International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Detecting Main Topics using Dictionary-based Topic Analysis

Luca Pavan
Institute of Foreign Languages, Vilnius University, Vilnius, Lithuania; Language Studies Centre, Faculty of Creative Industries, Vilnius Tech, Vilnius, Lithuania.
Corresponding Author: Luca Pavan, **E-mail**: pavan@panservice.it

| ABSTRACT
This paper describes a dictionary-based software for topic analysis written by the author. The dictionary was created manually. Many studies showed the advantages of using dictionaries to analyze texts. The software described here works in English and Italian languages, and it does not make use of probabilistic methods. In natural language processing, the use of a lexicon to reveal topics in a text is often avoided. Topics depend very much on the context. Assigning unique words to each topic does not help to check the topics in different contexts. However, the software, with a dictionary of about 5,500 topic words described in the paper, in many cases, allows the same word to fall into different topics. This approach allows one to find the main topics in a text, which corresponds to the most frequent topic words detected by the software. Advantages and disadvantages are discussed in the paper, along with examples. The software was extensively tested on large texts, such as Internet news corpora and classics of English and American literature, showing very high reliability in detecting the main topics. Analysis of topics in literary works demonstrates almost the same conclusions as were reached by critics.

| KEYWORDS
Computational linguistics, Topic analysis, English literature, American literature, Italian literature.

| ARTICLE INFORMATION
ACCEPTED: 28 November 2022 **PUBLISHED:** 04 December 2022 **DOI:** 10.32996/ijllt.2022.5.12.6

1. Introduction

In the last 20 years, probabilistic topic models have become very popular among researchers. For example, a model like LDA (Latent Dirichlet Allocation) is the main feature of many tools for the analysis of topics. LDA tries to 'reverse' the generative process of a text, revealing its hidden structure (Blei, 2012, p.79). Basically, the idea is to conceive a document as a mixture of topics (Steyvers, Griffiths, 2007, p. 4). Probabilistic topic models allow researchers to decide how many latent topics should be found in a text using a *bag-of-words* approach: the order of words is ignored, taking into account the frequency of each term (Kherwa, P., Bansal, P., 2019, pp. 2-5). This approach, however, has several limitations. LDA analysis performed by LDA often focuses only on hidden or latent topics. Also, the results can be 'noisy', showing a number of words that are not correlated with the topics. Another problem is the labelling of topics. Some automatic techniques to label topics were also proposed (Mei, Shen, Zhai, 2007). In general, many of these techniques make use of probabilistic methods. Finally, LDA performs well only by analysing long texts or corpora.

The software described in this paper makes use of a classic deterministic model using a dictionary created manually. Using a dictionary is not a new idea (Schone, P., Nelson, D., 1996, p. 295). More recently, it has become common to use both probabilistic methods and target dictionaries to perform the analysis (Watanabe, K., Zhou, Y., 2019, p. 1). The use of dictionaries is also quite common in the analysis of sentiment (Pavan, 2022).

The software proposed in this paper searches for the main topics in a text only based on the frequency of topic words. This approach can be useful for searching and classifying words according to a list of general topics. If the dictionary and the number of topics are quite large, it will be possible to obtain good results when searching for the main topics in a text. Contrary to LDA, in

this approach, one already has a labelled list of topics, each one containing a number of topic words. The software will search for words belonging to the topics and will perform some statistics, which, among other things, will provide a sorted list of topics, allowing one to know the main topics in the input text. In this approach, there are advantages and disadvantages, both of which shall be discussed in the paper.

2. Method

Topicword is a dictionary-based software for performing topic analysis in English or Italian. The software is written in C language. The dictionary used in the software was written manually by the author and included about 5,500 topic words. Each word is labeled with its corresponding topic. In total, there are 41 topics. Among others, some topics are: 'business', 'finance', 'education', 'computers', 'medicine', 'architecture', 'religion', 'plants', 'animals', 'weather', 'music', and 'accidents'. Words can be included in more than one topic. For example, the word *visibility* appears in 'weather' and 'accidents'. The software will output five text files. The first file will show the input text in lowercase with the topic labels near the topic words. The second file is the list of found topic words sorted in descending order according to their frequency (Fig. 1).

medical[medicine]	337
disease[medicine]	234
good[religion]	159
patient[medicine]	155
fever[medicine]	155

Fig. 1 – Example showing some topic words with frequency sorted in descending order

The third file shows the same list with the number of input text lines where the word was found. The fourth file, the most important, shows the list of topics and their frequency sorted in descending order. Here, it will be possible to see the main topics of the input text (Fig. 2). Finally, the fifth file shows the topics found, each topic containing the topic words in the input text.

Medicine	2356
Education	2109
Law	1118
Business	990
Religion	877
Human body	591
Food	549

Fig. 2 – Example showing several topics with topic words frequency sorted in descending order

The software described here, when outputting the fourth file, works with the frequencies of the topic words. Topic words can be ambiguous and can belong to different topics. However, the overall frequency of a topic includes ambiguous and not ambiguous words. Possibly, the main topics will end up at the top of the list, while several minor or even false topics will be down. Typically, the first topic of the list is the most important and shows the main argument of an input text. Topicword was tested with news corpora and many other books belonging to different fields of knowledge. Usually, it never fails to detect the main topics unless they fall outside the list of 41 topics.

3. Results and discussion

In this section, some examples will be discussed. Topicword was used at first to analyse Internet news corpora. A corpus of news from Harvard Dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GMFCTR), including online news in the years 2016-2017 was analysed, giving; as a result, the most popular topics (Fig. 3).

Business	56256
Politics	47436
Law	33785
Finance	29078
Education	24627
Accounting	17714
Transports	17157

Fig. 3 – The main topics from a news corpus with topic words frequency sorted in descending order

Fig. 3 demonstrates that the most popular topics are about the economy and politics. The most used topic words in the corpus are listed in Fig. 4.

President[politics]	4659
government[politics]	3273
house[architecture]	2699
state[politics]	2696
court[law]	1755
court[politics]	1755

Fig. 4 – The most popular topic words from a news corpus with frequency sorted in descending order

Analysis of other news corpora gives more or less similar results, sometimes including also "accidents" as a relevant topic. Although some topic words can be used in different contexts, statistically, it is possible to get good results relying on not ambiguous words. However, the whole topic words belonging to a topic will show the predominance or not of that topic. The frequency of topic words is not absolute because some words can be included in different topics. Therefore, several topic words can be labelled with different topics showing the same frequency, as in Fig. 4. Topicword will help only detect the main topics. If researchers need to search for subtopics, it will be necessary to update the dictionary by adding specific topic words. This can be, on one side, a disadvantage. On the other side, the software will perform a general analysis, giving the user the main structure of an input text. The software works well if the dictionary contains a large number of topic words. It also works for short texts, contrary to LDA, and it does not need to set stop words.

Another field of application of Topicword is literature. Some historical works of English and American literature were analysed with this software. All examples were taken from Project Gutenberg (https://www.gutenberg.org/). It was possible to reveal several known and unknown features in the style of the authors. For example, the software performed well in analysing English literature from the Romantic period. In analysing the *Lyrical Ballads* by William Wordsworth, the most popular topics show some typical romantic features of the author (Fig. 5).

Plants	278
Human body	261
Weather	258
Architecture	203
Religion	181
Sea	171

Fig. 5 – Main topics of the *Lyrical Ballads* by William Wordsworth sorted in descending order

Among the main topics in Wordsworth, the topics 'plants', 'sea', and 'weather' are correlated with the romantic theme of Nature (Roe, 1992). In fact, the philologists came to the same conclusions. In particular, the topic 'plants' shows Wordsworth as a 'landscape gardener' (Thompson, I. H., 2007, p. 196). The topic of 'religion' is also common in the poetry of Wordsworth (Borkowska, 2021). The topic 'human body' is also one of Wordsworth's favourites (Youngqvist, 1999, p. 152). Fig. 6 shows a cloud containing the 10 most popular topic words in Wordsworth's poetry. It is possible to see that the most used topic words in Wordsworth's poetry are 'heart', 'house', 'green', 'stone' (Fig. 6).

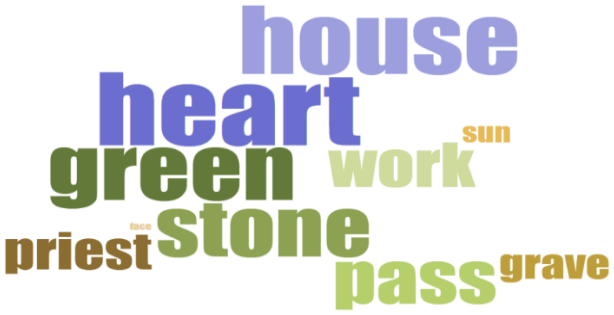

Fig. 6 – A cloud showing the most popular topic words in William Wordsworth's poetry

Topicword is able to analyse in the same manner all kinds of literature work. For example, by analysing a science fiction novel by Isaac Asimov ("Let's get together", 1957), it is possible to have a view of the main topics (Fig. 7).

Computer and technology	58
Business	49
Education	46
Politics	36
Transports	34
Human body	33

Fig. 7 – Main topics from a science fiction novel by Isaac Asimov sorted in descending order

In the same novel, the main topic words are listed in Fig. 8.

robotics[computer and technology]	33
security[transports]	19
science[education]	10
eyes[human body]	10
presidential[politics]	8
assistant[business]	8
scientists[astronomy]	7

Fig. 8 – Most popular topic words from a science fiction novel by Isaac Asimov sorted in descending order

In some cases, the context of these words is different in the text. However, by looking at the frequency, there is a high probability that the topic labels are in the right place. Another problem is false topics. Using topic words used for different topics, sometimes a false topic can appear at the bottom of the sorted list. However, the topics at the top of the list are usually correct. Topicword was created to analyse all kinds of texts except those containing very specific arguments. The dictionary can be enlarged to include any specific topic, but, in general, it is already big enough to work well in most cases. Finally, the software provided good results also by analysing literary works in the Italian language (from the website www.liberliber.it). For example, in analysing a Futurist work by Filippo Tommaso Marinetti ("La cucina futurista", 1932), the main topics are listed in Fig. 9.

Food	164
Food nutrition	84
Human body	59
Education	55
Plants	51
architecture	43

Fig. 9 – Main topics from work by Filippo Tommaso Marinetti sorted in descending order

4. Conclusions

The paper described Topicword - a software for topic analysis, based on a dictionary of about 5,500 entries which was written manually by the author. Topicword is able to detect the main topics of an input text file, performing some statistics according to the frequency of topic words. It was found that repeating some words in the dictionary with a different topic label gives very good results in topic detection. The software showed a high grade of reliability in all tests. The software can also be used to analyse corpora or literary works. In the case of the literature, the results are similar to the opinion of philologists. The software also works well with corpora, showing the main arguments of the collected texts. Unlike other probabilistic methods, Topicword is unable to detect subtopics. However, the dictionary could be enlarged to include specific topics. Further studies could include the possibility of creating larger dictionaries in different languages. Another possibility is to use target dictionaries for more specific purposes. The software could be used, among others, in the field of business (to analyse documents) or in the field of education (to analyse literary works and corpora).

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors and the reviewers.

References

[1] Blei, D. (2012). Probabilistic topic models. Communications of the ACM. *55*(4), 77-84.

[2] Borkowska, E. (2021). *The Presence of God in the Works of William Wordsworth*. New York: Routledge.

[3] Kherwa, P. and Bansal, P. (2019). Topic Modeling: A Comprehensive Review. *EAI*. Endorsement Transaction on Scalable Information Systems.

[4] Mei, Q., Shen, X and Zhai, C. (2007). Automatic labeling of multinomial topic models. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*. 490–499.

[5] Pavan, L. (2022). A Survey of Some Italian Literature Works using Sentiment Analysis. *International Journal of Linguistics, Literature and Translation*, *5*(1), 117-121.

[6] Roe, N. (1992). *The Politics of Nature. Wordsworth and Some Contemporaries*. London: Macmillan.

[7] Schone, P and Nelson, D. (1996). A dictionary-based method for determining topics in text and transcribed speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, p. 295.

[8] Steyvers, M., Griffiths, T. (2007). Probabilistic topic models. In T. K. Landauer, D. (eds.), *Handbook of latent semantic analysis*: 427–448. Mahwah (NJ): Lawrence Erlbaum Associates Publishers.

[9] Tagore, P. (2000). Keats in an Age of Consumption: The 'Ode to a Nightingale'. *Keats-Shelley Journal*, *49*, 67–84.

[10] Thompson, I. H. (2007). William Wordsworth, Landscape Architect. *The Wordsworth Circle*, *38*(4), 196–203.

[11] Watanabe, K. and Zhou, Y. (2019). Making a topic dictionary for semi-supervised classification of the UN speeches. *Quantitative Text Analysis Dublin Workshop*. 1-26.

[12] Youngquist, P. (1999). Lyrical bodies: Wordsworth's physiological aesthetics. *European Romantic Review*, *10*(1-4), 152-162.