 0.0202 – task · 18th with 0.0157 – stopwatch · 26th with 0.0087 – countdown · 32nd ... → 3 values set to zero × **Topic 85** · 2nd with 0.0909 – office · 36th with 0.0419 – pdf · 67th ... → 14 values set to zero
K=150	✓ **Topic 60** · 2nd with 0.0846 – calendar · 9th with 0.0208 – alarm · 12th with 0.0191 – stopwatch · 19th with 0.0157 – note · 34th... → 7 values set to zero × **Topic 117** · 3rd with 0.0906 – converter · 12th with 0.0210 – calculator · 17th with 0.0140 – currency converter · 19th with 0.0101 – to-do · X ... → ALL values set to zero	× **Topic 60** · 2nd with 0.0749 – alarm · 3rd with 0.0665 – calculator · 5th with 0.0556 – world clock · 6th with 0.0436 – organizer · 7th with 0.0418 – calendar · 10th with 0.0333 – converter · 11th with 0.0326 – note · 13th with 0.0306 – stopwatch · 14th with 0.0257 – task · 15th with 0.0219 – memo · 18th with 0.0161 – countdown · X ... → ALL values set to zero

	× **Topic 137** · 3rd with 0.1214 – office · 5th with 0.0607 – pdf · 46th ... → 13 values set to zero	✓ **Topic 123** · 2nd with 0.1634 – calendar · 3rd with 0.0501 – alarm · 4th with 0.0411 – organizer · 7th with 0.0324 – note · 8th with 0.0250 – to-do · 14th with 0.0170 – currency · 21st with 0.0122 – task · 24th with 0.0117 – calculator · 60th ... → 8 values set to zero
K=200	× **Topic 137** · 1st with 0.0986 – alarm · 2nd with 0.0829 – calculator · 5th with 0.0609 – world clock · 6th with 0.0588 – calendar · 8th with 0.0456 – converter · 9th with 0.0455 – organizer · 11th with 0.0374 – stopwatch · 19th with 0.0191 – note · 21st with 0.0142 – memo · 25th with 0.0091 – currency converter · 32nd ... → 5 values set to zero	× **Topic 47** · 1st with 0.0598 – alarm · 2nd with 0.0494 – note · 3rd with 0.0453 – calculator · 4th with 0.0423 – calendar · 5th with 0.0402 – organizer · 6th with 0.0340 – memo · 9th with 0.0265 – stopwatch · 12th with 0.0251 – task · 21st with 0.0159 – countdown · X ... → ALL values set to zero × **Topic 48** · 3rd with 0.0692 – world clock · 4th with 0.0493 – calculator · 5th with 0.0475 – converter · 8th with 0.0354 – calendar · 9th with 0.0346 – alarm · 12th with 0.0282 – organizer · 14th with 0.0237 – stopwatch · 15th with 0.0229 – note · 17th with 0.0196 – task · 19th with 0.0172 – memo · 28th with 0.0082 – currency · 30th with 0.0075 – countdown · X ... → ALL values set to zero ✓ **Topic 99** · 4th with 0.0626 – calendar · 8th with 0.0401 – task · 9th with 0.0370 – note · 19th with 0.0191 – alarm · 27th with 0.0040 – office · X ... → ALL values set to zero × **Topic 140** · 1st with 0.0829 – currecncy · 2nd with 0.0660 – calendar · 4th with 0.0623 – converter · 5th with 0.0473 – alarm · 10th with 0.0342 – calculator · 11th with 0.0325 – organizer · 29th with 0.0103 – note · 51st ... → 6 values set to zero

Table 11-6. Best scored topics of LDA considering the *organizer* product feature on paragraphs.

11.3. *Multimedia*

11.3.1. Documents

	Without	With
K=50	× **Topic 13** · 17th with 0.0073 – mp3 · 18th with 0.0069 – radio · 28th with 0.0061 – video · 53rd ... → 26 values set to zero × **Topic 21** · 7th with 0.0145 – video · 19th with 0.0073 – divx · 25th with 0.0063 – multimedia · 60th ... → 21 values set to zero	× **Topic 23** · 5th with 0.0111 – video · 12th with 0.0068 – multimedia · 66th ... → 25 values set to zero × **Topic 32** · 8th with 0.0096 – video · 16th with 0.0076 – radio · 19th with 0.0062 – fm · 38th ... → 31 values set to zero
K=100	× **Topic 27** · 5th with 0.0142 – video · 24th with 0.0054 – multimedia · 60th ... → 22 values set to zero × **Topic 60** · 2nd with 0.0121 – video · 14th with 0.0071 – multimedia · 43rd ... → 15 values set to zero	× **Topic 25** · 7th with 0.0149 – audio · 15th with 0.0092 – video · 20th with 0.0079 – fm · 26th with 0.0074 – radio · 53rd ... → 24 values set to zero × **Topic 63** · 7th with 0.0195 – video · 40th ... → 22 values set to zero
K=150	× **Topic 54** · 3rd with 0.0214 – video · 47th ... → 22 values set to zero × **Topic 95** · 3rd with 0.0123 – video · 8th with 0.0110 – multimedia · 25th with 0.0060 – divx · 159th ... → 21 values set to zero	× **Topic 27** · 10th with 0.0107 – radio · 16th with 0.0085 – fm · 26th with 0.0064 – music player · 29th with 0.0055 – video · 81st ... → 33 values set to zero × **Topic 97** · 2nd with 0.0125 – video · 7th with 0.0112 – multimedia · 30th with 0.0055 – divx · 181st ... → 27 values set to zero ✓ **Topic 147** · 11th with 0.0146 – mp3 · 18th with 0.0098 – music player · 34th ... → 38 values set to zero
K=200	× **Topic 134** · 2nd with 0.0363 – video · X ... → ALL values set to zero ✓ **Topic 146** · 9th with 0.0195 – battery · 27th with 0.0109 – talk time · 38th ... → 21 values set to zero ✓ **Topic 161** · 8th with 0.0213 – music player · 15th with 0.0090 – stereo · 36th ... → 30 values set to zero	× **Topic 19** · 5th with 0.0117 – video · 23rd with 0.0060 – music player · 28th with 0.0054 – radio · 57th ... → 31 values set to zero × **Topic 117** · 3rd with 0.0125 – video · 7th with 0.0111 – multimedia · 24th with 0.0062 – divx · 176th ... → 20 values set to zero

Table 11-7. Best scored topics of LDA considering the *multimedia* product feature on documents.

11.3.2. Sections

	Without	With
K=50	✓ **Topic 10** · 3rd with 0.0334 – music player · 4th with 0.0273 – video · 6th with 0.0256 – radio · 7th with 0.0210 – fm · 17th with 0.0112 – multimedia · 23rd with 0.0073 – video player · 33rd ... → 9 values set to zero × **Topic 11** · 1st with 0.0467 – video · 7th with 0.0181 – multimedia · 27th with 0.0064 – audio · 31st ... → 7 values set to zero	✓ **Topic 12** · 3rd with 0.0379 – battery · 4th with 0.0225 – talk time · 9th with 0.0115 – battery · 10th with 0.0147 – battery · 13th with 0.0134 – battery · 28th with 0.0065 – battery · 31st ... →1 values set to zero × **Topic 20** · 1st with 0.0351 – video · 3rd with 0.0216 – divx · 4th with 0.0183 – xvid · 10th with 0.0127 – multimedia · 37th ... → 29 values set to zero
K=100	✓ **Topic 10** · 3rd with 0.0401 – video · 7th with 0.0189 – music player · 9th with 0.0153 – multimedia · 18th with 0.0111 – radio · 22nd with 0.0087 – fm · 29th with 0.0064 – video player · 36th ... → 4 values set to zero × **Topic 12** · 1st with 0.0524 – video · 7th with 0.0188 – divx · 8th with 0.0180 – multimedia · 10th with 0.0160 – audio · 12th with 0.0153 – fm · 13th with 0.0152 – radio · 15th with 0.0137 – xvid · 18th with 0.0110 – video player · 20th with 0.0102 – music player · 36th ... → 19 values set to zero × **Topic 19** · 6th with 0.0269 – video · 74th with 0.0079 – multimedia · 21st with 0.0064 – xvid · 25th with 0.0058 – divx · 55th ... → 28 values set to zero	× **Topic 11** · 3rd with 0.0402 – music player · 5th with 0.0243 – radio · 8th with 0.0197 – rm · 14th with 0.0142 – video · 22nd with 0.0102 – multimedia · 35th ... → 28 values set to zero ✓ **Topic 12** · 1st with 0.0823 – video · 7th with 0.0228 – multimedia · 9th with 0.0206 – divx · 10th with 0.0190 – audio · 11th with 0.0151 – xvid · 12th with 0.0145 – video playback · 18th with 0.0103 – radio · 19th with 0.0101 – video player · 20th with 0.0098 – fm · 21st with 0.0087 – music player · 28th with 0.0067 – mpeg4 · 33rd ... → 15 values set to zero × **Topic 13** · 3rd with 0.0406 – video · 8th with 0.0157 – multimedia · 13th with 0.0131 – mp3 · 19th with 0.0096 – h263 · 22nd with 0.0087 – aac · 24th with 0.0085 – mpeg4 · 25th with 0.0085 – audio · 27th with 0.0082 – wma · 31st ... → 23 values set to zero

<table>
<tr>
<td>K=150</td>
<td>
× Topic 11
· 1st with 0.0523 – video
· 10th with 0.0130 – multimedia
· 13th with 0.0105 – mp3
· 14th with 0.0101 – h263
· 15th with 0.0097 – mpeg4
· 21st with 0.0079 – music player
· 23rd with 0.0072 – h264
· 24th with 0.0072 – aac
· 25th with 0.0071 – wma
· 26th with 0.0069 – wmv
· 31st ... → 8 values set to zero
× Topic 12
· 1st with 0.0601 – video
· 4th with 0.0263 – divx
· 6th with 0.0209 – xvid
· 7th with 0.0204 – multimedia
· 11th with 0.0149 – audio
· 13th with 0.0141 – video playback
· 39th ... → 23 values set to zero
✓ Topic 55
· 1st with 0.1736 – radio
· 2nd with 0.1453 – fm
· 6th with 0.0338 – rds
· 12th with 0.0181 – audio
· 16th with 0.0128 – music player
· 20th with 0.0106 – fm transmitter
· 22nd with 0.0090 – video player
· 33rd ... → 29 values set to zero
</td>
<td>
× Topic 12
· 1st with 0.0686 – video
· 4th with 0.0263 – divx
· 5th with 0.0262 – multimedia
· 9th with 0.0193 – xvid
· 11th with 0.0153 – video playback
· 18th with 0.0096 – audio
· 20th with 0.0086 – mpeg4
· 23rd with 0.0070 – h264
· 35th ... → 18 values set to zero
× Topic 51
· 1st with 0.1049 – radio
· 2nd with 0.0899 – fm
· 7th with 0.0234 – rds
· 9th with 0.0166 – video
· 16th with 0.0112 – audio
· 17th with 0.0107 – music player
· 28th with 0.0072 – fm transmitter
· 38th ... → 20 values set to zero
</td>
</tr>
<tr>
<td>K=200</td>
<td>
× Topic 10
· 2nd with 0.0427 – video
· 11th with 0.0151 – music player
· 13th with 0.0143 – multimedia
· 21st with 0.0097 – video player
· 25th with 0.0080 – h263
· 30th with 0.0067 – mpeg4
· 35th ... → 9 values set to zero
✓ Topic 11
· 1st with 0.0736 – video
· 5th with 0.0280 – multimedia
· 8th with 0.0221 – divx
· 10th with 0.0200 – xvid
· 12th with 0.0147 – video playback
· 13th with 0.0135 – audio
· 24th with 0.0078 – mpeg4
· 30th with 0.0066 – h264
· 34th ... → 15 values set to zero
</td>
<td>
✓ Topic 12
· 3rd with 0.0335 – video
· 6th with 0.0204 – radio
· 11th with 0.0161 – fm
· 13th with 0.0144 – music player
· 18th with 0.0119 – multimedia
· 23rd with 0.0090 – h263
· 26th with 0.0080 – video player
· 27th with 0.0080 – mpeg4
· 39th ... → 4 values set to zero
✓ Topic 14
· 1st with 0.0762 – video
· 5th with 0.0244 – multimedia
· 6th with 0.0238 – divx
· 8th with 0.0212 – xvid
· 10th with 0.0170 – audio
· 12th with 0.0160 – video playback
· 15th with 0.0113 – fm
· 17th with 0.0099 – radio
· 23rd with 0.0085 – video player
· 36th ... → 24 values set to zero
</td>
</tr>
</table>

Table 11-8. Best scored topics of LDA considering the *multimedia* product feature on sections.

11.3.3. Paragraphs

	Without	With
K=50	✓ **Topic 5** · 3rd with 0.0350 – music player · 8th with 0.0203 – video · 14th with 0.0130 – audio · 19th with 0.0105 – multimedia · 26th with 0.0060 – mp3 · 34th ... → 12 values set to zero × **Topic 8** · 1st with 0.1173 – video · 9th with 0.0150 – divx · 15th with 0.0115 – xvid · 23rd with 0.0097 – mpeg4 · 34th ... → 17 values set to zero	× **Topic 5** · 4th with 0.0291 – music player · 9th with 0.0161 – audio · 16th with 0.0123 – multimedia · 25th with 0.0067 – mp3 · 44th ... → 9 values set to zero × **Topic 32** · 2nd with 0.0663 – radio · 4th with 0.0511 – fm · 14th with 0.0168 – music player · 19th with 0.0128 – rds · 28th with 0.0077 – stereo · 31st ... → 36 values set to zero ✓ **Topic 35** · 1st with 0.3426 – video · 11th with 0.0207 – video player · 13th with 0.0186 – video playback · 14th with 0.0167 – mpeg4 · 19th with 0.0113 – h263 · 29th with 0.0065 – mp4 · 45th ... → 35 values set to zero
K=100	× **Topic 2** · 1st with 0.0661 – video · 8th with 0.0146 – divx · 12th with 0.0112 – xvid · 22nd with 0.0071 – mpeg4 · 29th with 0.0061 – h264 · 33rd ... → 20 values set to zero × **Topic 9** · 6th with 0.0344 – music player · 16th with 0.0118 – audio · 22nd with 0.0072 – multimedia · 6239th ... → 22 values set to zero ✓ **Topic 40** · 1st with 0.2234 – video · 6th with 0.0398 – video player · 10th with 0.0248 – multiemdia · 16th with 0.0142 – audio · 20th with 0.0088 – mpeg4 · 21st with 0.0087 – mp3 · 22nd with 0.0083 – wmv · 42nd ... → 10 values set to zero	✓ **Topic 22** · 1st with 0.1561 – video · 7th with 0.0241 – divx · 11th with 0.0187 – xvid · 12th with 0.0165 – video player · 15th with 0.0155 – mpeg4 · 17th with 0.0135 – video playback · 20th with 0.0108 – multimedia · 21st with 0.0103 – h263 · 22nd with 0.0102 – h264 · 35th ... → 6 values set to zero × **Topic 42** · 2nd with 0.1313 – radio · 3rd with 0.0987 – fm · 8th with 0.0264 – rds · 10th with 0.0225 – stereo · 37th ... → 37 values set to zero

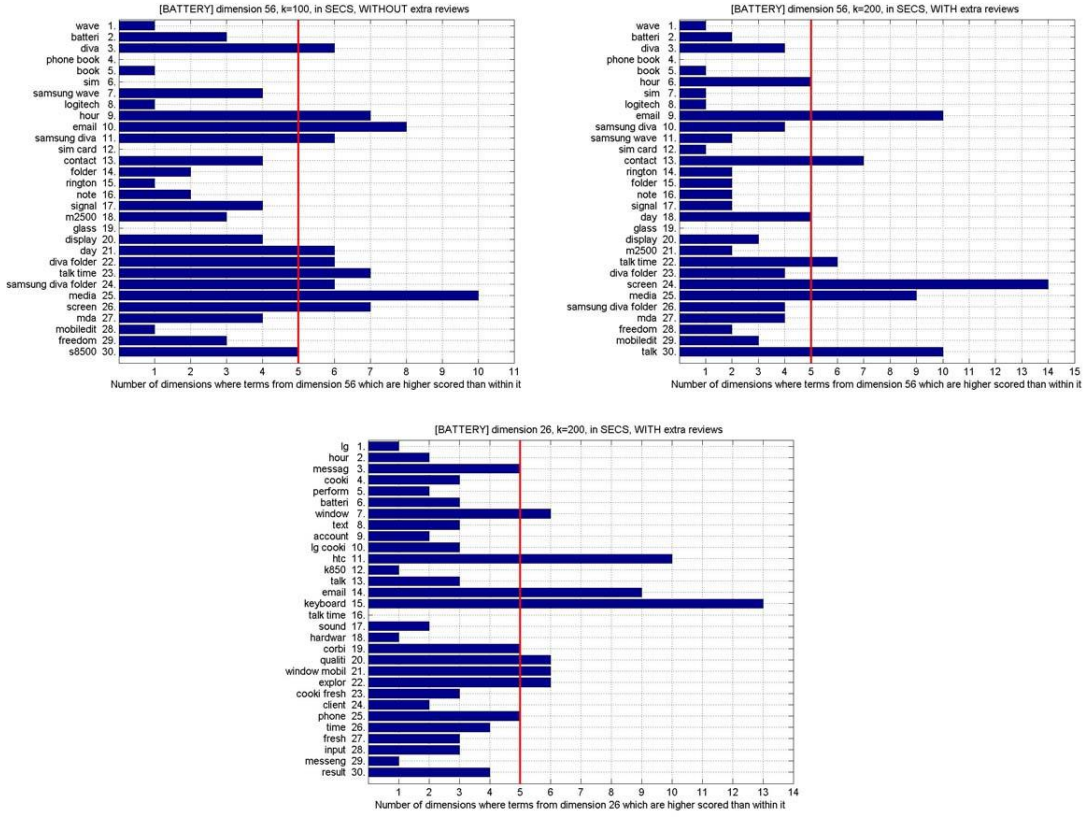
K=150	× **Topic 2** · 1st with 0.0513 – video · 3rd with 0.0197 – divx · 6th with 0.0157 – xvid · 18th with 0.0072 – h264 · 22nd with 0.0068 – video playback · 47th ... → 22 values set to zero × **Topic 5** · 10th with 0.0176 – music player · 17th with 0.0105 – audio · 25th with 0.0064 – mp3 · 32nd ... → 16 values set to zero ✓ **Topic 56** · 1st with 0.2338 – video · 6th with 0.0384 – mpeg4 · 7th with 0.0355 – video player · 9th with 0.0282 – h263 · 23rd with 0.0114 – wmv · 31st ... → 26 values set to zero	✓ **Topic 20** · 1st with 0.1671 – video · 4th with 0.0430 – divx · 6th with 0.0353 – battery · 9th with 0.0243 – xvid · 11th with 0.0232 – mpeg4 · 14th with 0.0218 – video playback · 16th with 0.0180 – video player · 17th with 0.0175 – h263 · 22nd with 0.0130 – h264 · 24th with 0.0113 – multimedia · 32nd ... → 7 values set to zero × **Topic 57** · 4th with 0.0636 – video · 8th with 0.0474 – mp3 · 12th with 0.0258 – aac · 13th with 0.0191 – wav · 15th with 0.0183 – multimedia · 20th with 0.0078 – video player · 24th with 0.0061 – amr · X ... → ALL values set to zero
K=200	× **Topic 3** · 1st with 0.0901 – video · 3rd with 0.0363 – divx · 4th with 0.0323 – xvid · 15th with 0.0125 – video playback · 16th with 0.0116 – h264 · 25th with 0.0082 – mpeg4 · 32nd ... → 23 values set to zero ✓ **Topic 46** · 2nd with 0.0632 – video · 10th with 0.0219 – mp3 · 13th with 0.0206 – aac · 15th with 0.0185 – wma · 17th with 0.0178 – multimedia · 18th with 0.0174 – wav · 23rd with 0.0129 – audio · 29th with 0.0106 – video player · 30th with 0.0103 – aac+ · 32nd ... → 17 values set to zero	× **Topic 2** · 1st with 0.0586 – video · 4th with 0.0189 – divx · 6th with 0.0182 – xvid · 24th with 0.0073 – video playback · 26th with 0.0066 – h264 · 33rd ... → 27 values set to zero × **Topic 20** · 7th with 0.0257 – wma · 8th with 0.0183 – audio · 11th with 0.0168 – video · 13th with 0.0155 – aac+ · 15th with 0.0139 – aac · 18th with 0.0125 – album art cover · 20th with 0.0114 – mp3 · 23rd with 0.0094 – wmv · 26th with 0.0088 – eaac+ · 27th with 0.0082 – mp4 · 29th with 0.0077 – multimedia · 33rd ... → 2 values set to zero ✓ **Topic 58** · 2nd with 0.0761 – battery · 10th with 0.0334 – talk time · 12th with 0.0257 – battery · 16th with 0.0188 – battery · 17th with 0.0184 – battery · 21st with 0.0156 – battery · 23rd with 0.0125 – battery · 26th with 0.0112 – battery · 31st ... → 31 values set to zero

Table 11-9. Best scored topics of LDA considering the *multimedia* product feature on paragraphs.

12. Annex D: Discarding method applied to LSI dimensions

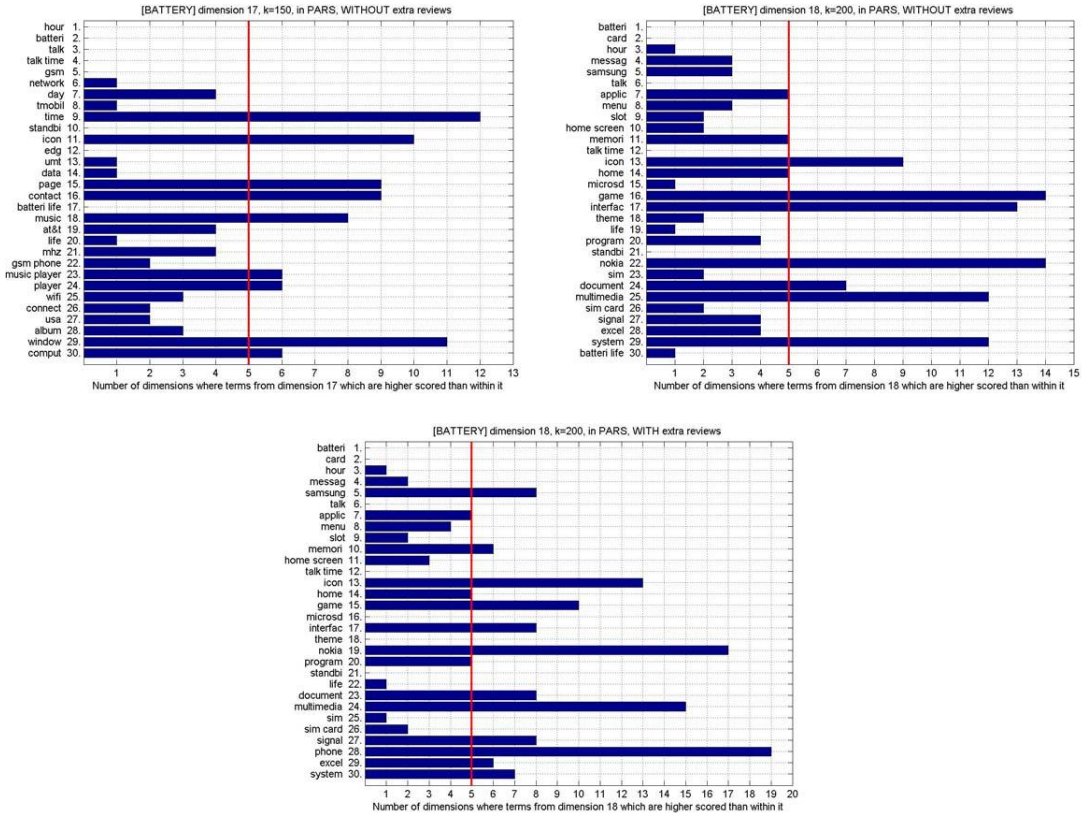
12.1. *Battery*

12.1.1. Sections



Graphic 12-1. Discarding method applied to top 30 terms from *battery* dimensions on sections.

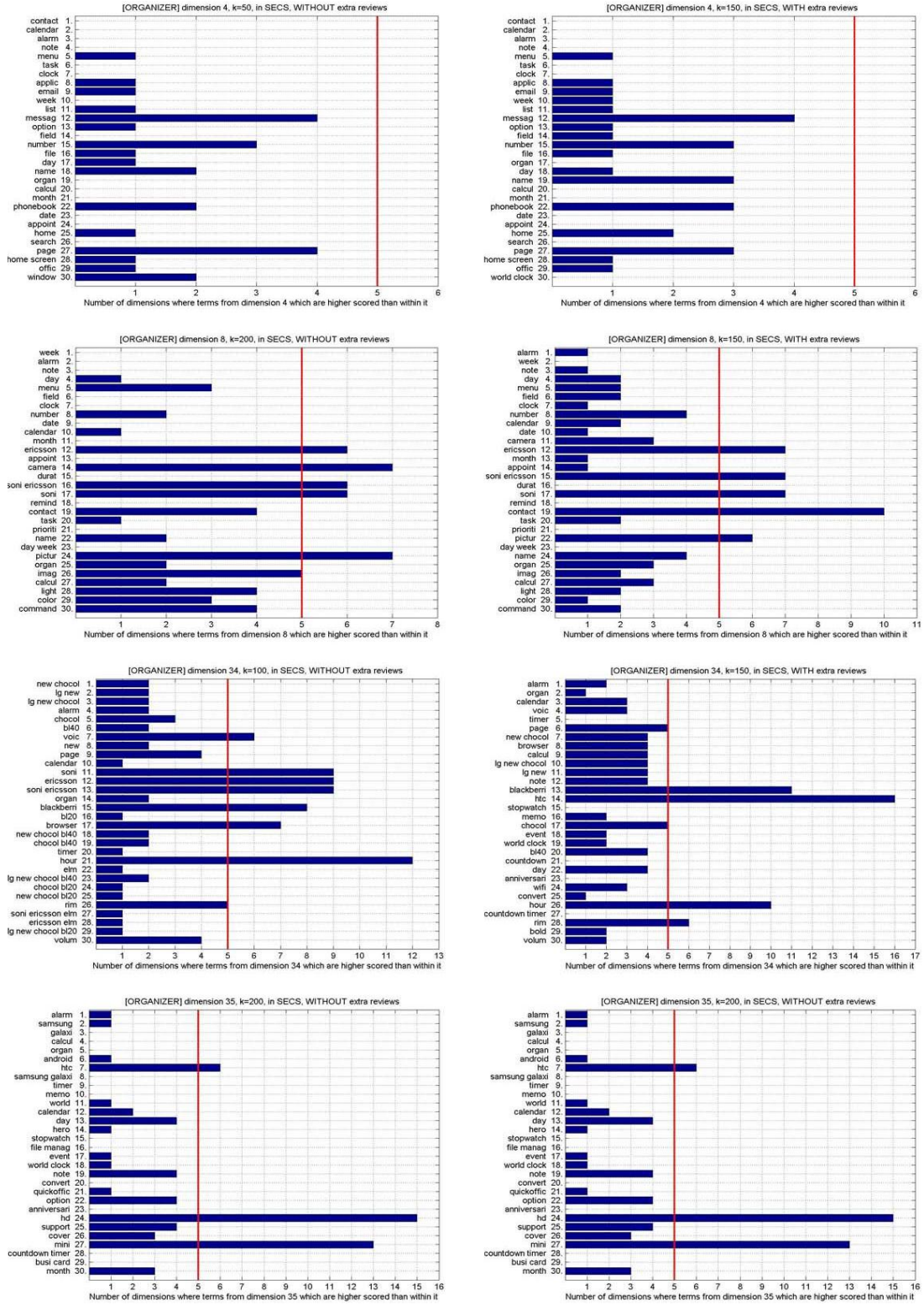
12.1.2. Paragraphs



Graphic 12-2. Discarding method applied to top 30 terms from *battery* dimensions on paragraphs.

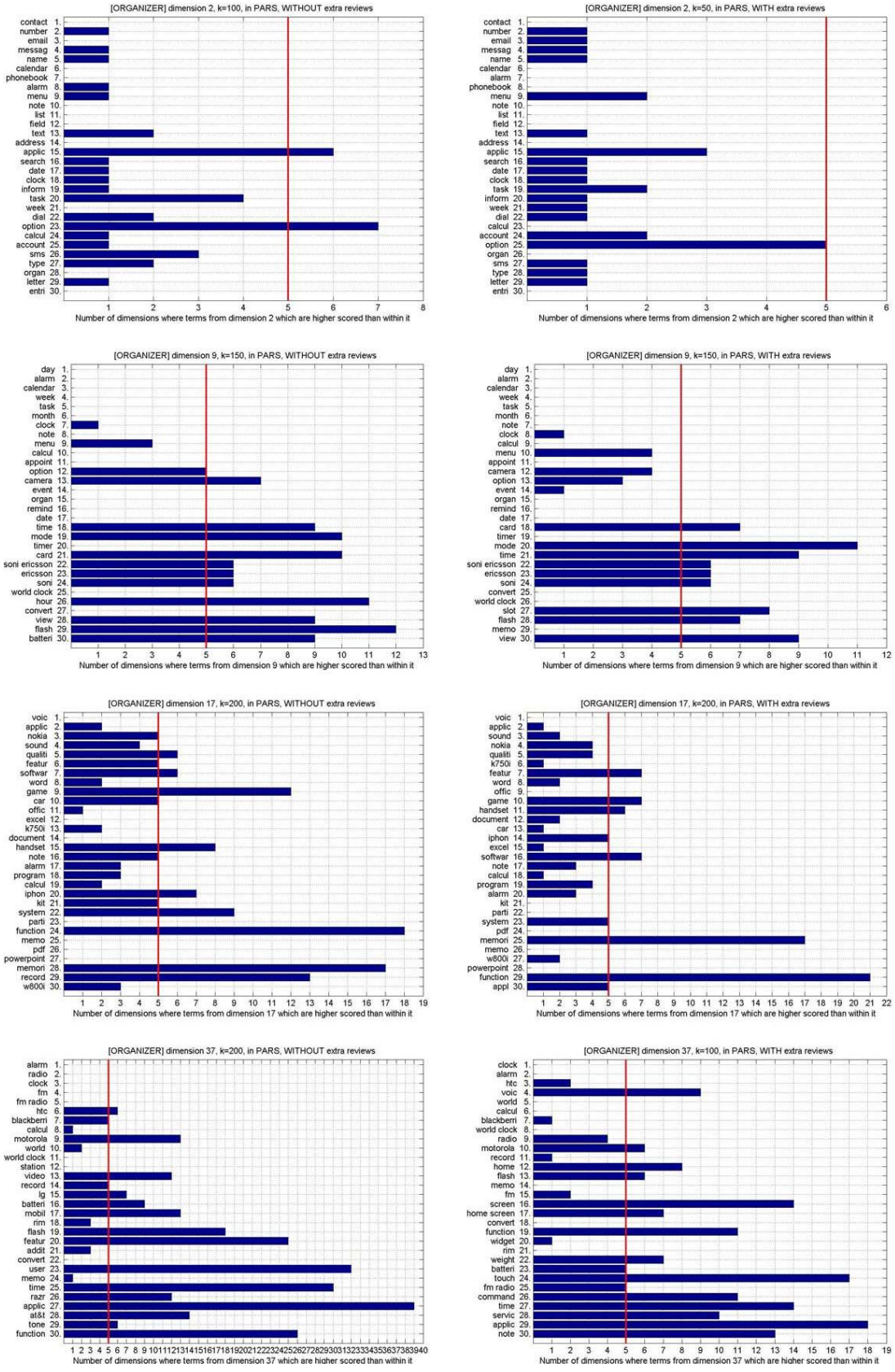
12.2. Organizer

12.2.1. Sections



Graphic 12-3. Discarding method applied to top 30 terms from *organizer* dimensions on sections.

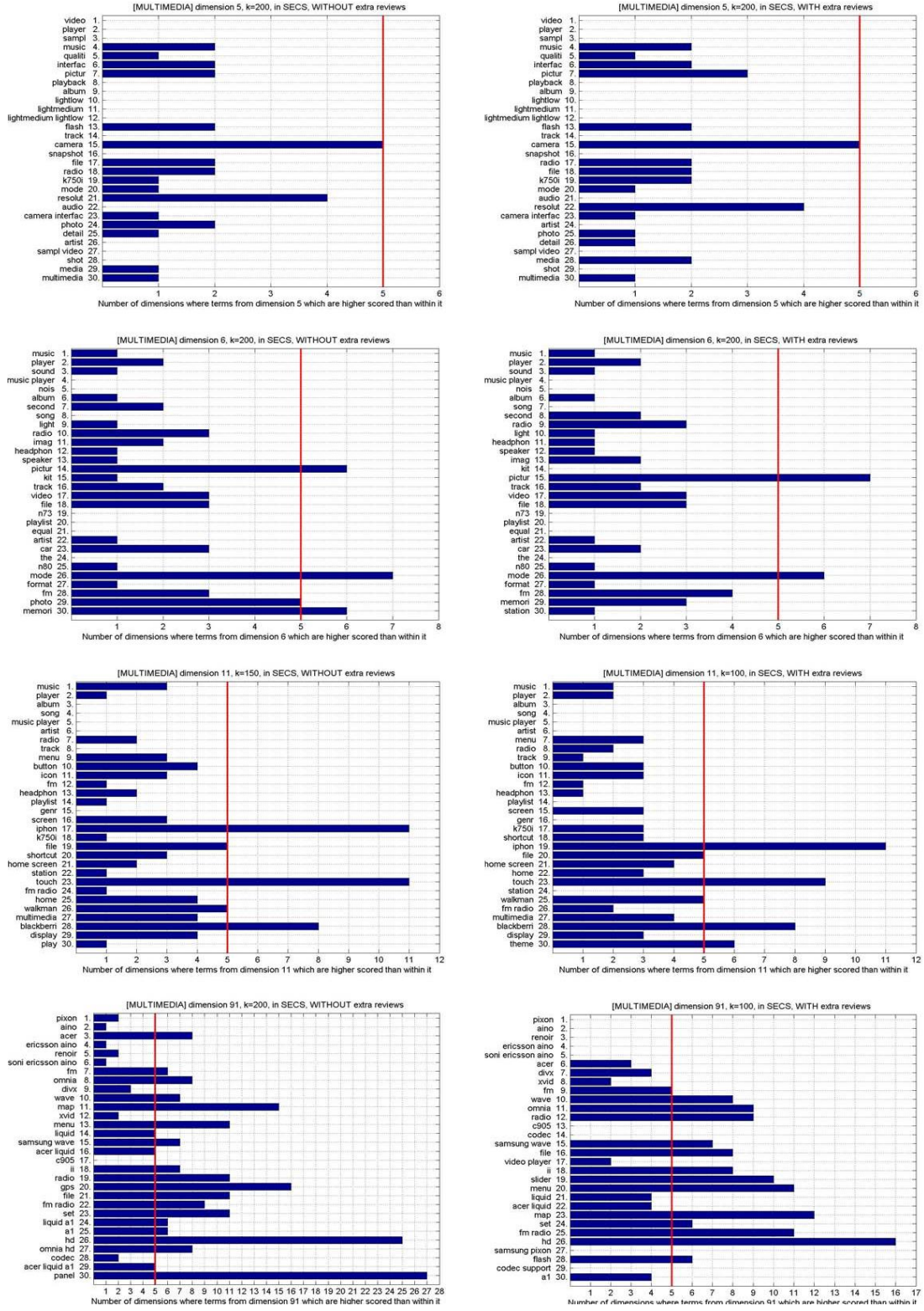
12.2.2. Paragraphs



Graphic 12-4. Discarding method applied to top 30 terms from *organizer* dimensions on paragraphs.

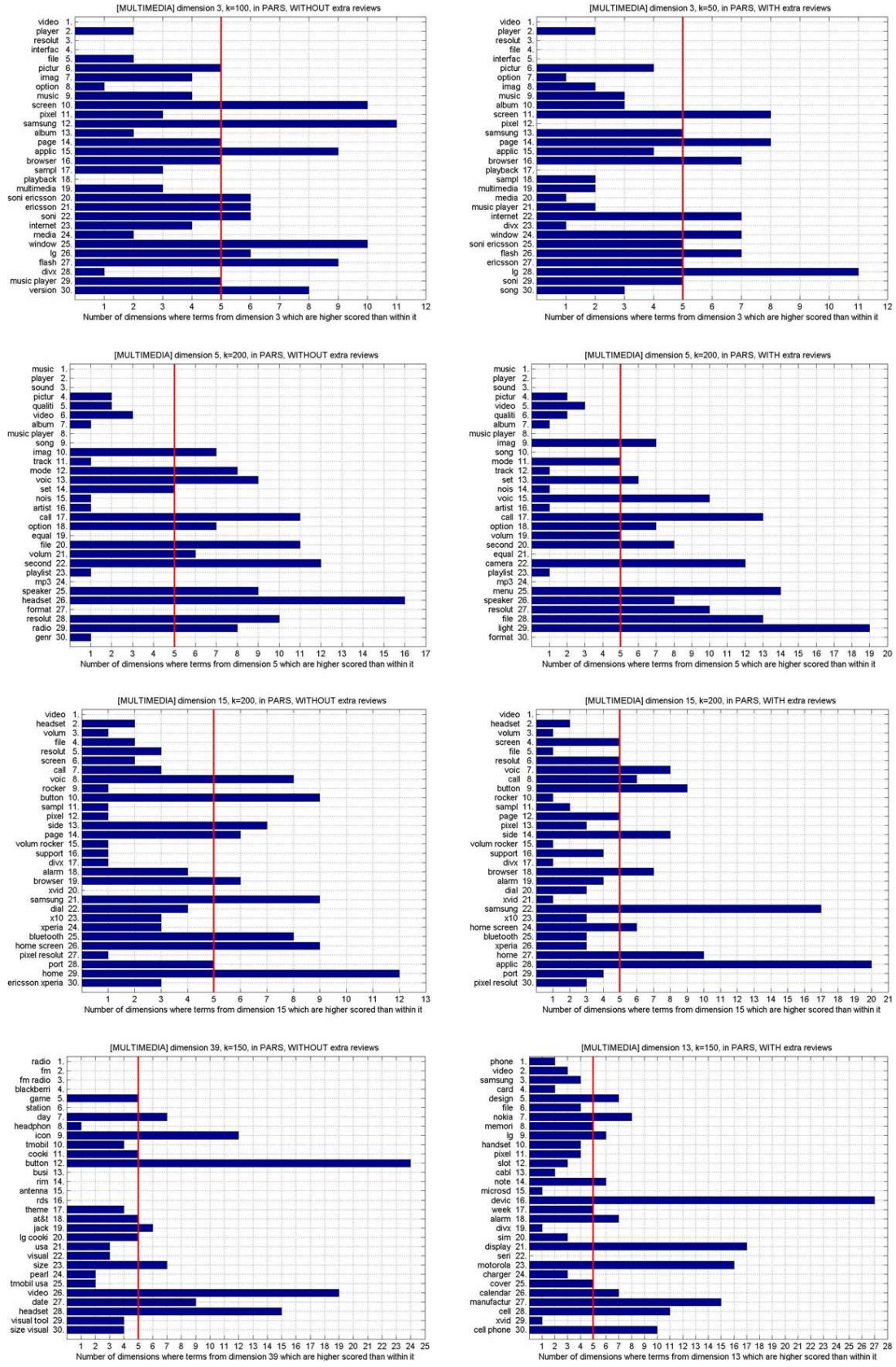
12.3. Multimedia

12.3.1. Sections



Graphic 12-5. Discarding method applied to top 30 terms from *multimedia* dimensions on sections.

12.3.2. Paragraphs

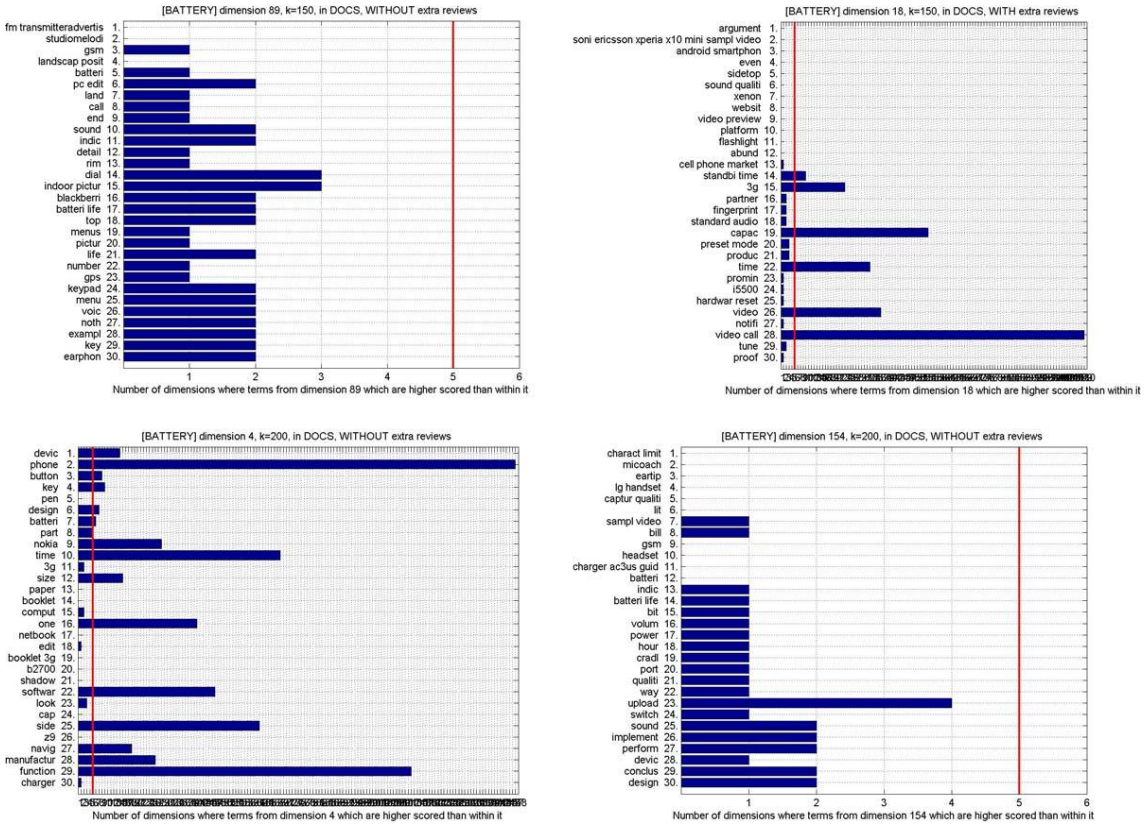


Graphic 12-6. Discarding method applied to top 30 terms from *multimedia* dimensions on paragraphs.

13. Annex E: Discarding method applied to PLSI topics

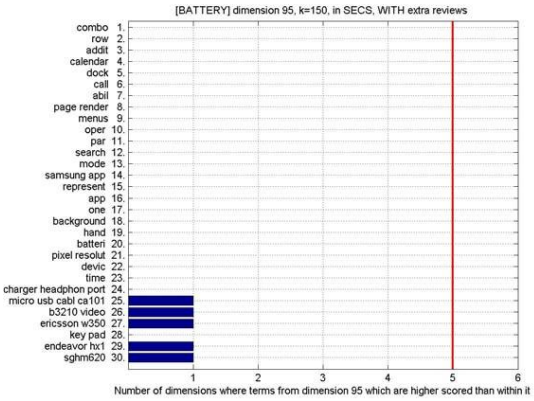
13.1. *Battery*

13.1.1. Documents



Graphic 13-1. Discarding method applied to top 30 terms from *battery* dimensions on documents.

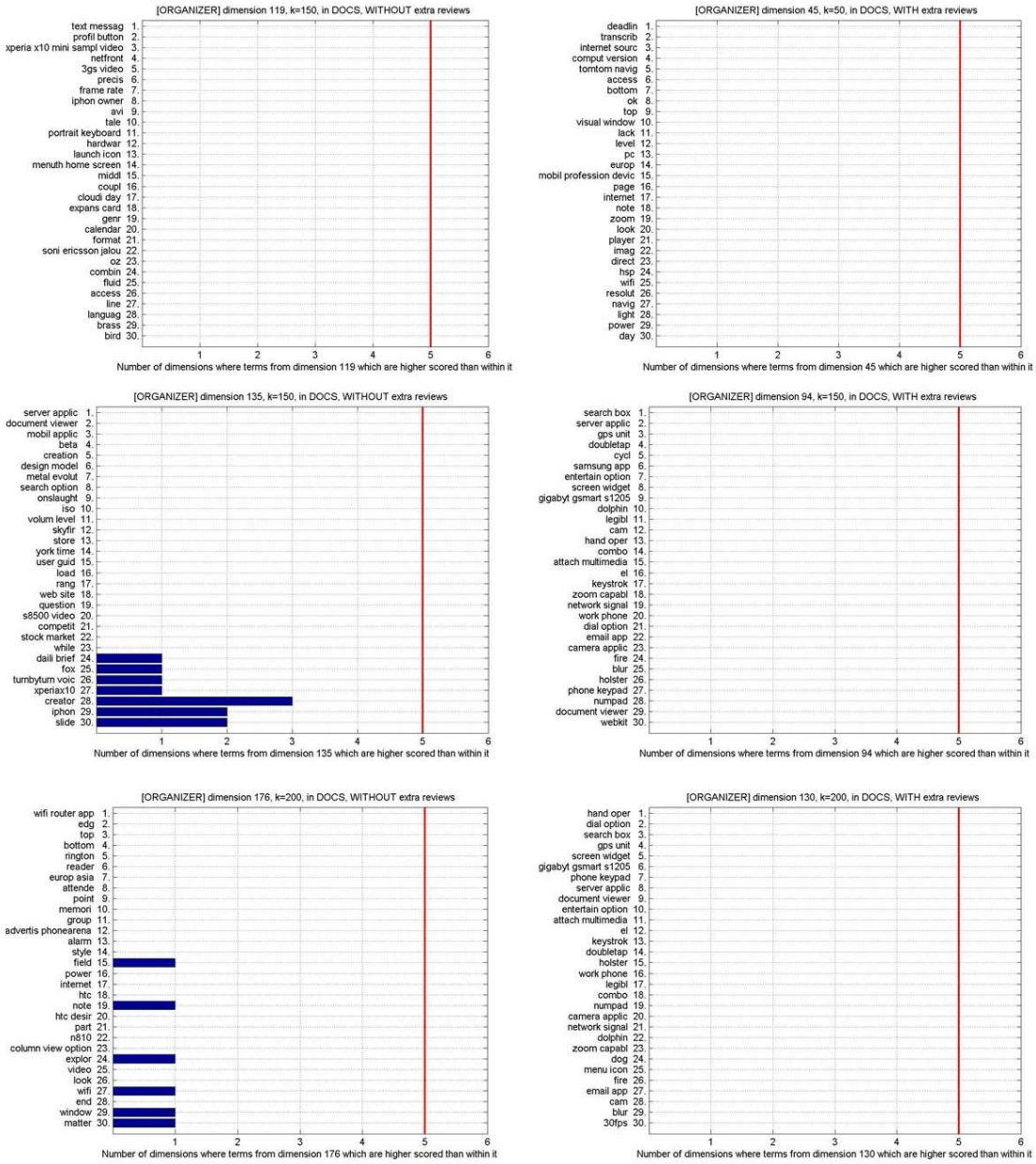
13.1.2. Sections



Graphic 13-2. Discarding method applied to top 30 terms from *battery* dimensions on sections.

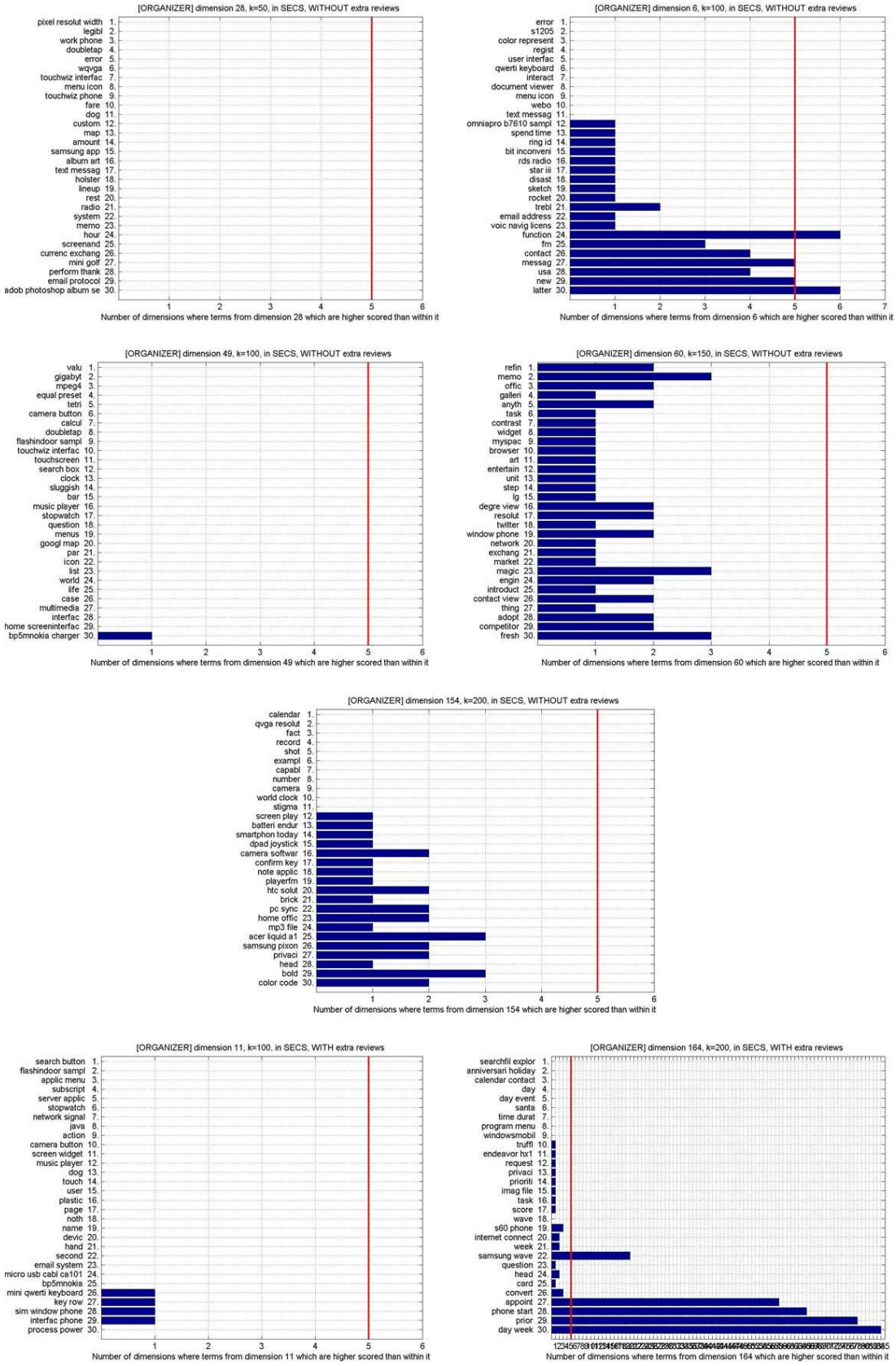
13.2. Organizer

13.2.1. Documents



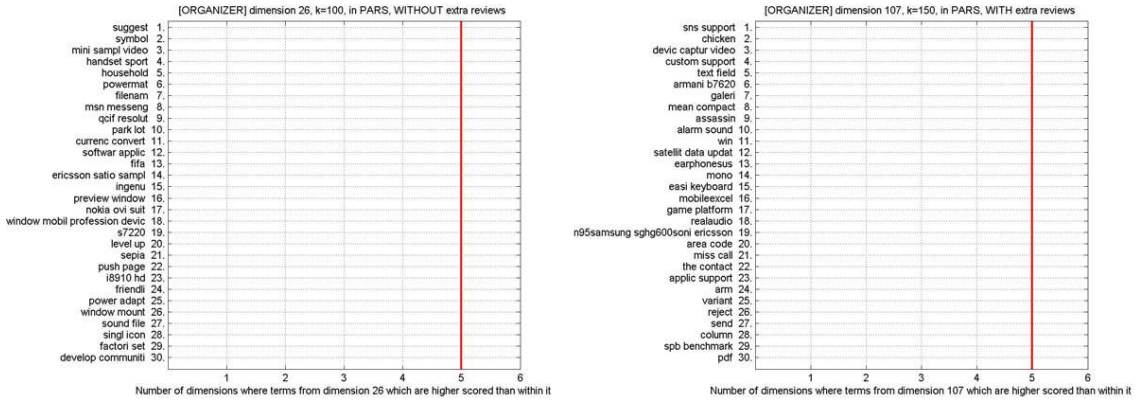
Graphic 13-3. Discarding method applied to top 30 terms from *organizer* dimensions on documents.

13.2.2. Sections



Graphic 13-4. Discarding method applied to top 30 terms from *organizer* dimensions on sections.

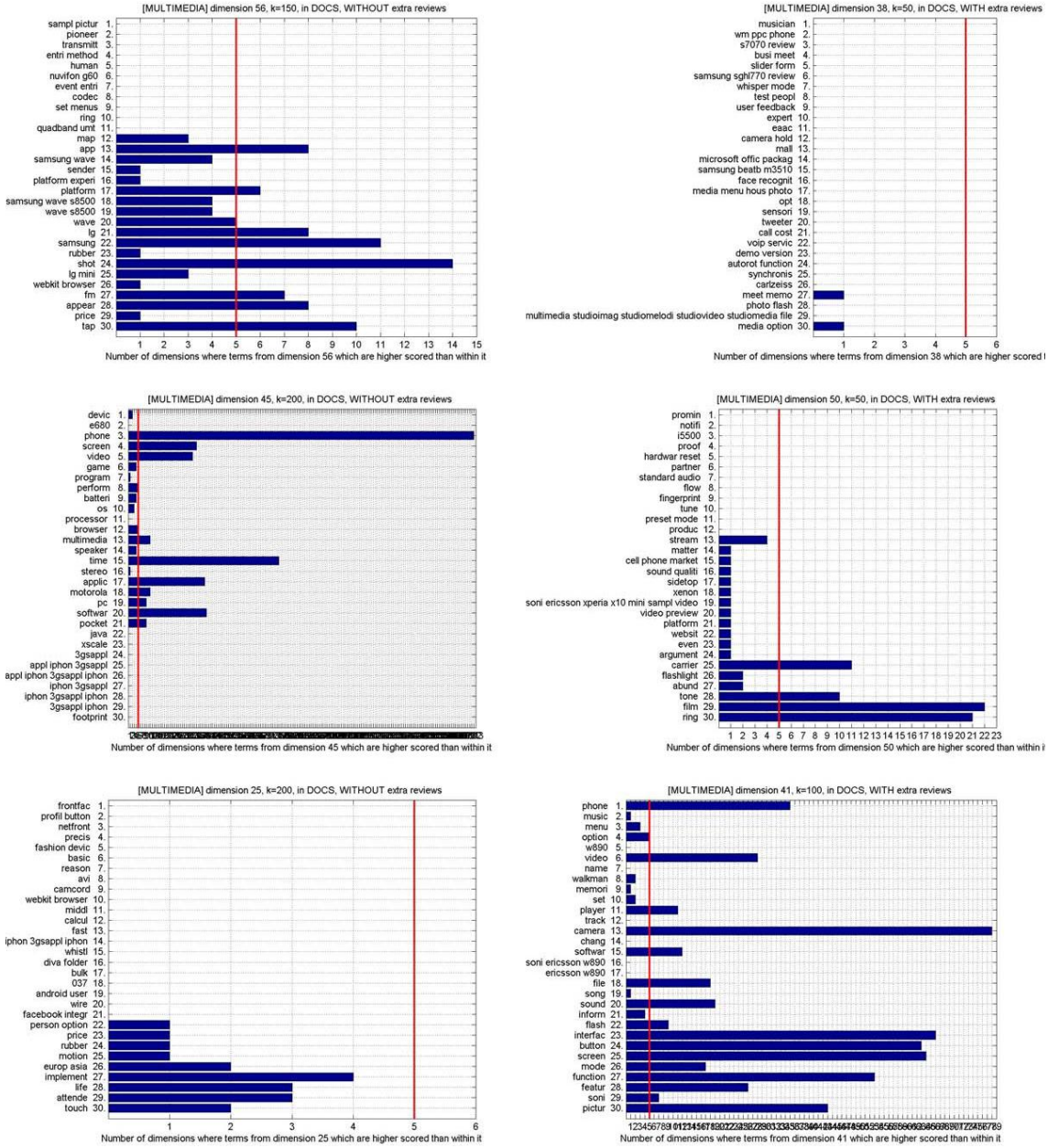
13.2.3. Paragraphs



Graphic 13-5. Discarding method applied to top 30 terms from *organizer* dimensions on paragraphs.

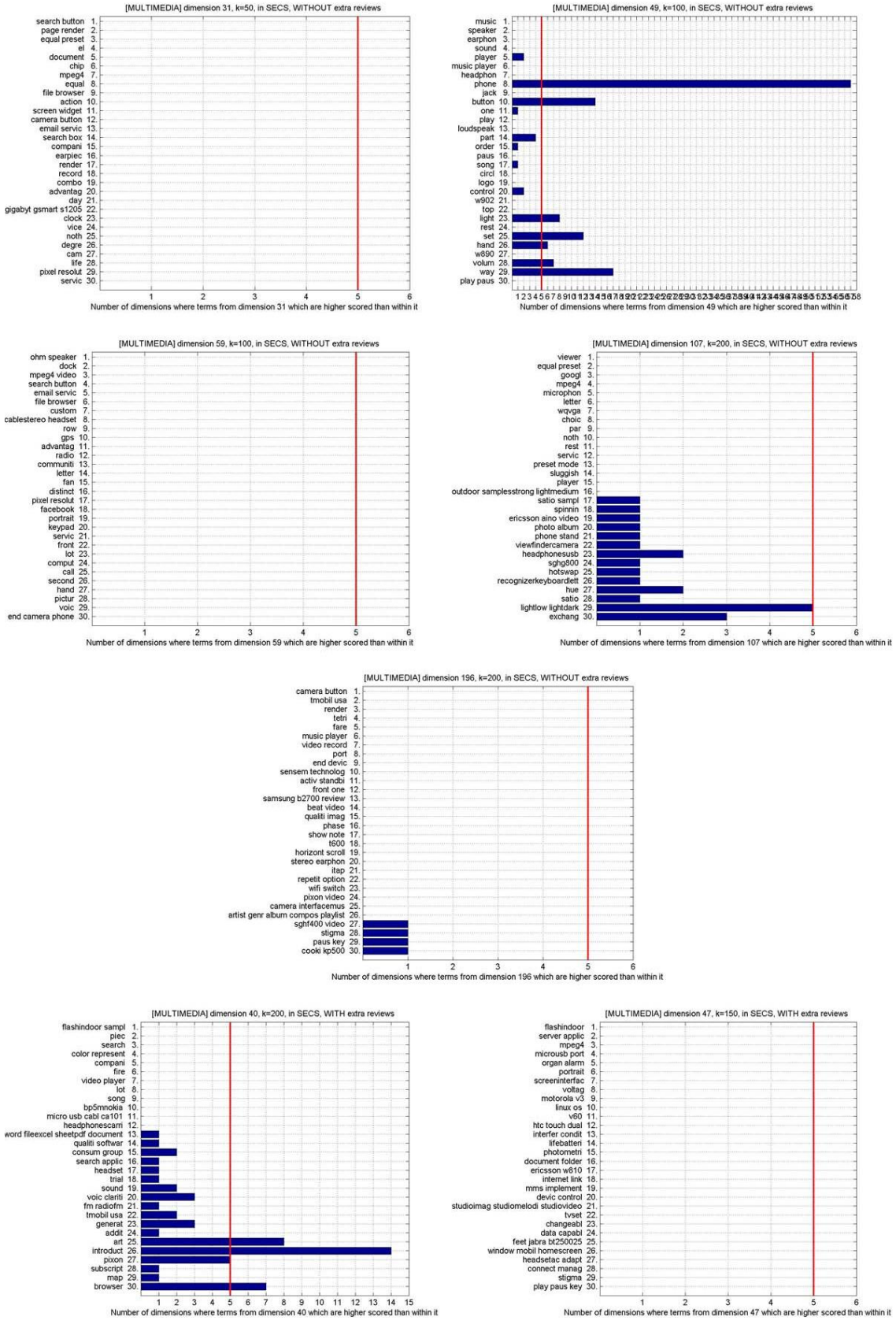
13.3. *Multimedia*

13.3.1. Documents



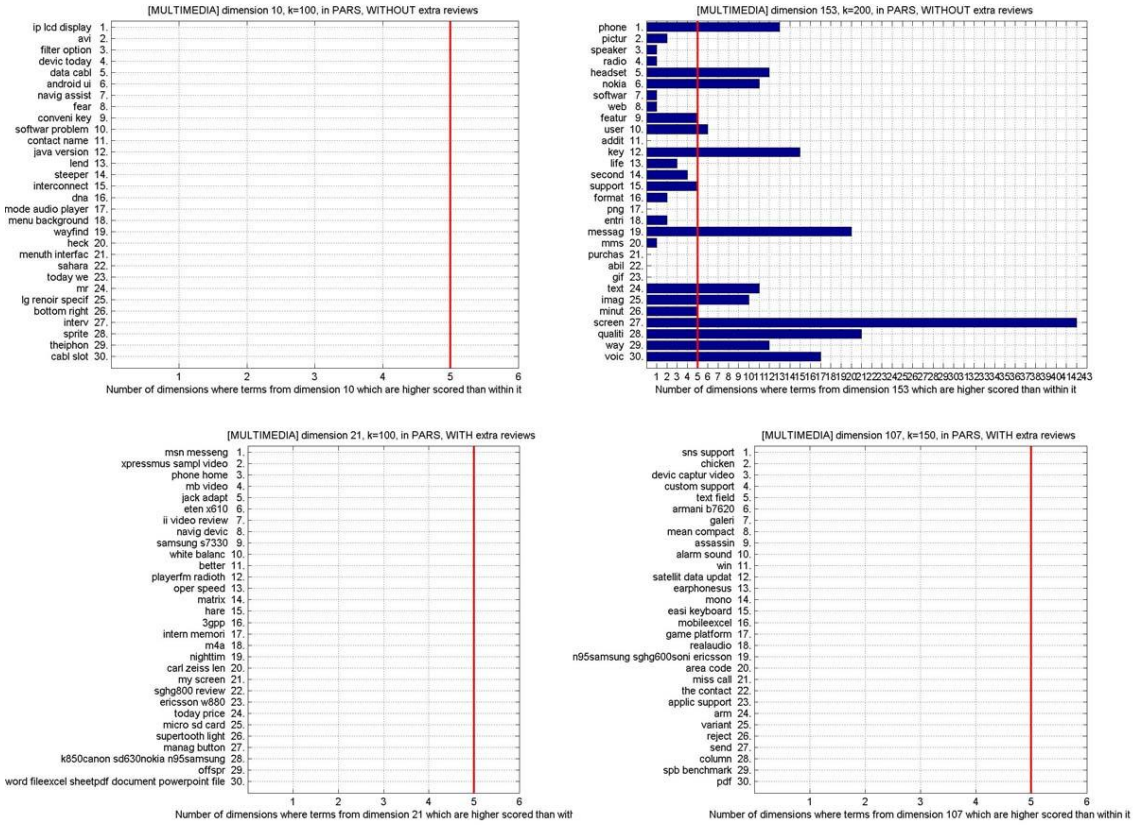
Graphic 13-6. Discarding method applied to top 30 terms from *multimedia* dimensions on documents.

13.3.2. Sections



Graphic 13-7. Discarding method applied to top 30 terms from *multimedia* dimensions on sections.

13.3.3. Paragraphs

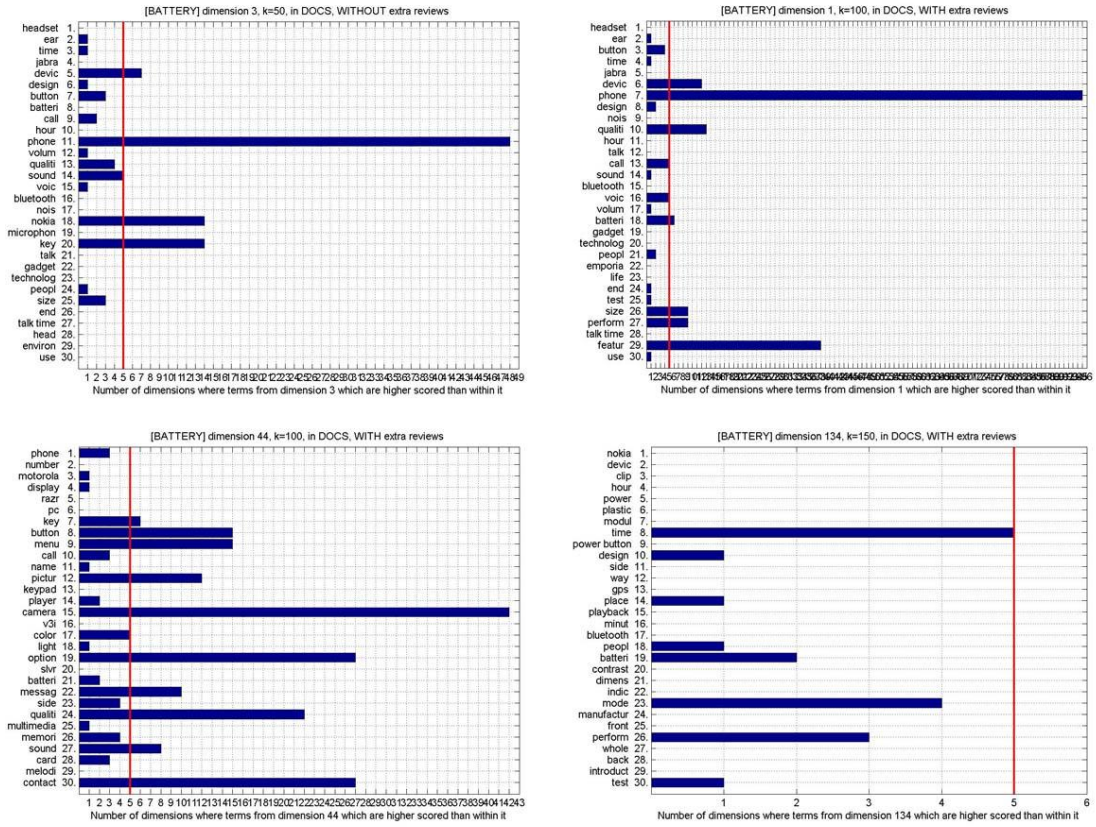


Graphic 13-8. Discarding method applied to top 30 terms from *multimedia* dimensions on paragraphs.

14. Annex F: Discarding method applied to LDA topics

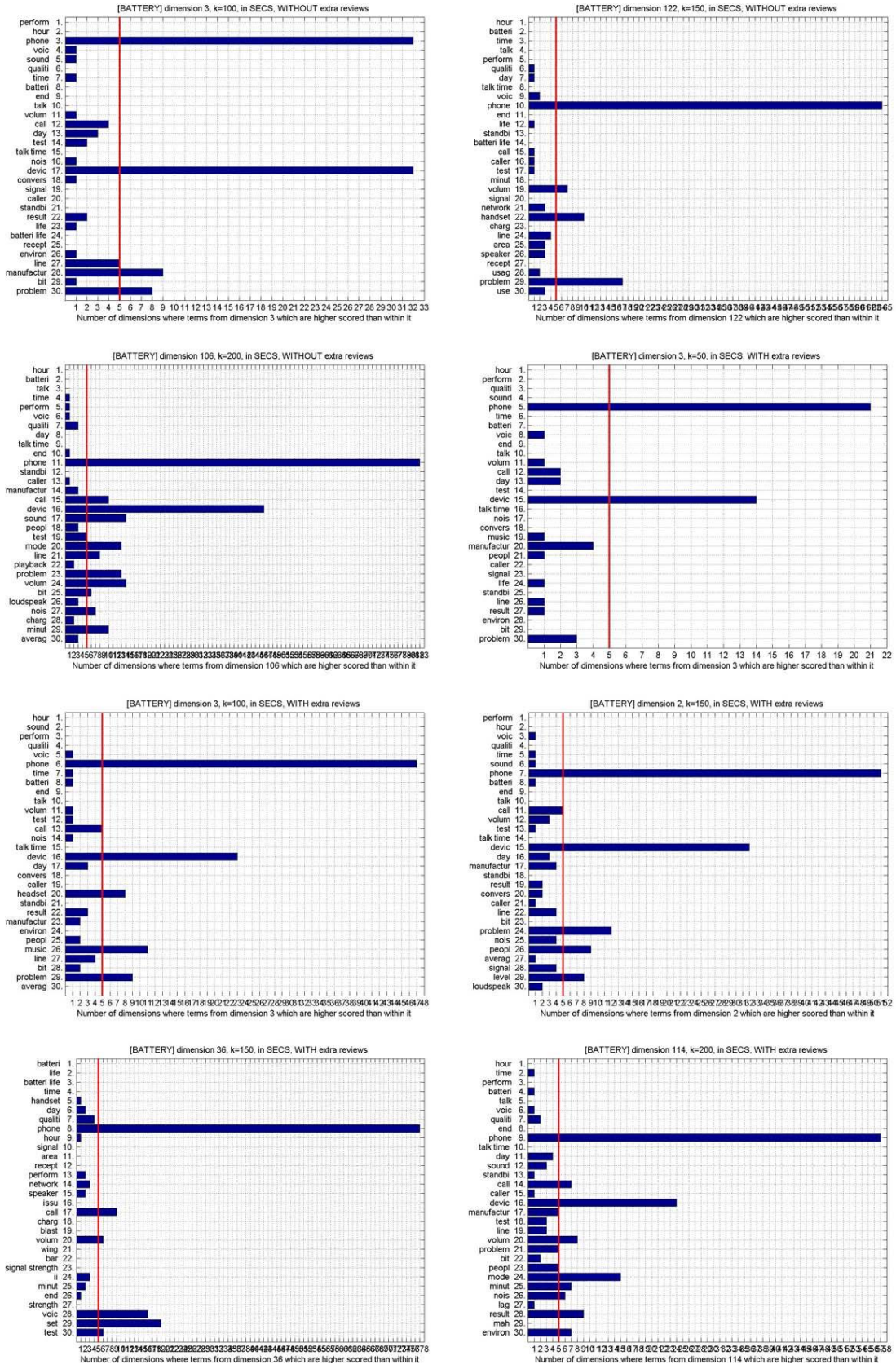
14.1. *Battery*

14.1.1. Documents



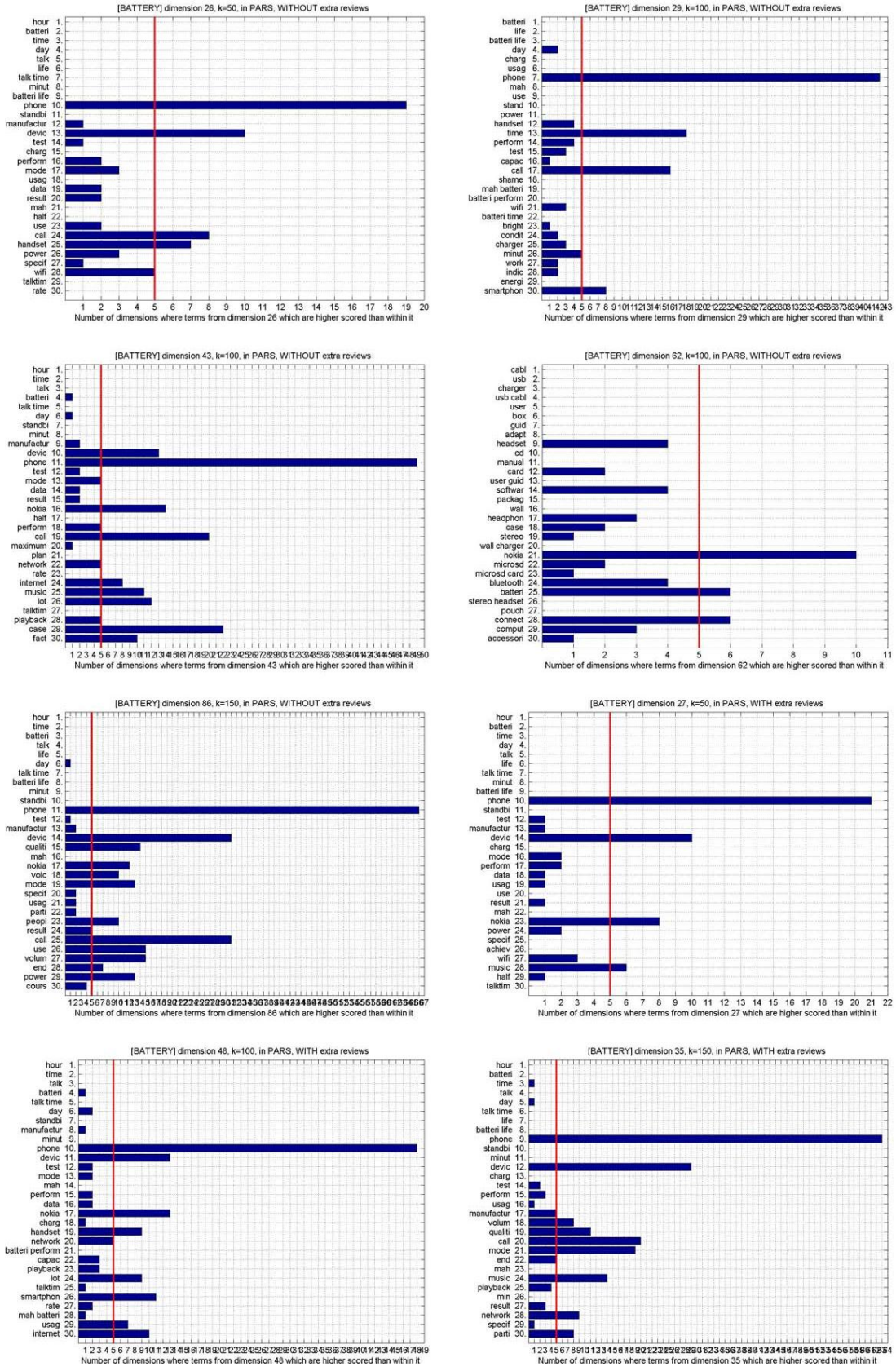
Graphic 14-1. Discarding method applied to top 30 terms from *battery* dimensions on documents.

14.1.2. Sections



Graphic 14-2. Discarding method applied to top 30 terms from *battery* dimensions on sections.

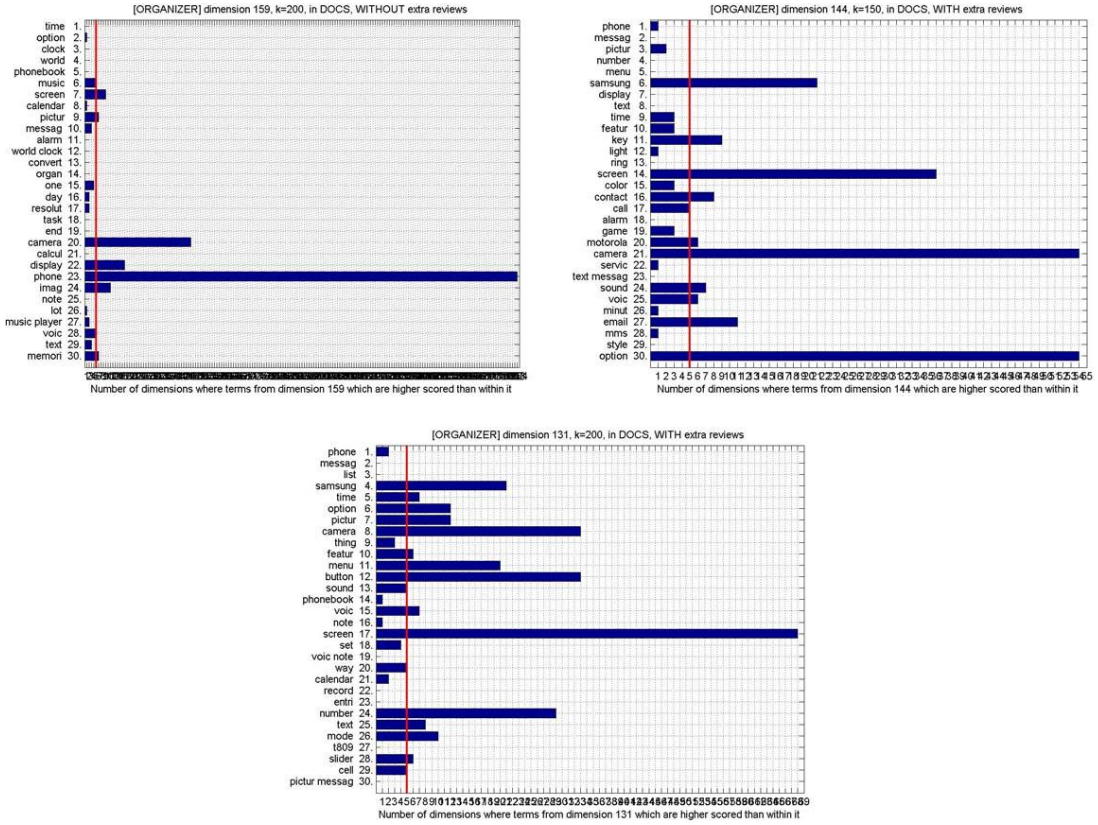
14.1.3. Paragraphs



Graphic 14-3. Discarding method applied to top 30 terms from *battery* dimensions on paragraphs.

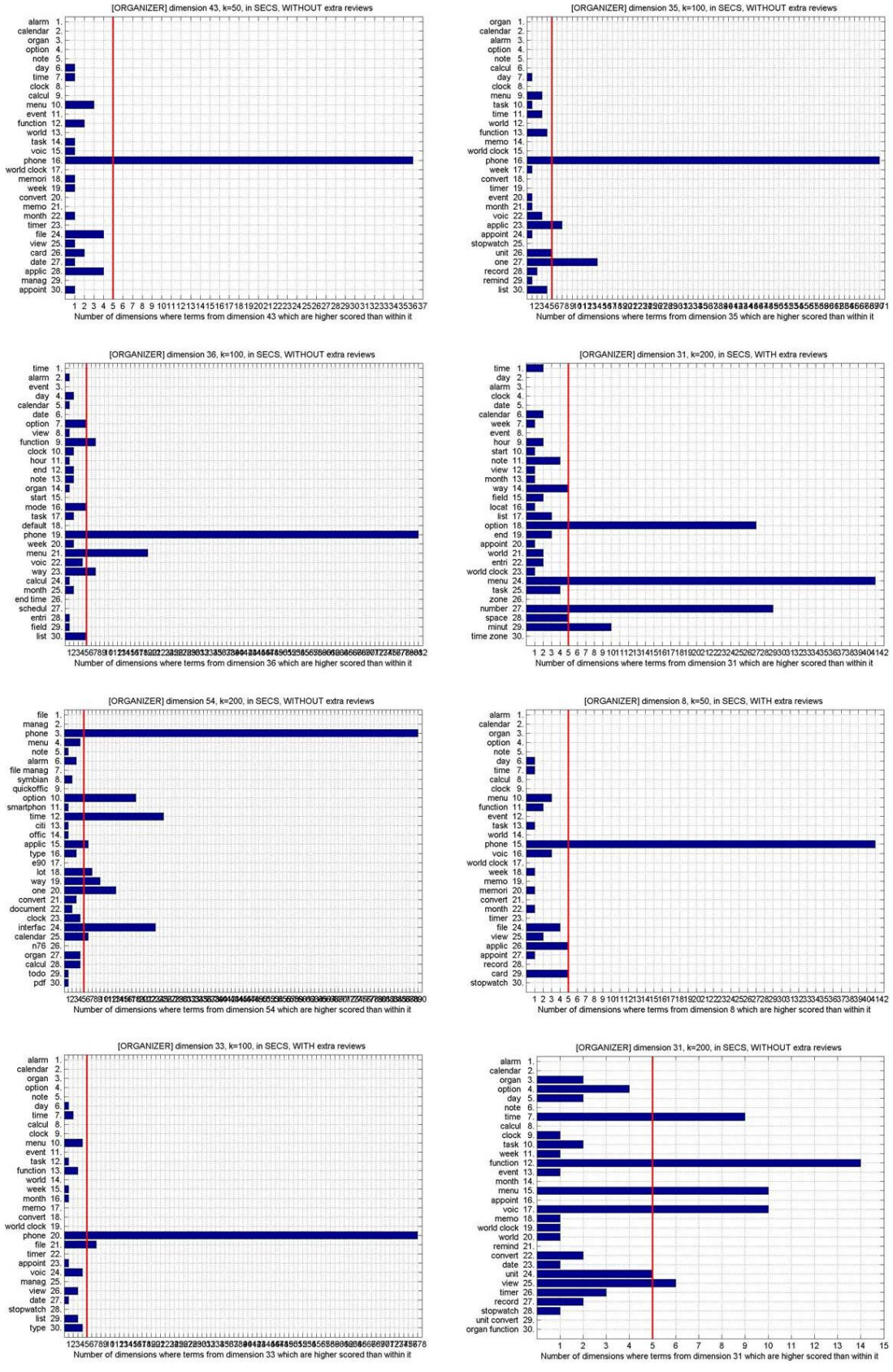
14.2. *Organizer*

14.2.1. Documents



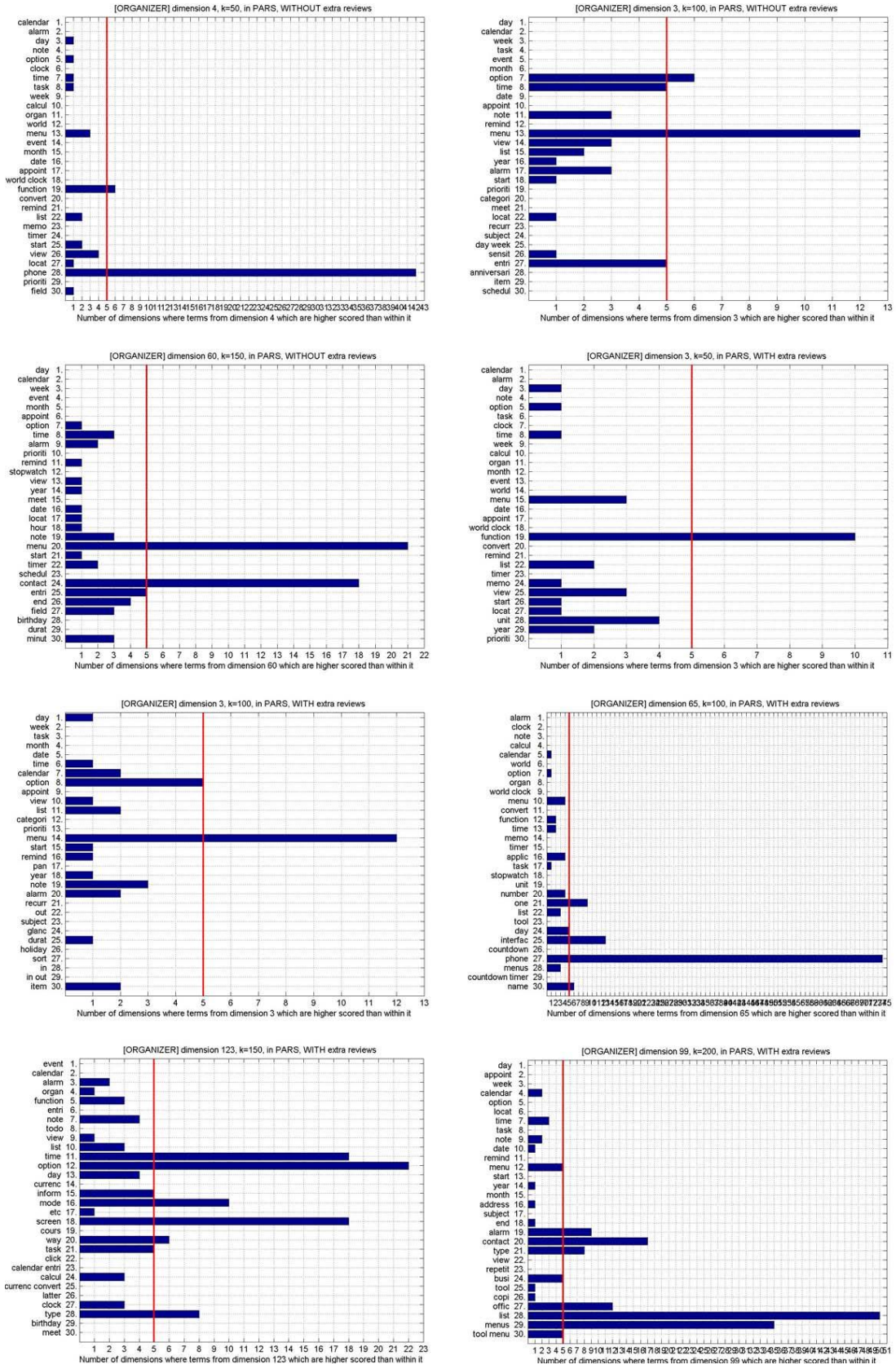
Graphic 14-4. Discarding method applied to top 30 terms from *organizer* dimensions on documents.

14.2.2. Sections



Graphic 14-5. Discarding method applied to top 30 terms from *organizer* dimensions on documents.

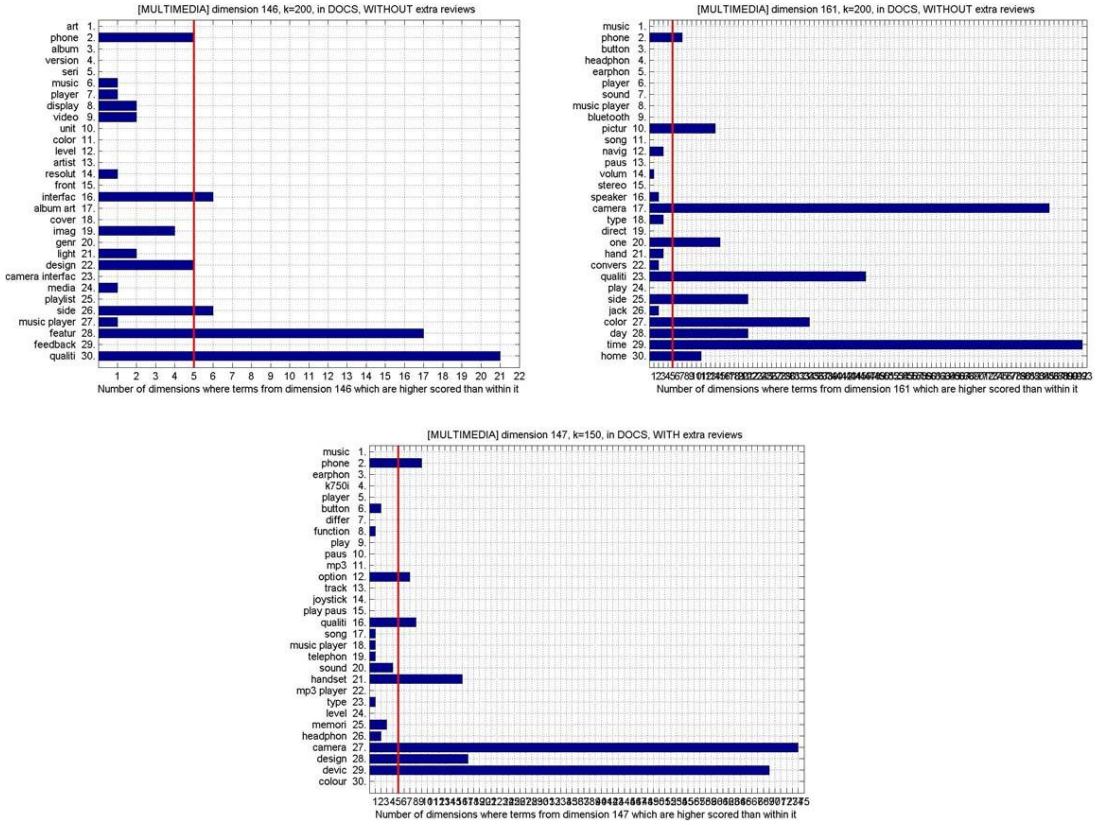
14.2.3. Paragraphs



Graphic 14-6. Discarding method applied to top 30 terms from *organizer* dimensions on paragraphs.

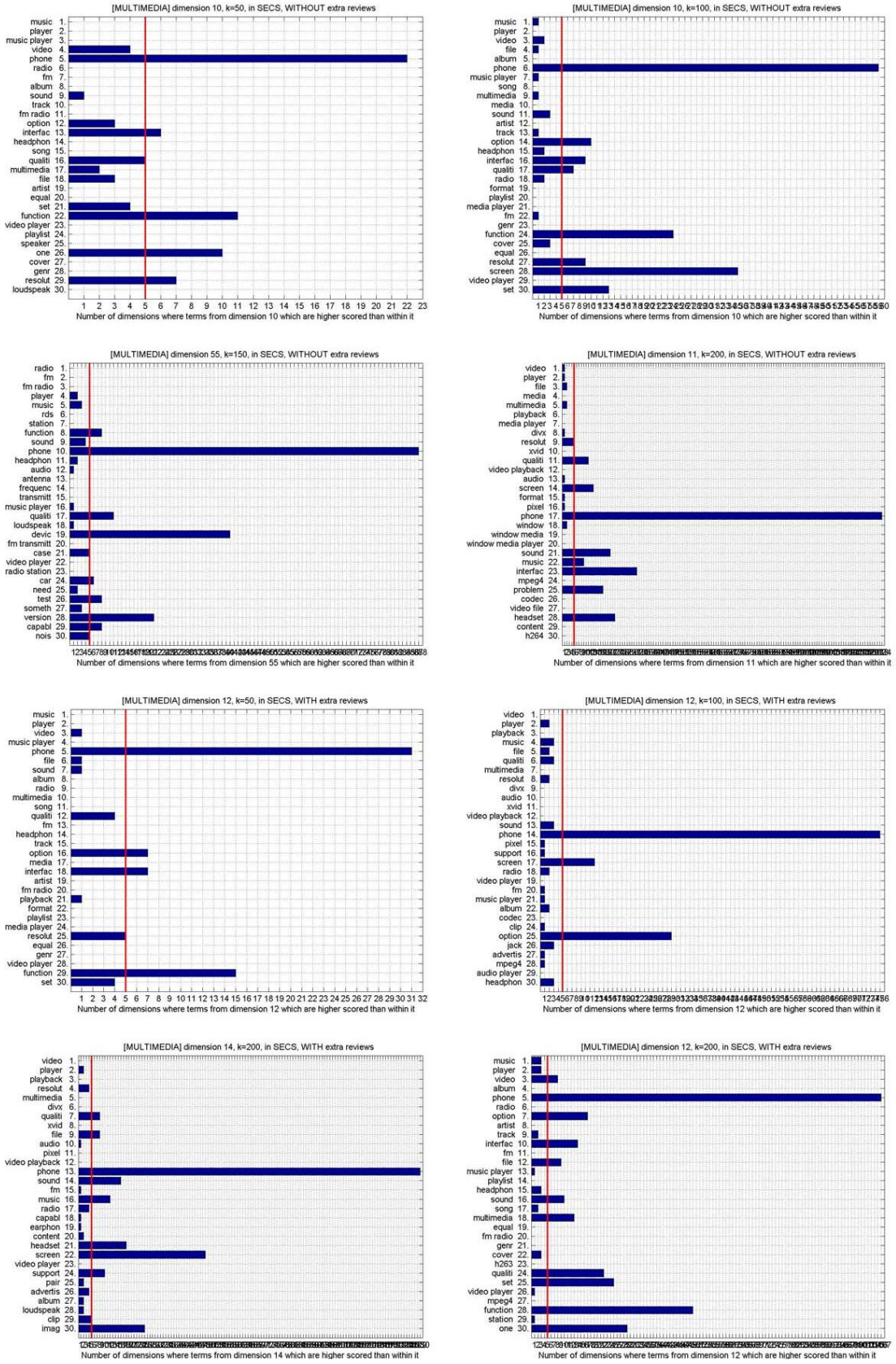
14.3. *Multimedia*

14.3.1. Documents



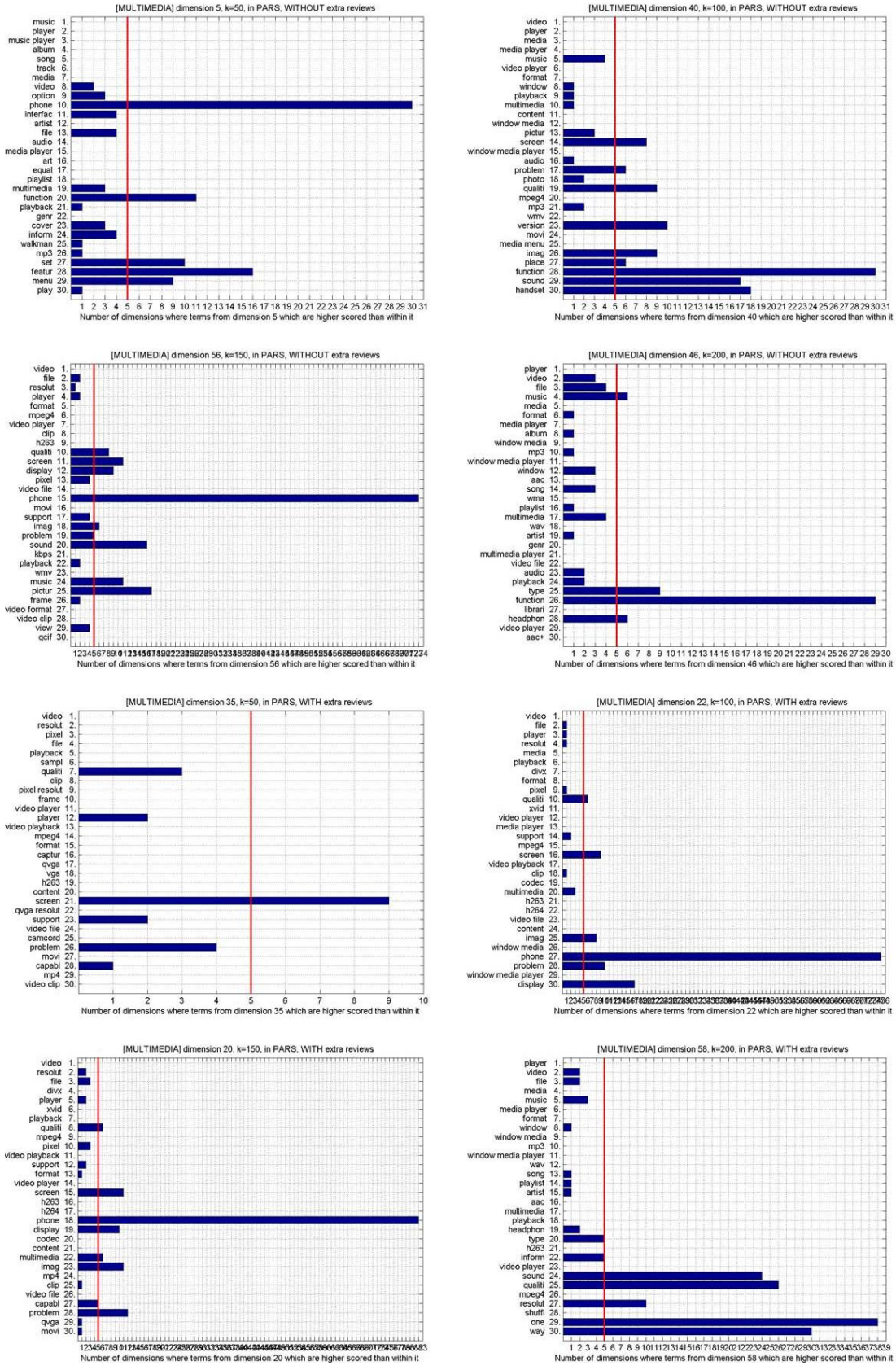
Graphic 14-7. Discarding method applied to top 30 terms from *multimedia* dimensions on documents.

14.3.2. Sections



Graphic 14-8. Discarding method applied to top 30 terms from *multimedia* dimensions on sections.

14.3.3. Paragraphs

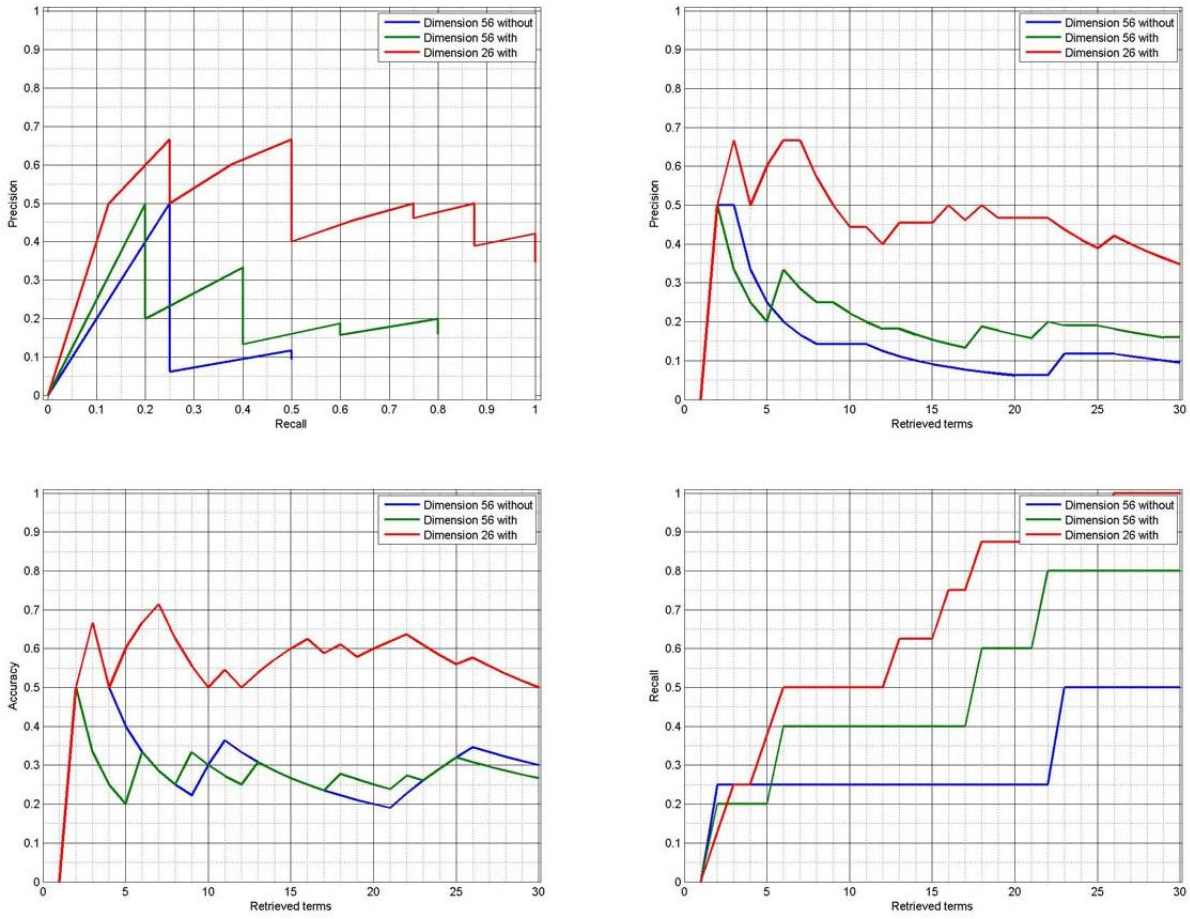


Graphic 14-9. Discarding method applied to top 30 terms from *multimedia* dimensions on paragraphs.

15. Annex G: Precision, recall and accuracy graphs of LSI

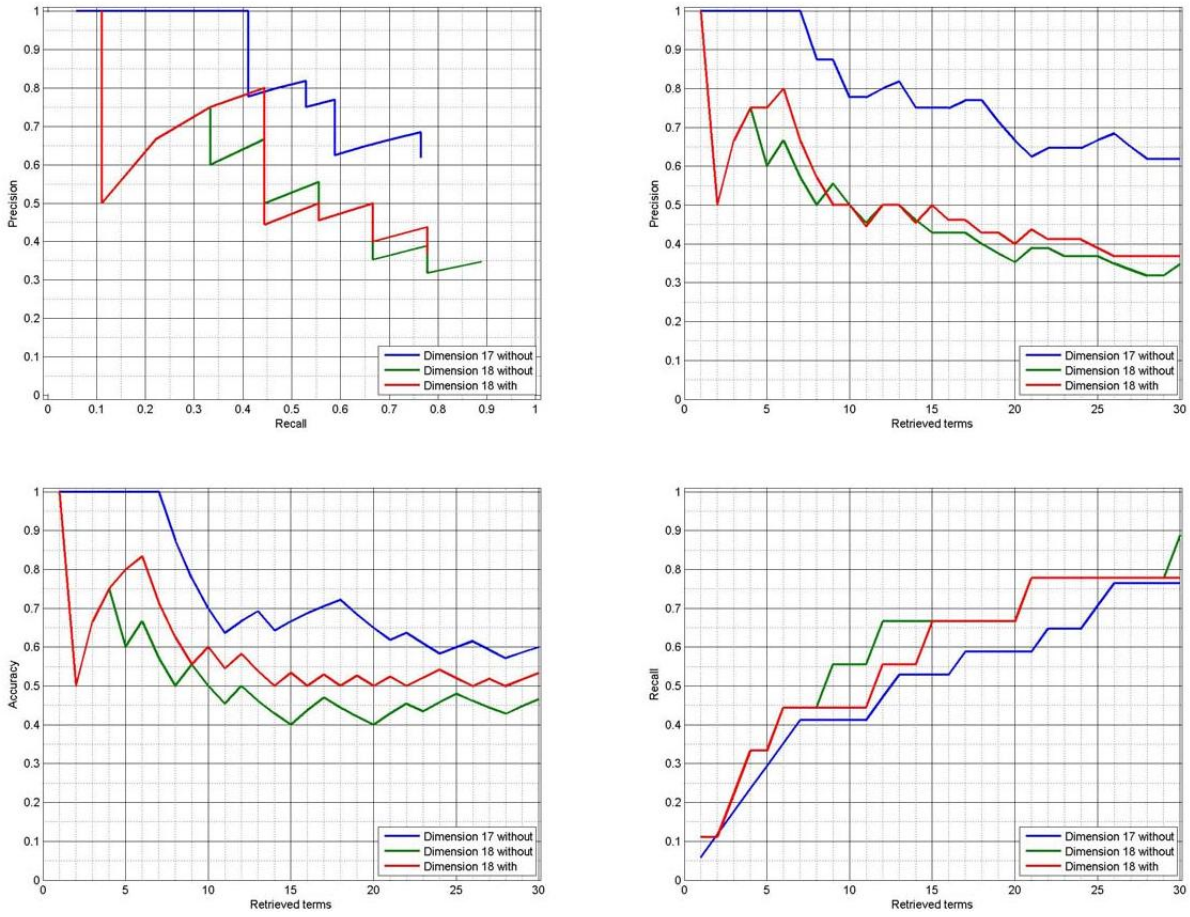
15.1. *Battery*

15.1.1. Sections



Graphic 15-1. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *battery* on sections.

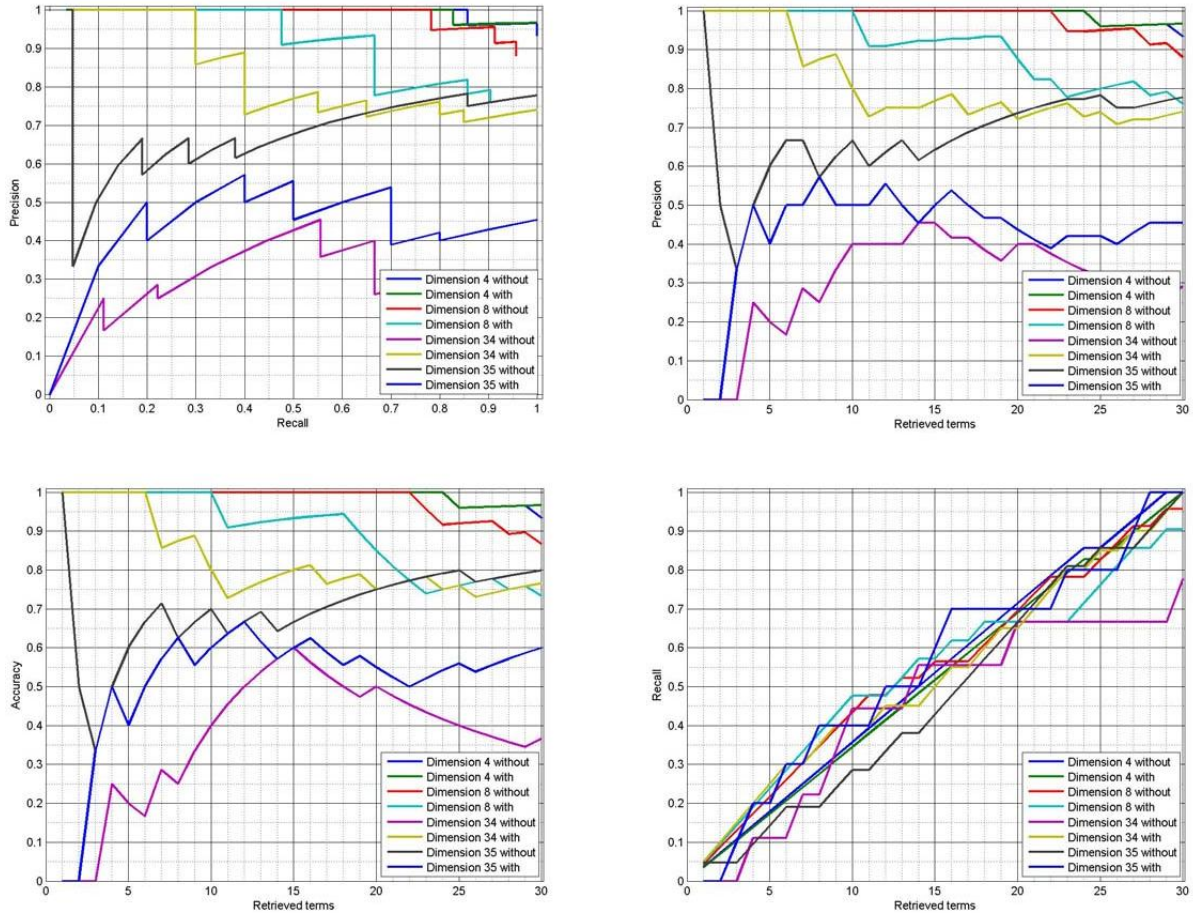
15.1.2. Paragraphs



Graphic 15-2. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *battery* on paragraphs.

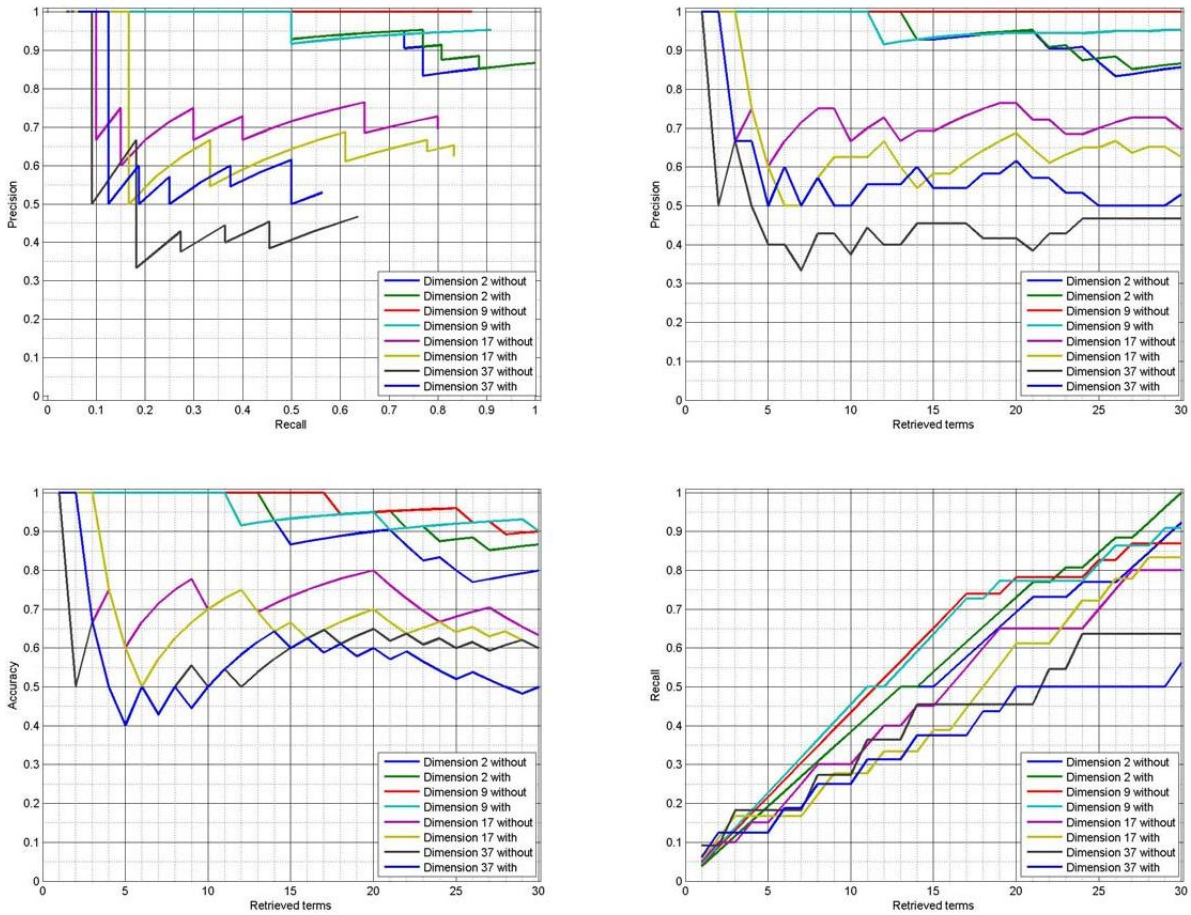
15.2. *Organizer*

15.2.1. Sections



Graphic 15-3. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *organizer* on sections.

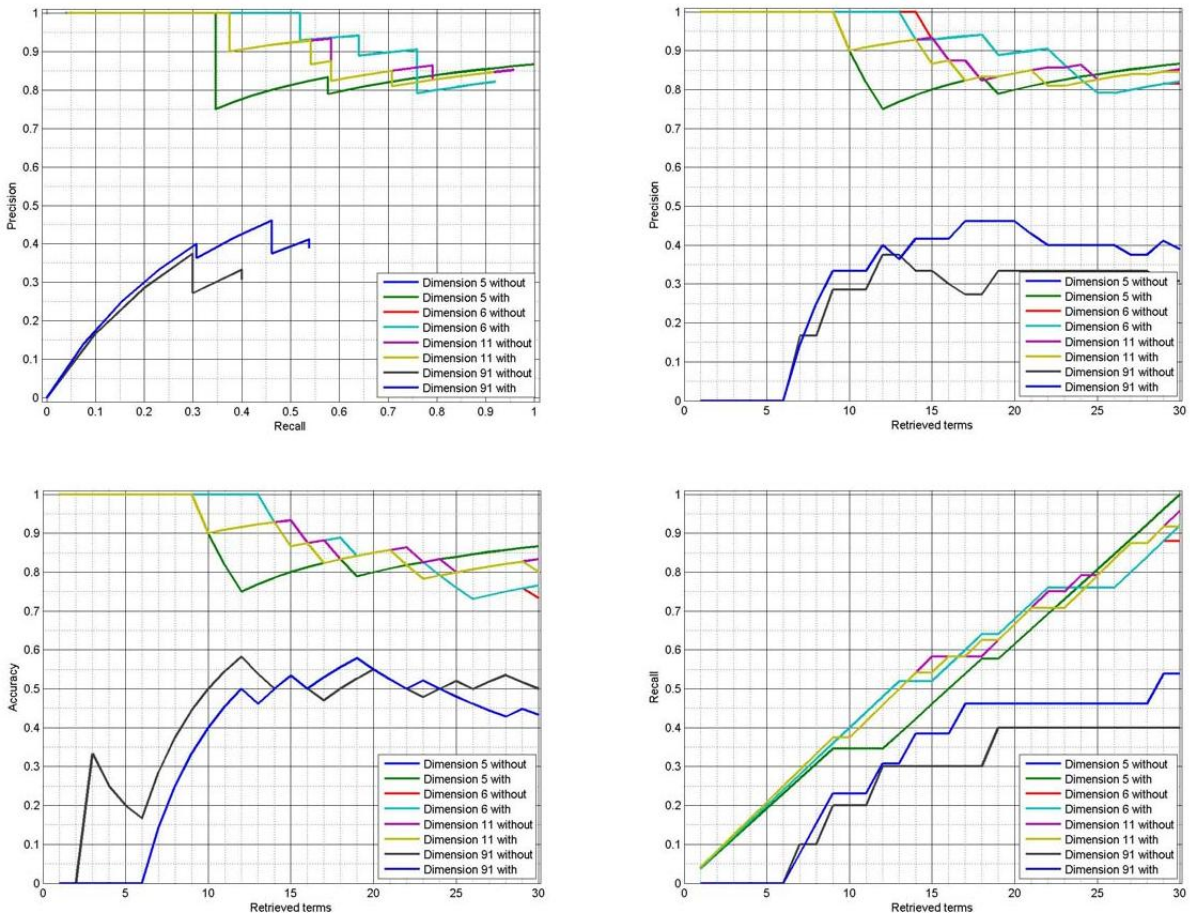
15.2.2. Paragraphs



Graphic 15-4. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *organizer* on paragraphs.

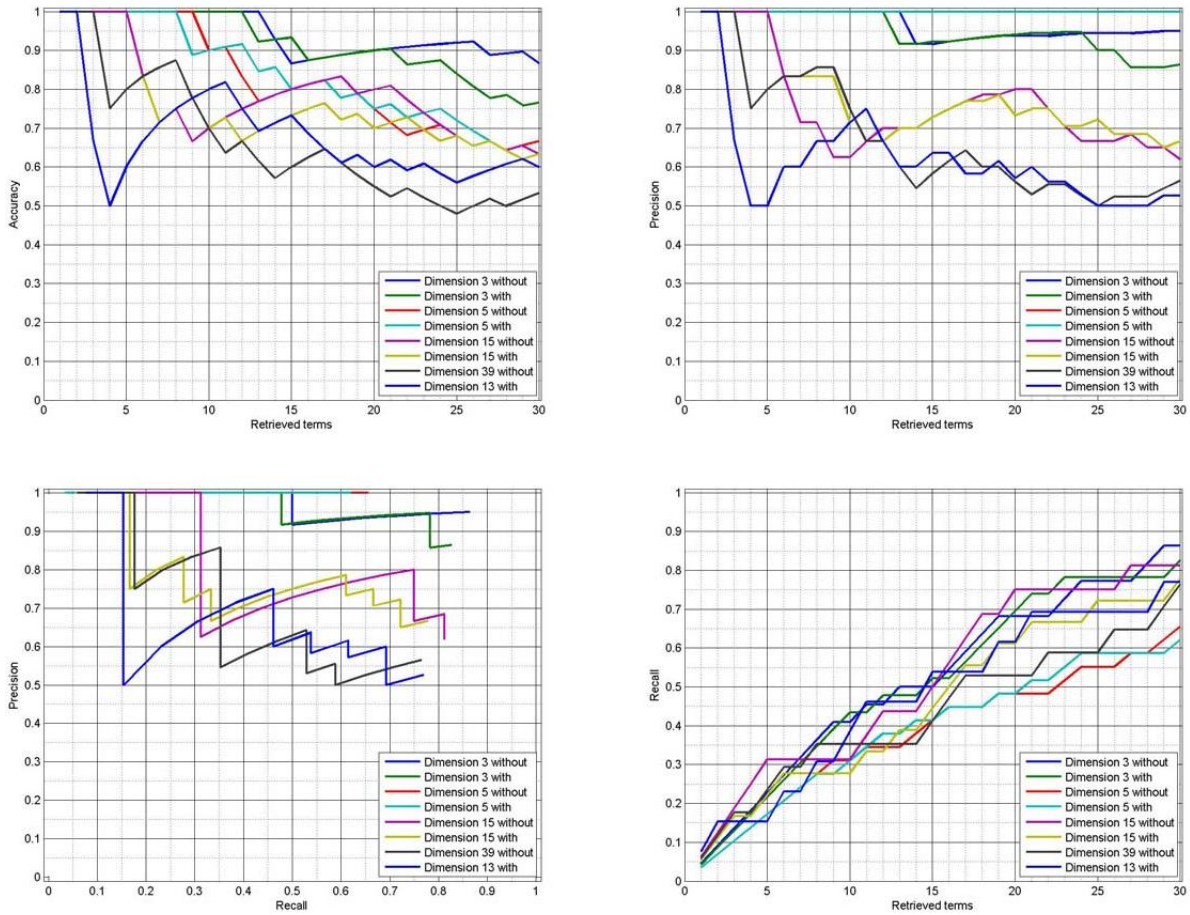
15.3. *Multimedia*

15.3.1. Sections



Graphic 15-5. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *multimedia* on sections.

15.3.2. Paragraphs

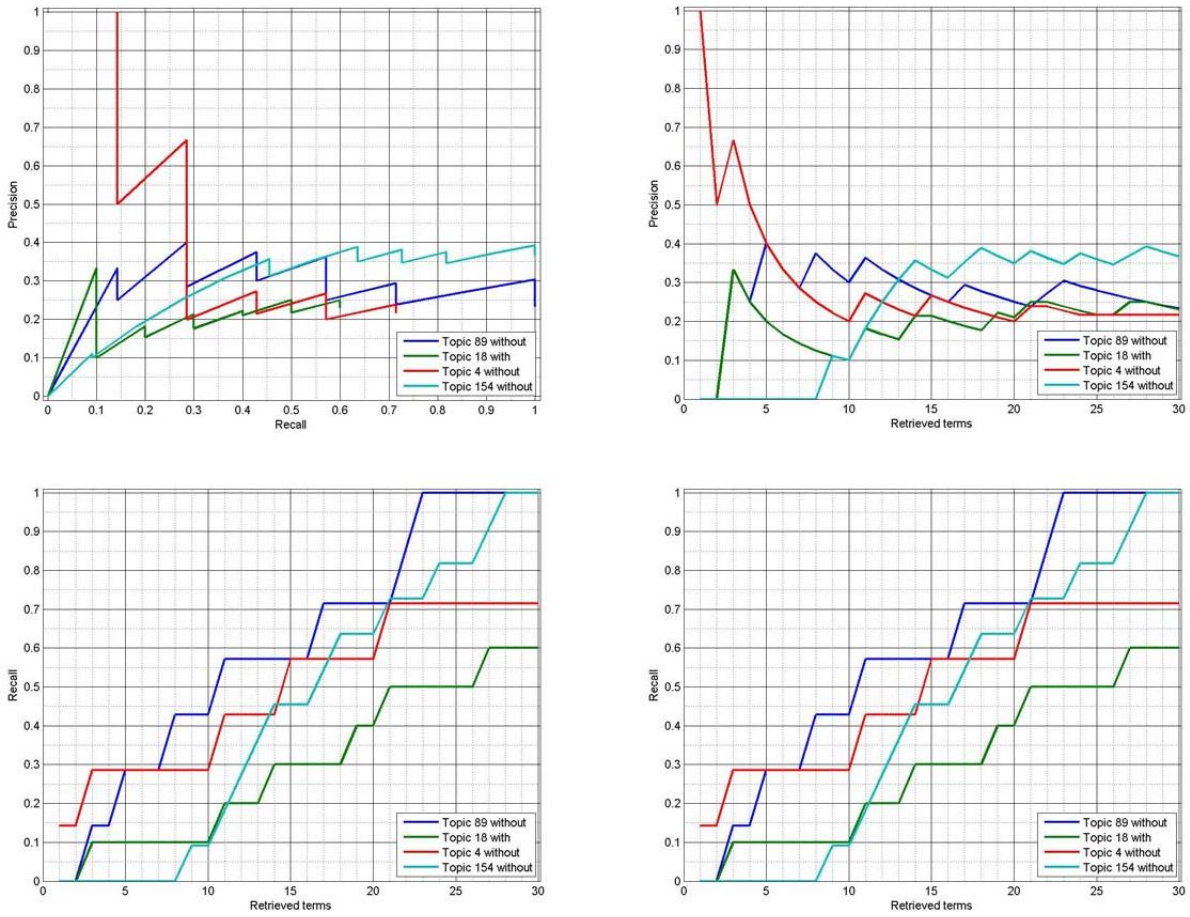


Graphic 15-6. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *multimedia* on paragraphs.

16. Annex H: Precision, recall and accuracy graphs of PSLI

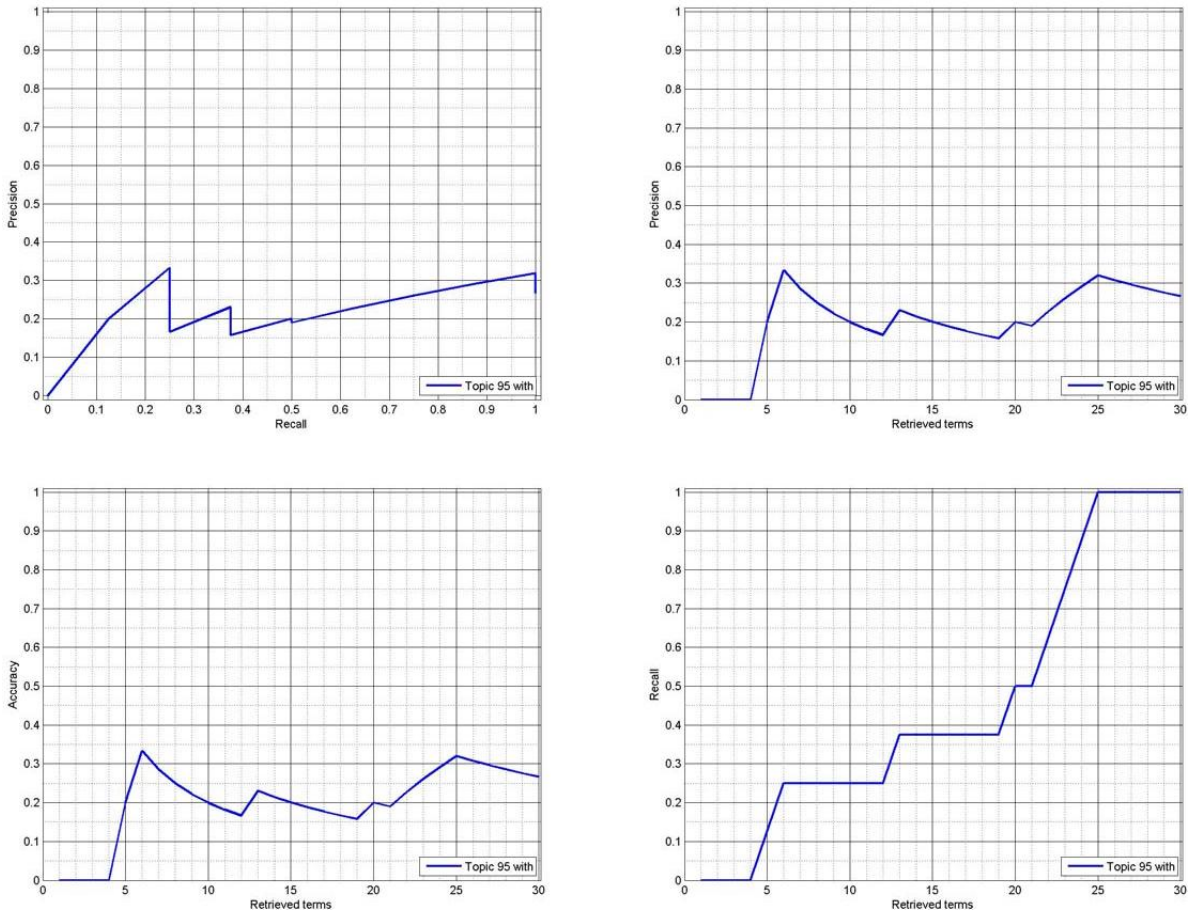
16.1. *Battery*

16.1.1. Documents



Graphic 16-1. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *battery* on documents.

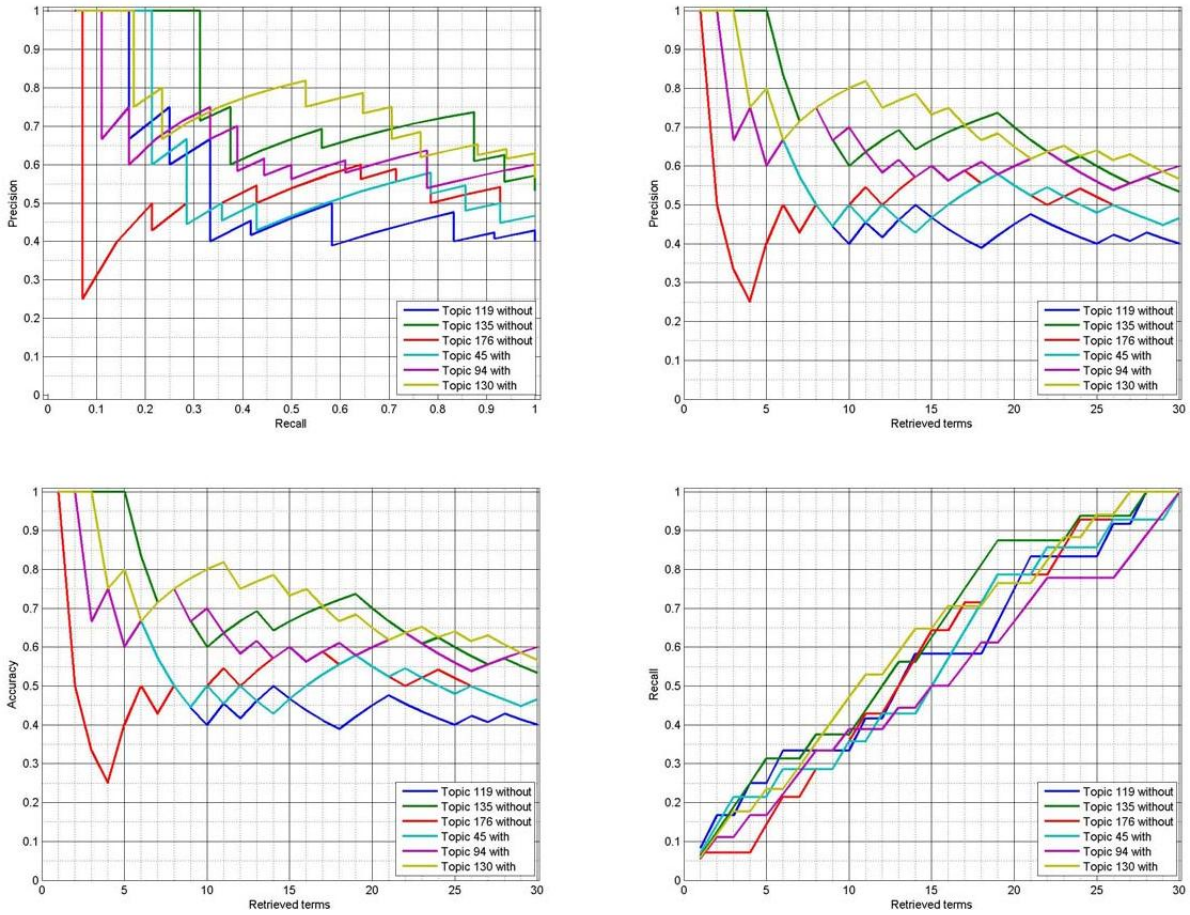
16.1.2. Sections



Graphic 16-2. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *battery* on sections.

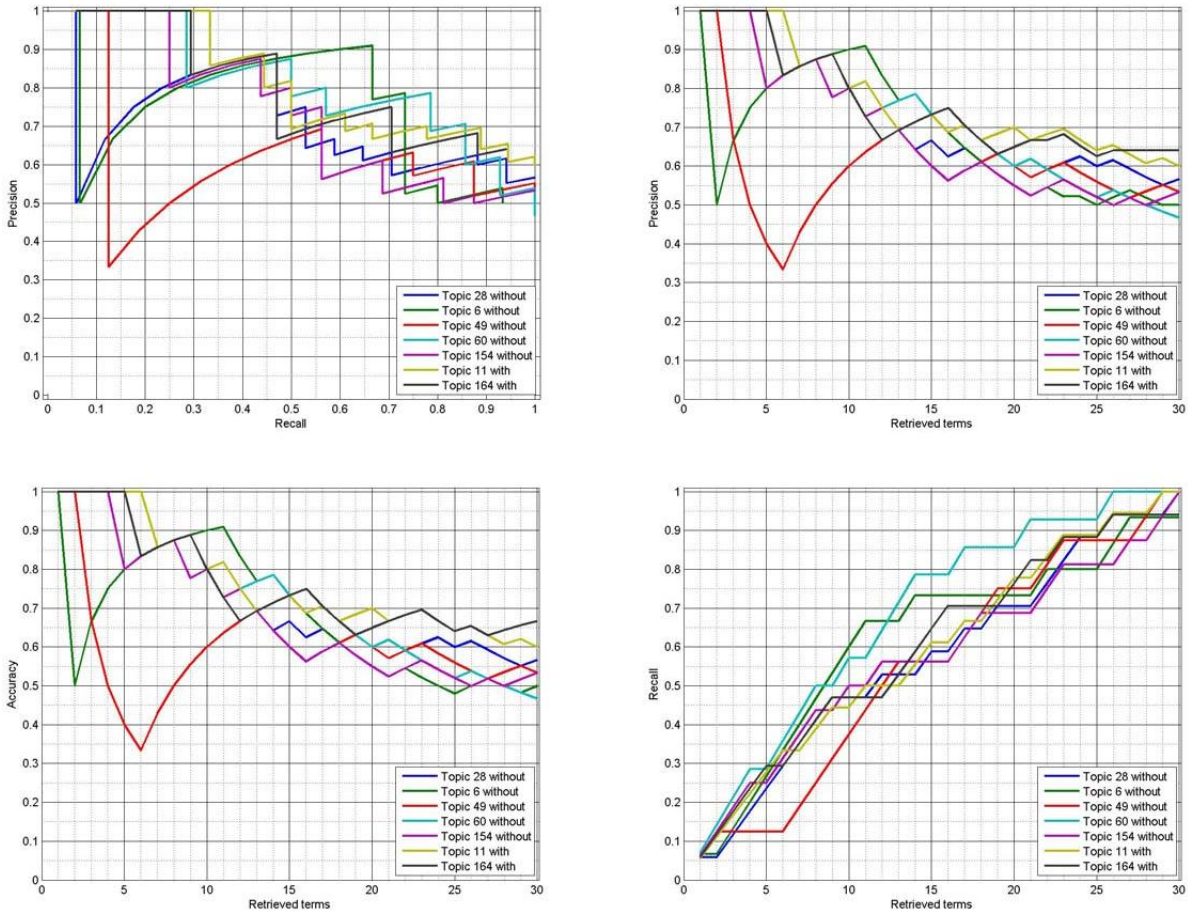
16.2. *Organizer*

16.2.1. Documents



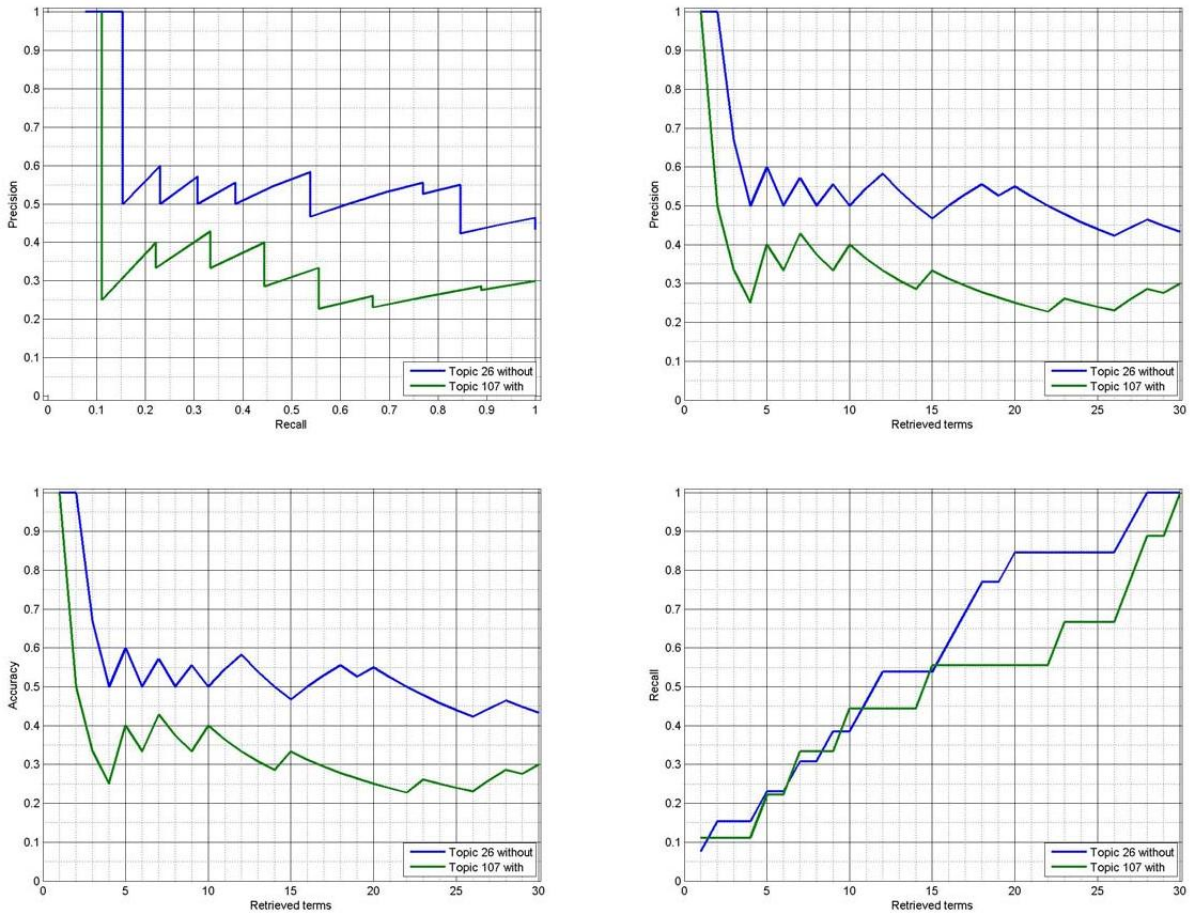
Graphic 16-3. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *organizer* on documents.

16.2.2. Sections



Graphic 16-4. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *organizer* on sections.

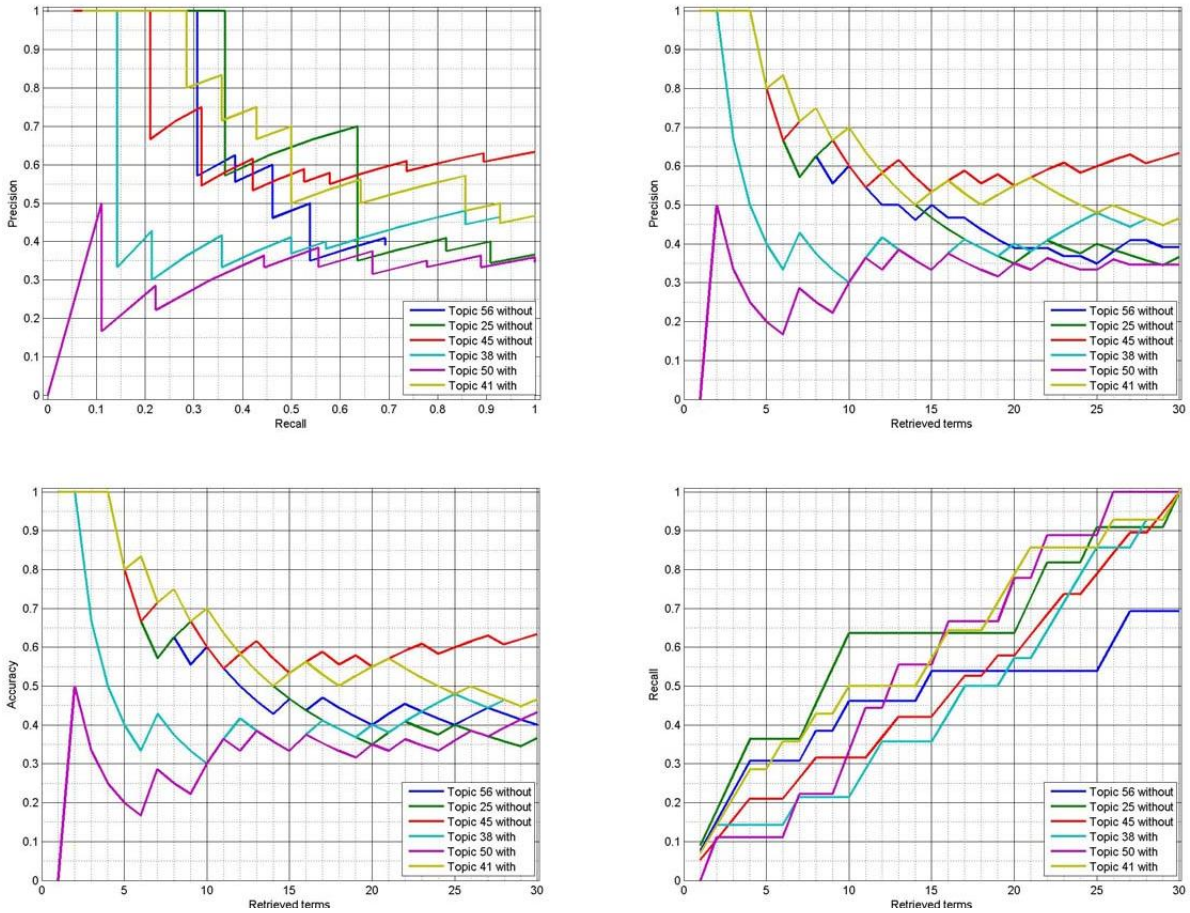
16.2.3. Paragraphs



Graphic 16-5. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *organizer* on paragraphs.

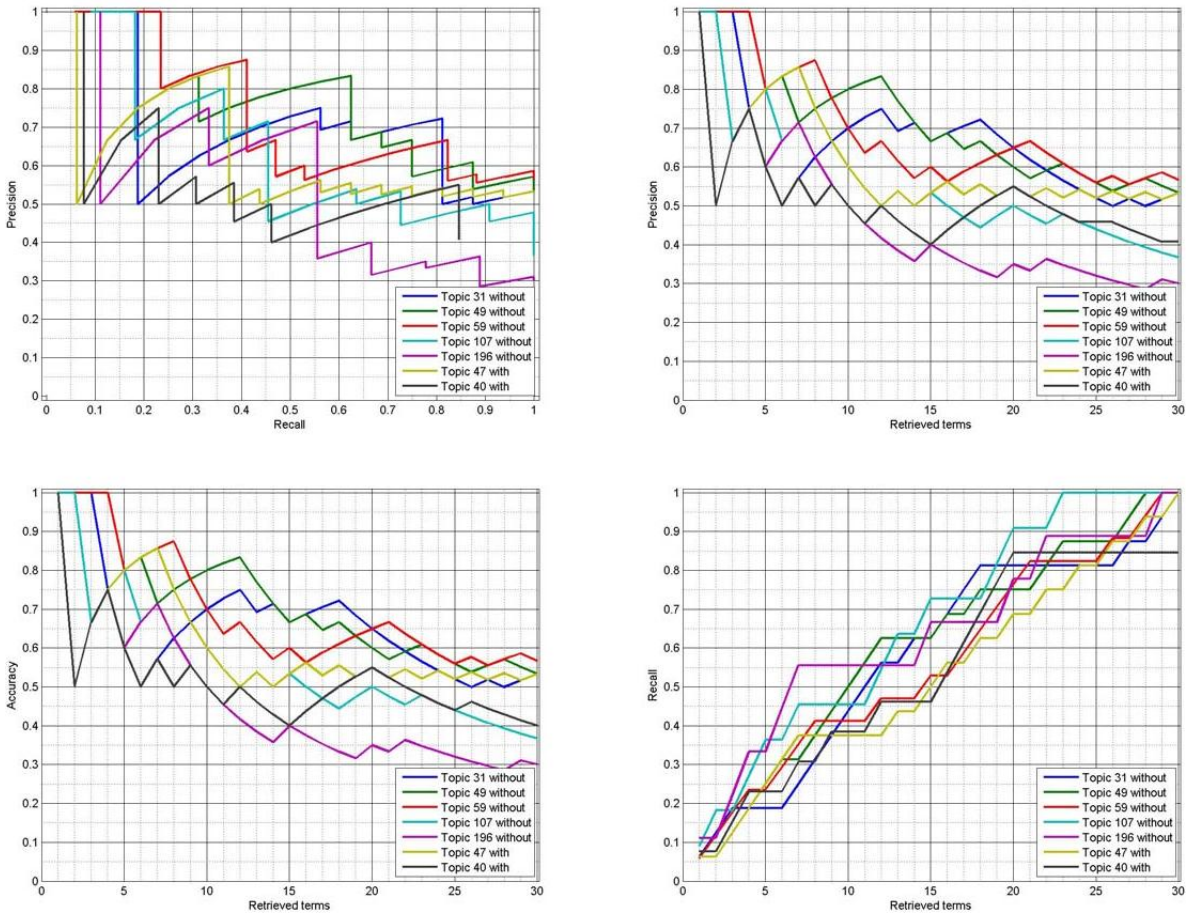
16.3. *Multimedia*

16.3.1. Documents



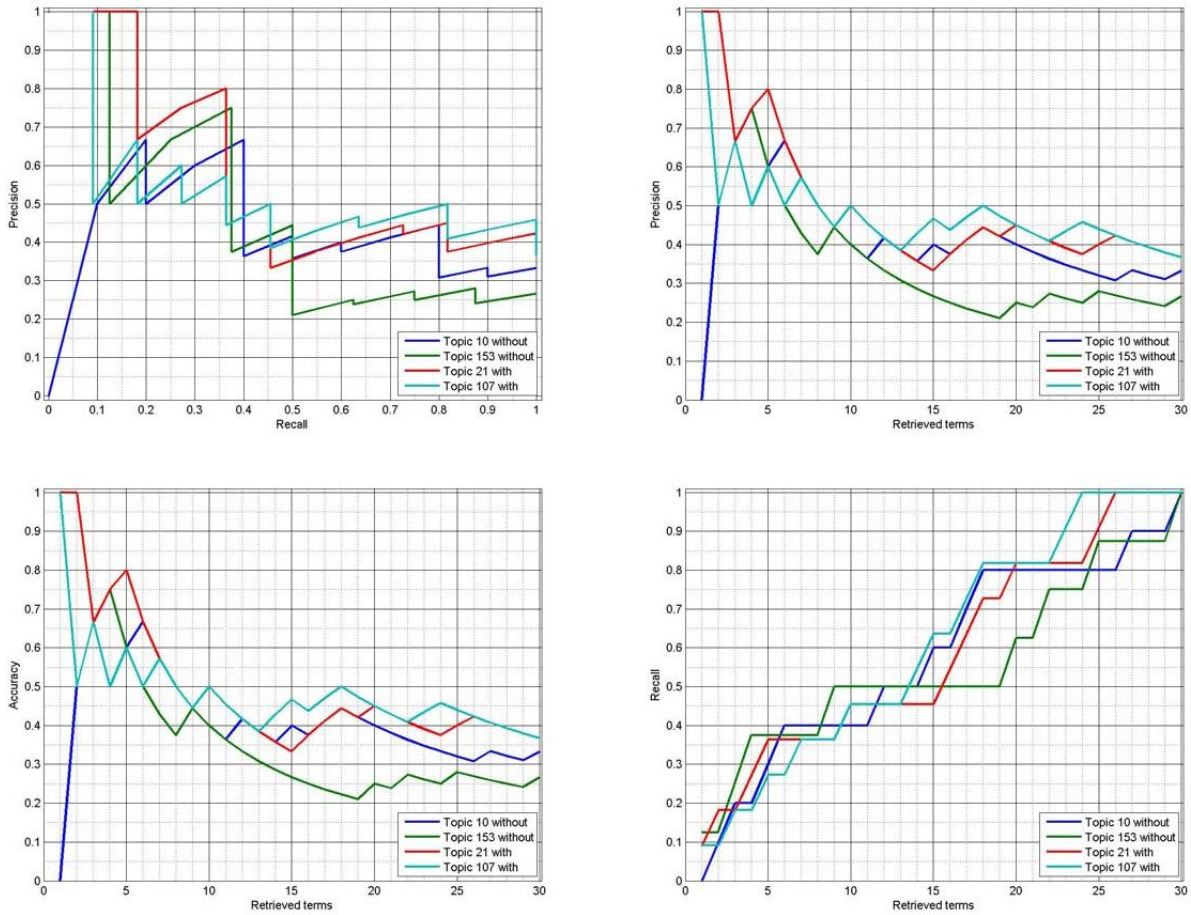
Graphic 16-6. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *multimedia* on documents.

16.3.2. Sections



Graphic 16-7. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *multimedia* on sections.

16.3.3. Paragraphs

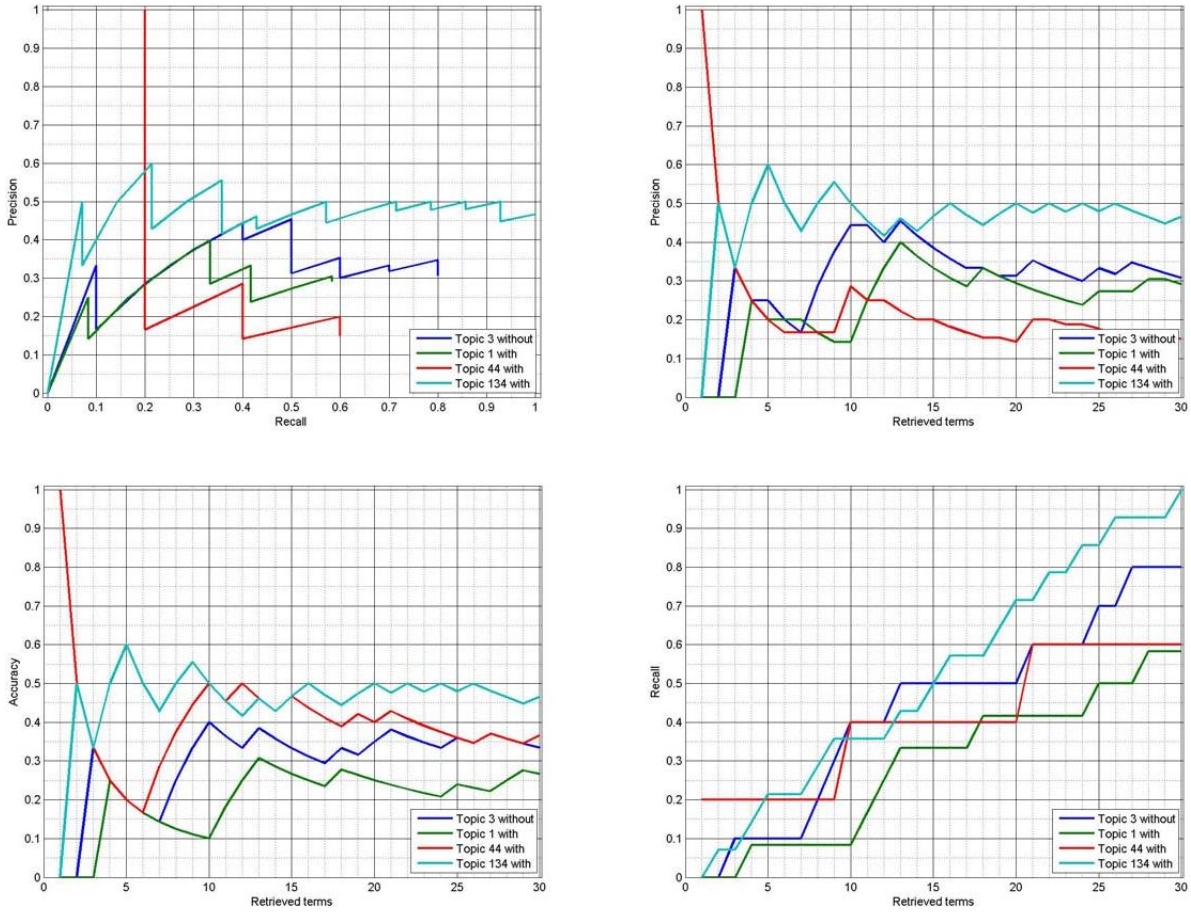


Graphic 16-8. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *multimedia* on paragraphs.

17. Annex I: Precision, recall and accuracy graphs of LDA

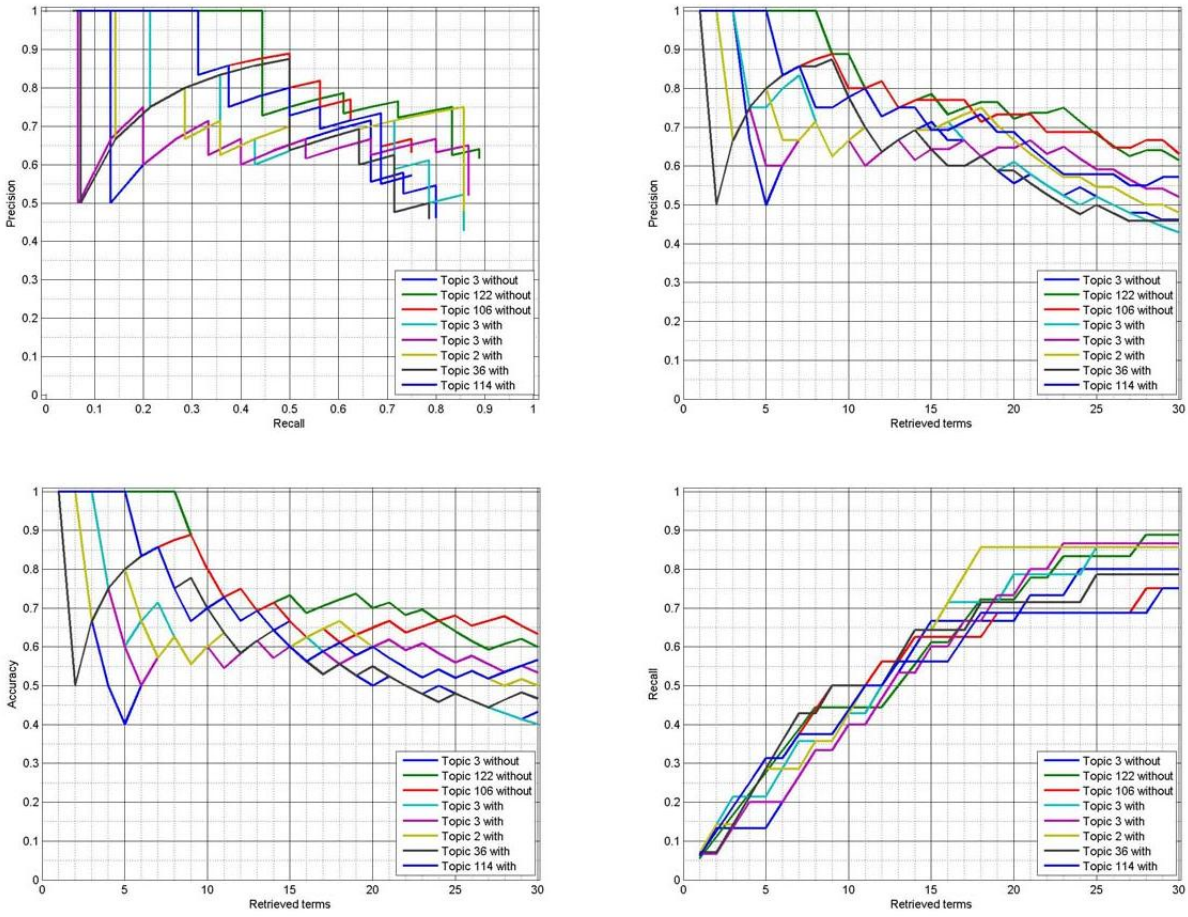
17.1. *Battery*

17.1.1. Documents



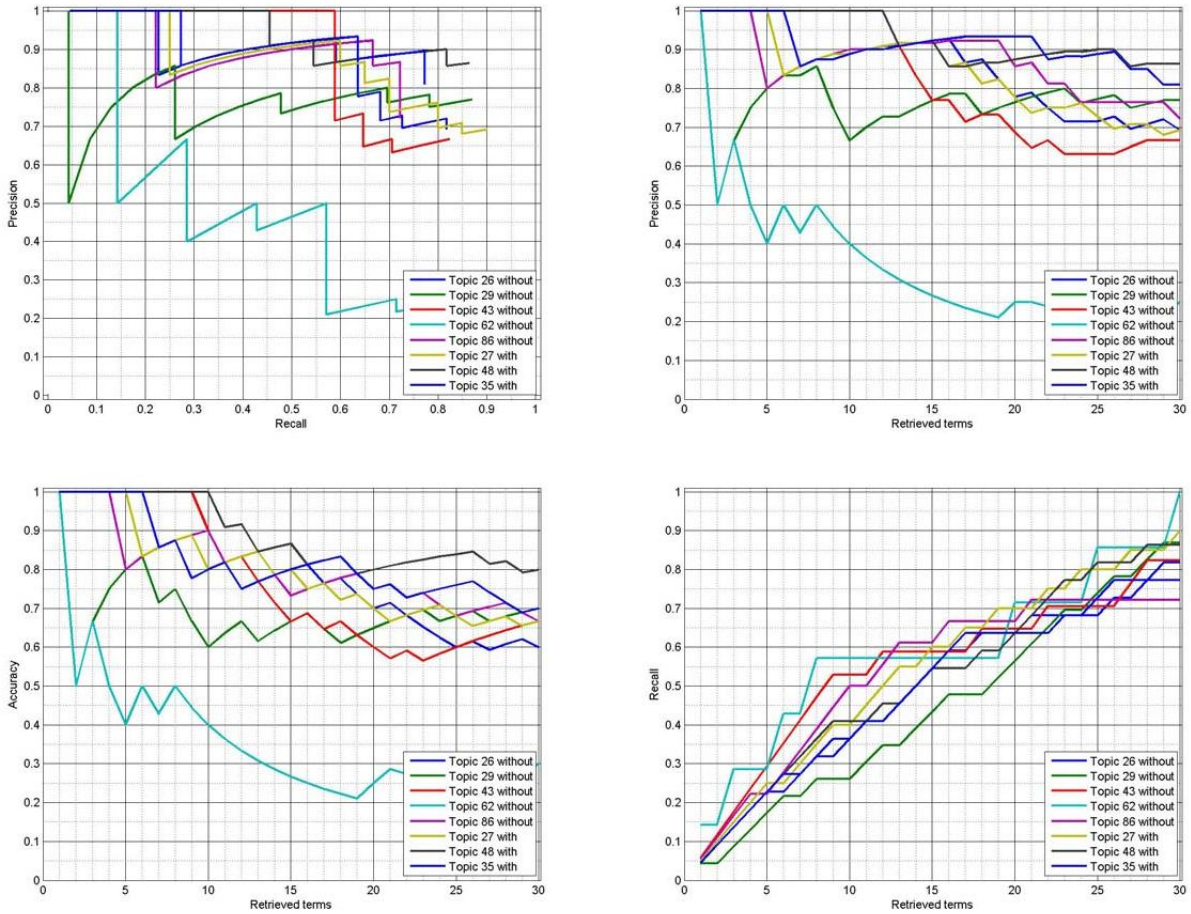
Graphic 17-1. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *battery* on documents.

17.1.2. Sections



Graphic 17-2. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *battery* on sections.

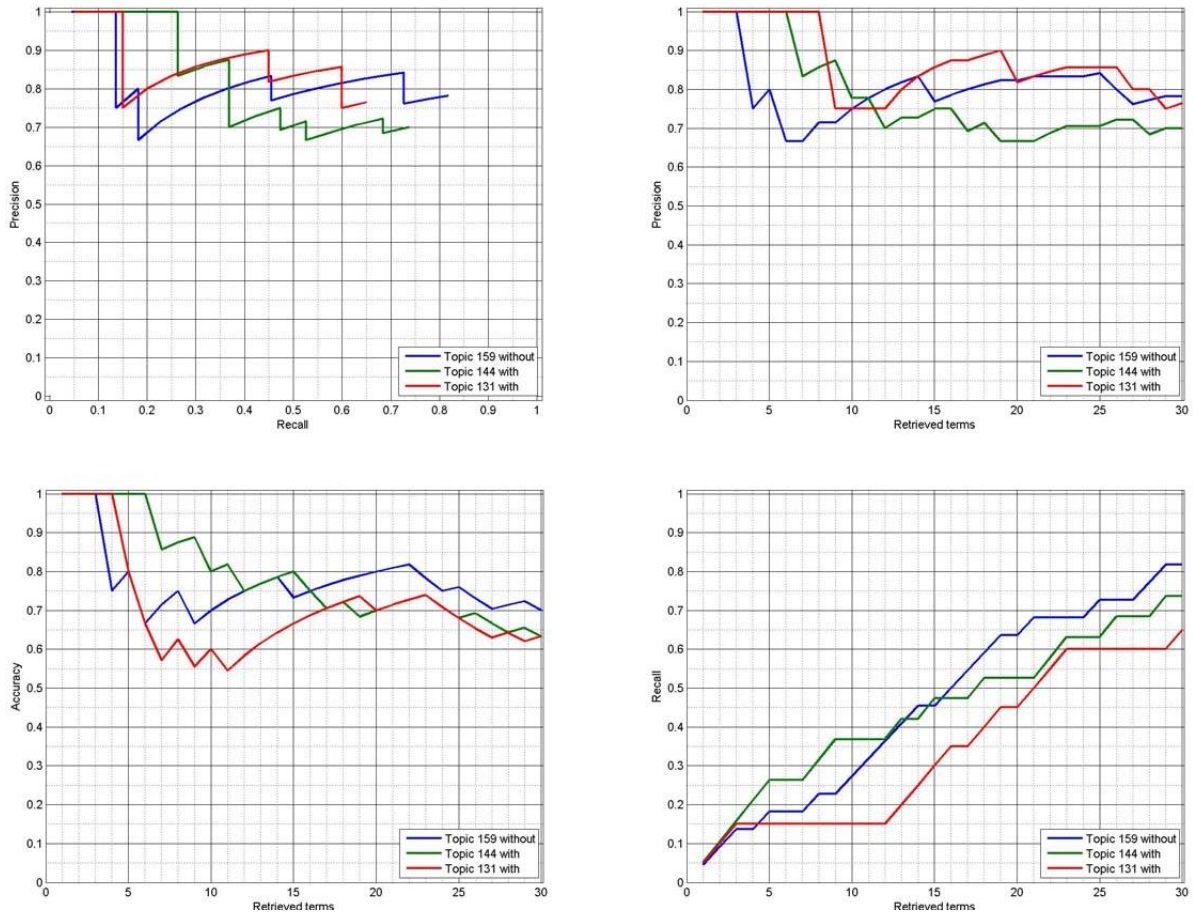
17.1.3. Paragraphs



Graphic 17-3. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *battery* on paragraphs.

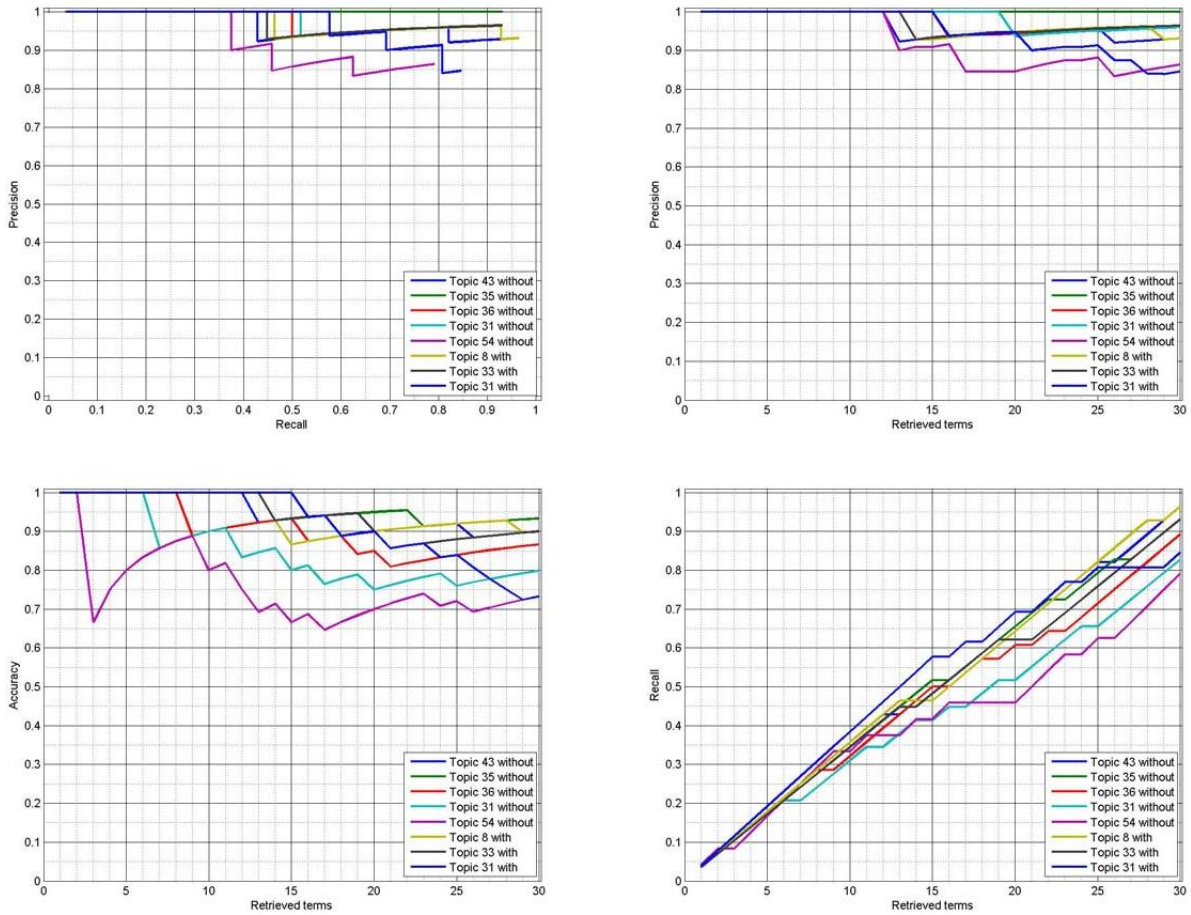
17.2. *Organizer*

17.2.1. Documents



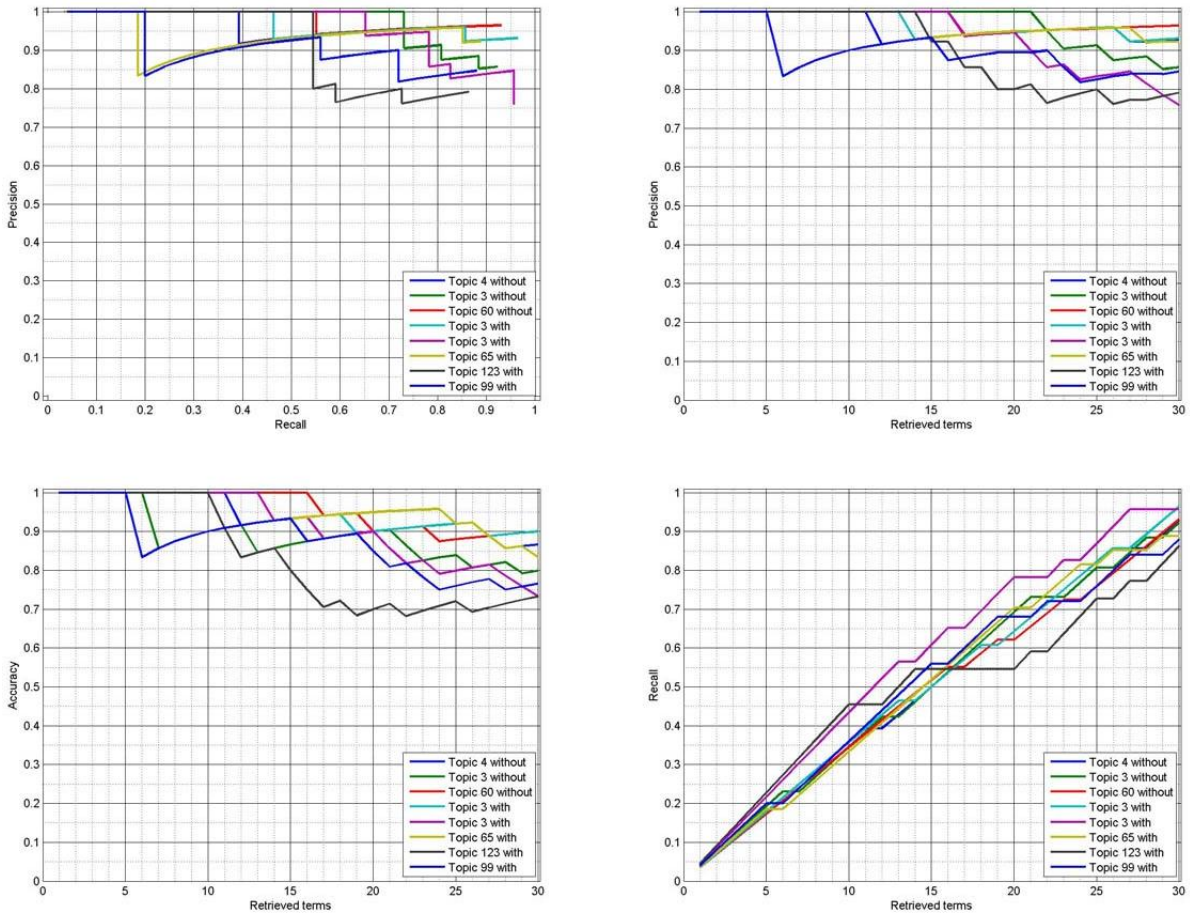
Graphic 17-4. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *organizer* on documents.

17.2.2. Sections



Graphic 17-5. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *organizer* on sections.

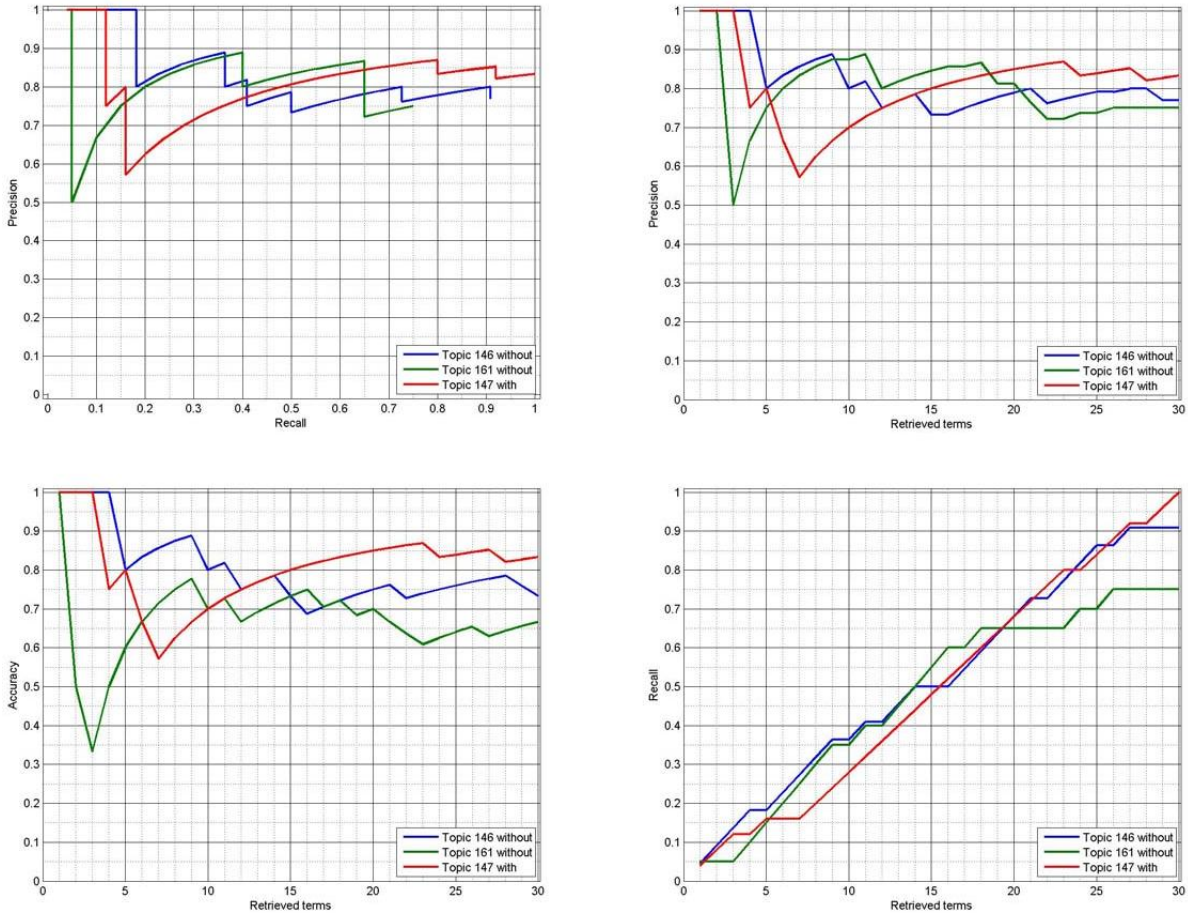
17.2.3. Paragraphs



Graphic 17-6. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *organizer* on paragraphs.

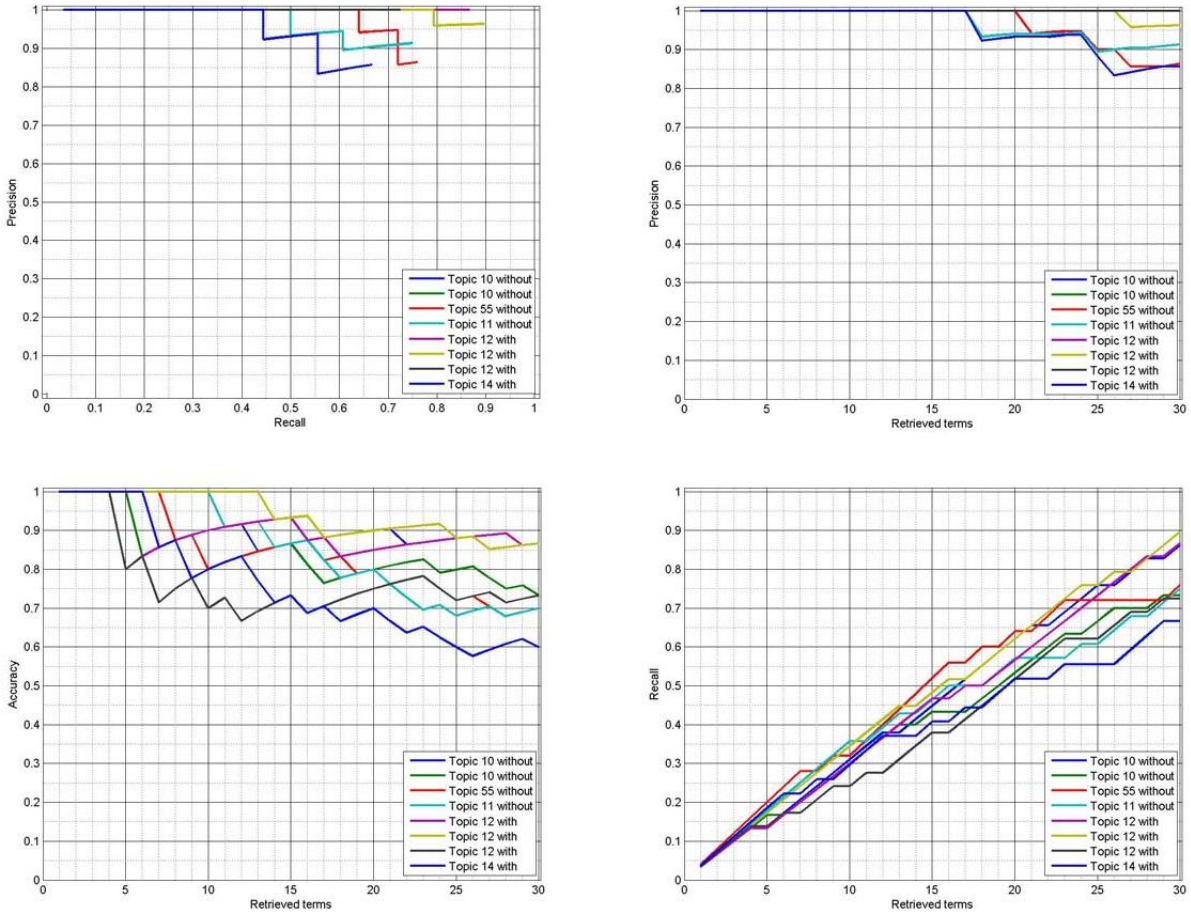
17.3. *Multimedia*

17.3.1. Documents



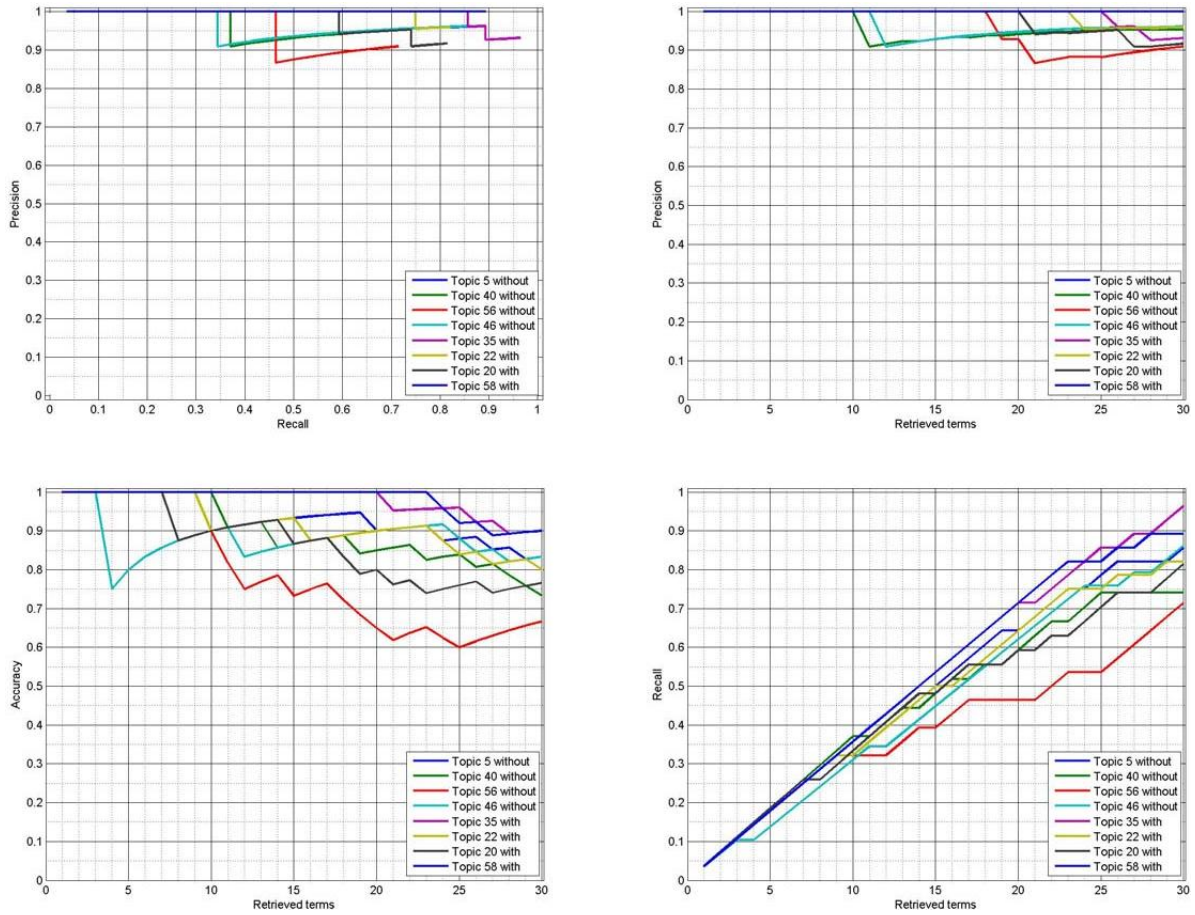
Graphic 17-7. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *multimedia* on documents.

17.3.2. Sections



Graphic 17-8. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *multimedia* on sections.

17.3.3. Paragraphs



Graphic 17-9. Precision vs recall, precision, accuracy and recall graphics of selected dimensions of *multimedia* on paragraphs.