# Topic Modeling - A Case Study with Scientific Production

Rui Sarmento

Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal
mail@ruisarmento.com

**Abstract.** The automatic extraction of topics from text is an area of research with a fast evolution in recent years. We use two different methods for topic extraction, Latent Dirichlet Allocation (LDA), and Term Frequency-Inverse Document Frequency (TF-IDF) and compare them in several experimental situations. We also test the use of stemming processing of data in the context of our dataset and with topic extraction in mind. We study these methods by applying them to a dataset composed by scientific literature production. These documents are the 2014 work of researchers from the Czech Republic. The results obtained are conclusive and provide decision support for the choice between these methods with our dataset. TF-IDF provided better results when compared with LDA. Additionally, TF-IDF with clustering approached LDA with increasing size of analyzed text data. Finally, all methods increase quality when applying stemming to the data. The goal is to provide, in a near future, a complete prototype developed for the purpose of the study and visualization of affinity between authors and their topics of studies or research.

**Keywords:** NLP · Researchers Dataset · Topic Modeling

## 1 Introduction

Some authors, for example, Blei [3], emphasize that generative statistical and distribution models for text have the potential to make important contributions to the statistical analysis of large document collections. One example of these generative models is the topic model. These models enable the use of refined statistical methods to identify the structure that underlies a set of words combined in phrases. Topic models might use many of the key assumptions behind Latent Semantic Analysis (LSA) but enabling the identification of a set of interpretable probabilistic topics rather than a semantic space. Investigating these models provides the opportunity to expand both the practical benefits and the theoretical understanding of statistical language learning.

Our contribution with this work is the study of clustering and its implications when applied to the author's data. We experiment with the TF-IDF matrix, and after applying cosine similarity to it. We aim to find representative elements in the cluster division of the obtained authors similarity data and their topics. We use these extracted topics from the exemplar or central element of the cluster to generalize to every cluster's element. We provide a comparison with a well-known benchmark method and obtain results for our dataset. We also test the application of stemming in topic extraction and provide results of this experiment.

Regarding the structure of this document, we start by including milestones and breakthroughs regarding topic model in 2. Topic modeling will be the main task regarding the goals of this research. Then, we also provide a brief statistical analysis of the dataset used for these experiment tasks in 3. We introduce the methodology for our experiments with topic modeling in 4. The results we obtained with our data are presented in 5 and finally, we conclude this document in 6 providing also some thoughts on the future work or necessary studies to improve this research.

## 2 Related Work

Several studies have already addressed Topic Modeling. This section first introduces related work with a more generic approach including an overview of research in this area. The selected publications imply milestones or novelties regarding this subject of Natural Language Processing (NLP) related research.

In a second subsection we introduce related work regarding Topic Modeling applied to scientific research data. Thus, we describe the previous NLP applications developed, focusing on research text data.

### 2.1 Topic Modeling - Milestones and Approaches

Lin and Hovy claim that only about 30% of topic keywords are not mentioned in the text directly [9]. The authors conclude that only about 30% of the humans' abstracts in this domain derive from some inference processes. They also conclude that in a computational implementation, only about the same amount has to be derived by processes yet to be determined with further research. The authors wrote that the titles contain about 50% of the topic keywords; the title plus the two most rewarding sentences provide about 60%, and the next five or so add another 6%. The authors, therefore, conclude that a fairly small number of sentences provides 2/3 of the keyword topics.

Lin and Hovy provide empirical validation for the Position Hypothesis. The authors also describe a method of deriving an Optimal Position Policy for a collection of texts within a genre, as long as a small set of topic keywords is defined with each text. The Precision and Recall scores indicate the selective power of the Position method on individual topics. Additionally, the Coverage scores indicate a kind of upper bound on topics and related material as contained in sentences from human-produced abstracts.

Lean and Hovy evaluations treat the abstract as ideal - the authors rest on the assumption that the central topic(s) of a text are contained in the abstract made of it. In many cases, this is a good assumption; it provides what one may call the author's perspective of the text. But this assumption does not support goal-oriented topic search, in which one wants to know whether a text pertains to some particular prespecified topics.

Topic models have also been extended to capture some properties of language, such as the hierarchical semantic relations between words [4], and the interaction between syntax and semantics [6].

Titov and McDonald [14] presented multi-grain topic models and claimed that they are superior to standard topic models when extracting ratable aspects from online reviews. According to the authors, these models are suited to this problem since they enable the identification of important terms, but also cluster them into consistent groups, which is a handicap of previously proposed methods.

AlSumait et al. [2], developed an online topic model for discrete data to model the temporal evolution of topics in data streams. The researchers used a non-Markov on-line LDA Gibbs sampler topic model (OLDA), in which the current model, along with the new data, guide the learning of a new generative process that reflects the dynamic changes in the data. They achieved this by using the generated model, at a given time, as a prior for LDA at the successive time slice, when a new data stream becomes available for processing. The weight of history in the generative process can be controlled by the weight matrix depending on the homogeneity of the domain. The authors claimed that the model results in an evolutionary matrix for each topic in which the evolution of the topic over time is captured. In addition, the authors proposed an algorithm to detect emerging topics based on the framework of OLDA. By processing small subsets of documents only, OLDA is claimed to enable the learning of meaningful topics, in some cases with higher quality than the LDA baseline. Additionally, the authors claim their method also outperforms LDA in detecting topics represented by a small set of documents at a certain point in time.

Sendhilkumar et al. [13], claim that their hPAM method is better to topic model research articles as the authors experienced better performance in terms of accuracy, precision and recall for retrieval of relevant documents. The authors include originality (inverse of similarity) as a parameter to define novelty in the documents. The described approach is not fully quantitative as it considers the semantics of concepts in the research article. The authors add they will be focused in further implementations and a qualitative approach for research articles, involving sentence importance and sentence contribution to novelty.

Gansner et al. [5], experiment streaming topic extraction with LDA and TF-IDF and argued that when extracting topics from short texts like twitter posts, the authors have better results with TF-IDF. Nonetheless, the authors did not experiment with the online version of LDA, the OLDA. According to the previous cited publication [2], when compared with LDA, OLDA presents better results than LDA for some use cases.

### 2.2 Topic Modeling of Scientific Production

The vast majority of the analysis of scientific production uses citations [12]. This includes the analysis of metrics as the frequency, patterns and graphs of citations in scientific written production like articles and books. Another approach already explored is the discovery of similarities between researchers. It was addressed by Price et al. [11], aiming to facilitate the process of paper distribution to reviewers. Their web-based methodology, called SubSift [11], retrieves researcher profiles based on their publications. These profiles enable a typical Information Retrieval task. The papers submitted to a scientific conference – playing the role of Query in IR – are compared with different profiles, in order to optimize the task of attributing articles for review.

Another approach was presented by Trigo et al. [15]. The authors present an Information Retrieval tool that facilitates the task of the user when searching for a particular information that is of interest to him. Trigo et al. [15] propose a system that processes a dataset of documents to produce a graph. This graph nodes represent documents and the links define similarities between nodes. The authors aim to offer the user a tool to navigate in the space of documents in an easy way. The authors present a case study that shows affinity groups based on the text production of researchers, beyond the previously established communities revealed by co-authorship. It characterizes the activity of each author by a set of automatically generated topics/keywords and by membership to a particular affinity group. The authors also provide validation methods of the most relevant information to be retrieved from researchers publications, analyse the impact of titles, abstracts and keywords on capturing the similarity between researchers.

Topic models such as LDA [4] and hierarchical models [8] have been successfully applied to various publications such as The American Political Science Review and Science. The work of Hall et al. [7] introduce the study of the history of ideas developments by using LDA and topic entropy. In [10] the authors extend over the work of Hall et al. [7] by adding two related fields (Linguistics and Education) and by employing various novel topic models for scientific research analysis.

## 3 Case Study

In this case study, we selected R&D publications from Czech Republic researchers. The dataset is publicly available in [1]. The dataset can be exported to a .html file or .xls. This dataset has a high amount of information including research area for each publication, authors, titles, abstracts, keywords, type of publication (which might include conference papers, book chapter, conference proceedings, patents, software, algorithms among many others), author's research institution and others. This is a complete source of information and there is information for dozens of years starting from around 1985 until 2015. The high quality and high organization of this structured data make it a good source for text mining or NLP tasks.

After exporting the data for 2014, we selected conference papers and book chapters only. Therefore, the number of publications was reduced from around 25000 to 5110 publications. For this amount of publications, we selected the titles and abstracts for our NLP tasks. In average, the titles have 11 words and the abstracts have 240 words. This amount of publications represents approximately 2800 different first authors.

Pre-processing the data is important in an information retrieval context since we are interested in reducing noise in data. Thus, by treating the text, we achieve less entropy in our models and achieve better results from automatic machine learning procedures. The pre-processing of data included the following sequence of procedures:

1. removal of whitespaces
2. removal of stopwords
3. removal of punctuation except hyphenated compounds
4. removal of numbering
5. convert every word to its lowercase version

This pre-processing was completely done with the *tm* package available for R language.

After text data pre-processing, there are 38544 terms in titles and 822888 terms in abstracts. It is visible that the titles have high sparsity with few words repeated many times and a high amount of words existing only one time. Fig. 1 representation of frequency distribution makes us conclude that it is approached by a power law distribution.

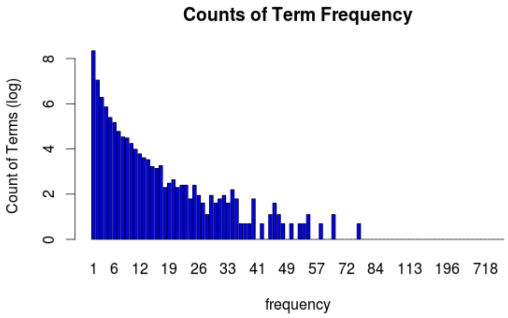

**Fig. 1.** Frequencies Counting for Titles Words

The distribution of words in abstracts is represented in Fig. 2. It is visible that the abstracts have high sparsity with few words repeated many times and a high amount of words existing only one time. Again, as previously with the titles, this representation of frequency distribution makes us conclude that it is approached by a power law distribution. This is an expected characteristic when studying text data.

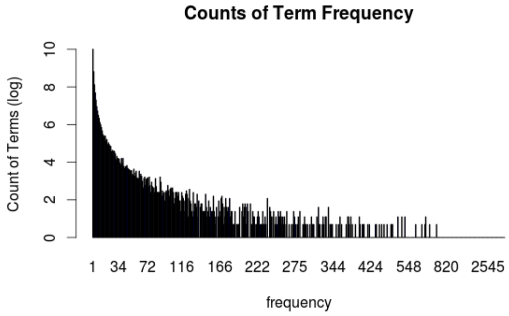

**Fig. 2.** Frequencies Counting for Abstracts Words

Both previous figures were smoothed by application of a logarithmic function to the values of frequency occurrences.

## 4 Methodology

With the described data we are interested in finding topics in the documents available in the dataset. Our approach to the data has several possibilities. Succinctly, we can study:

- Titles of publications
- Abstracts of publications
- Titles and Abstracts concatenated

Additionally, the data available enables the use of the provided authors keywords to validate all experimented methods and their results. Since we have the goal to find an affinity between authors and topics, we are interested in finding the topics each author approaches in their publications. Thus, we concatenate every author's publications in one document for each author. This is true either for titles, abstracts or titles plus abstracts. To evaluate, we proceed to group all author's keywords in one document for each author.

The procedure used to extract the topics relies on experimenting with two distinct methods, LDA and TF-IDF. For these methods we find the similarity of the extracted topics/keywords with the keywords provided manually by the authors. In the end, we have the similarity/overlapping results for every author and regarding both methods. Additionally, we extend the TF-IDF method, by finding clusters of authors in the cosine similarity matrix, obtained from TF-IDF matrix. Then, we generalize the extracted topics of the central author (the exemplar or centroid) in each cluster, for every author belonging to the individual cluster.

### 4.1 Extraction of Topics

We used two different methods to extract the topics from the data. Following some related work we opted for LDA [4] and TF-IDF. For this task, we used the R language implementation of LDA from the package *topicmodels*. The TF-IDF method was applied to the data by using the package *tm* also for R language.

**4.1.1 LDA** The intuition behind LDA is that documents have multiple topics. Furthermore, knowing that the document blends those topics would help you situate it in a, for example, collection of scientific articles. LDA is a statistical model of document collections that tries to capture this intuition. It can be described by the generative process associated to this method. This process involves randomly selecting a model assumed to be one the documents arose from.

Blei et al. [3, 4] formally define a topic to be a distribution over a fixed vocabulary. For example, the *streaming* topic has words about streaming with high probability and the *clusterization* topic has words about clusterization with high probability. Then, for each document in the collection, the authors generate the words with a process involving two stages.

1. Randomly choose a distribution over topics
2. For each word in the document
   (a) randomly choose a topic from the distribution over topics in step 1
   (b) randomly choose a word from the corresponding distribution over the vocabulary

This statistical model reflects the intuition that documents exhibit multiple topics. Each document exhibits the topics in different proportion (step 1). Additionally, each word in each document is drawn from one of the topics (step 2b), where the selected topic is chosen from the per-document distribution over topics (step 2a).

The distinguishing characteristic of LDA is all the documents in the collection share the same set of topics. Nonetheless, each document exhibits those topics in different proportion.

**4.1.2 TF-IDF** TF-IDF is typically a computationally less complex option to calculate similarity using word counts. Since this counting can be biased toward common words in the documents, some adjustments have to be done. Thus, the term counting is weighted by the inverse of the number of appearances of the same term in the documents. We can mathematically formalize this the following way:

- Let D be the set of documents, $d \in D$, a document consisting of a sequence of words (terms), and t a particular term of interest in d. Then the scaled word count based on TF-IDF is

$$tfidf(t, d) = tf(t, d) * idf(t, D) \tag{1}$$

where $tf(t, d)$ is the fraction of times the term t appears in d, and $idf(t, D)$ is the logarithm of the inverse of the proportion of documents containing the term.

**4.1.3 TF-IDF with Clustering** The similarity of documents can then be calculated by cosine similarity of the TF-IDF vectors. After calculating similarity, we can extract the clusters presented in the author's list with the *apcluster* package for R language. Finding clusters, in this case, will signify that we have similar authors grouped together. We assume these authors are similar because they approach similar topics of studies. Then, for each cluster, we find the exemplar i.e. the central author, and assume the cluster topics is represented by the exemplar topics. We validate this assumption by iterating, for each cluster's authors, and then comparing the cluster's topics with the cluster's authors keywords.

### 4.2 Stemming

For further research about the results from both models, and since we are validating the models with the author's keywords, we are interested in finding the root of the words resulting from the automatic extraction and the root version of the keywords themselves.

It is commonly described that stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected or sometimes derived words to their word stem, base or root form. Frequently, the stem is not identical to the morphological root of the word. Additionally, it is usually sufficient that related words map to the same stem, even if this stem is not, in itself, a valid root.

For example, a stemming algorithm reduces the words "streaming", "streamed", and "streamer" to the root word, "stream". Another example, the words "endue", "endued", "endues", "enduing" reduce to the stem "endu". This last case exhibits a situation where the stem is not itself a word or root. By using stemming, we expect that both models increase the similarity with the keywords. This is a result of normal use by the authors of different forms for the same stem in different textual situations like, for example, in titles, abstracts or keywords. Further developments will be presented in the following sections.

For this task, the R language was used, more specifically the *tm* package. This package provides the possibility of executing stemming of a Corpus of documents.

### 4.3 Evaluation of topics/keywords generated

To validate and compare both LDA and TF-IDF results, we use a similarity measure to compare the author's keywords and the model results for topics extraction. The following formula describes the method to find this similarity:

$$similarity = \frac{common^2}{c(w_K) * c(w_M)} \tag{2}$$

where *common* is the intersection of both groups i.e. the number of words that appear in both the results and the keywords. $c(w_K)$ and $c(w_M)$ are the total number of words the author's keywords group has and the method (LDA or TF-IDF) provides, respectively. The values for the TF-IDF method similarity are calculated with the topics extracted from each author's TF-IDF matrix row, and by selecting every term with value superior to 0. Additionaly, with LDA, we used all terms generated by the method for each author's topics.

## 5 Results

The results presented in this section were all obtained with the use of R language. The results are presented to provide a clear comparison between LDA and TF-IDF methods applied to our data. We provide a comparison of the LDA method and the TF-IDF method with and without clustering. Finally, we also provide the same experiments but applying stemming to the data in the validation process.

### 5.1 LDA vs TF-IDF

The average similarity results are presented here to provide a comparison for all methods. This average similarity is calculated with all first authors in the dataset.

**5.1.1 Without Clustering** In this section, we present a comparison between LDA and TF-IDF results for each author. Using only the titles, the average similarity for LDA is 0.176 and with TF-IDF is 0.217. For the abstracts, with LDA, the average similarity is 0.040. Additionally, with TF-IDF, the value is 0.105. Regarding the titles plus the abstracts, with the LDA method, the average similarity is 0.029. With TF-IDF, this value is 0.103.

**Table 1.** Average Similarity with LDA and TF-IDF without Clustering

| Average Similarity | | | |
|---|---|---|---|
| Topic Extraction Method | Titles | Abstracts | Titles + Abstracts |
| LDA | 0.176 | 0.040 | 0.029 |
| TF-IDF | 0.217 | 0.105 | 0.103 |

Table 1 presents results that indicate TF-IDF is better than LDA in the extraction of topics when these topics are compared with the author's keywords.

**5.1.2 With Cluster Exemplars** In this section, we provide the comparison between LDA results for each author and the TF-IDF results obtained by generalizing the cluster's exemplar and its extracted topics.

**Table 2.** Average Similarity with LDA and TF-IDF Clustering Exemplars Generalization

| Average Similarity | | | |
|---|---|---|---|
| Topic Extraction Method | Titles | Abstracts | Titles + Abstracts |
| LDA | 0.176 | 0.040 | 0.029 |
| TF-IDF | 0.066 | 0.029 | 0.029 |

Table 2 presents evidence the TF-IDF method equals LDA with more data in the studied dataset.

**5.1.3 Using Stemming - With and Without Clustering** In this section, we provide the results for LDA, TF-IDF with clustering and TF-IDF without clustering but using stemming when processing the extracted topics and also author's keywords.

**Table 3.** Average Similarity with stemming and for LDA and TF-IDF, with and without clustering

| Average Similarity with Stemming | | | |
|---|---|---|---|
| Topic Extraction Method | Titles | Abstracts | Titles + Abstracts |
| LDA | 0.206 | 0.044 | 0.033 |
| TF-IDF with Clustering | 0.076 | 0.031 | 0.032 |
| TF-IDF without Clustering | 0.256 | 0.120 | 0.118 |

Comparing table 3 results with previous tables 1 and 2, we emphasize that, by stemming the topic extraction results and also the keywords, leads to better values of average similarity. This result suggests that topic extraction of this dataset improves by using stemming. This is true for every variant of the method or dataset (titles or abstracts, or even titles plus abstracts).

### 5.2 Discussion of Results

Our results indicate that TF-IDF is the best method to automatically extract topics with our dataset. Additionally, with the titles dataset, we obtain better similarity values which indicate that both models extract more topics/keywords with less tentative provided terms. These topics have, therefore, higher similarity to the keywords provided by the authors. With abstracts or titles plus abstracts, the similarity values are lower. This happens because both models extract more terms and intersect less with the keywords provided by the authors.

Our results, by using clustering, were not conclusive to provide clear improvements in the extraction of topics when compared with the LDA method for each author. We stress that we have conducted a study with one year of data (2014). This short period of data might cause the clustering to have low density. Thus, the elements belonging to the same cluster might not have the expected similarity between them. Additionally, as the amount of data increased, for example, with titles plus abstracts the similarity decreased and is comparable to the similarity obtained with the LDA method. Thus, the exemplar or central element of the cluster might be representative of the cluster's generalization of topics if we had more data. Since we have only one year of data, we have less than 2 publications per year and by each author, which implies more variety of areas of research and, therefore, points in data clusters with higher dispersion. So, a question is, if we had more data, TF-IDF with clustering might be better than LDA?

Another interesting result we obtain from our experiments is the improvement exhibited by applying stemming to our extracted topics and keywords. This indicates that authors use different forms of words selected for keywords, and the same subject inside titles and abstracts text.

## 6 Conclusions and Future Work

This document provides an introduction to the subject of Topic Modeling. This NLP research area is an important task and can be a starting point to applications of NLP in need of documents clustering. In this work, we also do an introduction to the dataset we used in the testing of LDA and TF-IDF with and without clustering. The task to retrieve the topics with both methods from the titles and abstracts data was evaluated by using also the author's own keywords. Additionally, we also tested the methods with stemming processing experiments, to check if we obtain better topic modeling. Stemming proved to be useful when evaluating our model against the author's keywords.

As future work, we are expecting to improve experiments with more data. This might provide better results with the clustering TF-IDF method. The experimentation with LDA with clustering might also be needed to compare both methods. Additionally, we want to develop a prototype for visualization of networks of authors and their affinity regarding topics modeled from the dataset used for this task. We will eventually obtain a global matrix of authors and their affinities/similarity. As each author might be represented by their topics, we can obtain a method resulting in a graph to enable the visualization of the affinity between authors and topics. It is expected that the system will provide high intuitiveness or comprehensibility, in the knowledge or information extraction, from the visualization of results. Additionally, our goal is also enabling the system to be used with a streaming approach, and will be expected to provide scalability regarding large scale data and documents inputs.

## References

1. Published data from the r&d information system of the czech republic. http://www.isvav.cz/, 2015. Accessed: 2015-09-30.
2. Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *ICDM*, pages 3–12. IEEE Computer Society, 2008.
3. David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012.
4. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.
5. Emden R. Gansner, Yifan Hu, and Stephen C. North. Interactive visualization of streaming text data with dynamic maps. *J. Graph Algorithms Appl.*, 17(4):515–540, 2013.
6. T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.
7. David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 363–371, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.
8. Wei Li and Andrew McCallum. Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 577–584, New York, NY, USA, 2006. ACM.
9. Chin-Yew Lin and Eduard H. Hovy. Identifying topics by position. In *ANLP*, pages 283–290, 1997.
10. Michael J. Paul and Roxana Girju. Topic modeling of research fields: An interdisciplinary perspective. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov, editors, *RANLP*, pages 337–342. RANLP 2009 Organising Committee / ACL, 2009.
11. Simon Price, Peter A Flach, and Sebastian Spiegler. Subsift: a novel application of the vector space model to support the academic research process. In *WAPA*, pages 20–27, 2010.
12. R. Rubin. *Foundations of Library and Information Science.* Neal-Schuman Publishers, 2010.
13. S. Sendhilkumar, Nachiyar S N, and G. S. Mahalakshmi. Novelty detection via topic modeling in research articles.
14. Ivan Titov and Ryan T. McDonald. Modeling online reviews with multi-grain topic models. *CoRR*, abs/0801.1063, 2008.
15. Luis Trigo, Martin Vita, Rui Sarmento, and Pavel Brazdil. Retrieval, visualization and validation of affinities between documents. In *KMIS 2015 - Proceedings of the International Conference on Knowledge Management and Information Sharing, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Volume 3, Lisbon, Portugal, November 12-14, 2015*, pages 452–459, 2015.