Topic modeling in humanitarian researches

Andrey Maksimov

May 2024

Abstract

The article shows the possibilities of topic modeling as a tool to support humanitarian research related to text analysis. The two most well-known models are compared and some aspects of the use of this toolkit in semantic text research are considered.
The project code link: https://github.com/NLPforHumanitarianResearches/TopicModels

1 Introduction

A significant part of historiographical, sociological, cultural, bibliographic and many other studies in the humanitarian field of knowledge is more or less related to the analysis of textual documents. One of the widespread methods of supporting the implementation of such studies is the content analysis, which allows you to determine the frequency of occurrence of certain words (terms) in the texts under study, which makes it possible to obtain some semantic characteristics of the text in question [Neuendorf, 2017].

In order to expand the tools to support such research, topic modeling methods can be used, allowing each document of the studied collection of texts to be presented as a set of topics contained in it, each of which is characterized by a certain set of words (terms). Obviously, such a representation is a more semantically meaningful characteristic of the texts under study than the results of the content analysis. At the same time, the possibility of an unambiguous semantic interpretation of the thematic sets of words identified by algorithms is of particular importance, which largely determines the quality of the constructed thematic model. Sets of words satisfying this condition can be called "good topics".

The main task of this work is to compare the capabilities of the two most well-known topic modeling algorithms in order to obtain "good topics" and determine the sequence of actions necessary for this with relatively modest computing resources.

1.1 Team

Andrey Maksimov prepared this document.

2 Related Work

Topic modeling is one of the modern technologies of Natural language Processing (NLP), which has been actively developing since the late 90s. This technology does not require pre-markup of data and refers to unsupervised machine learning. Topic modeling is represented by a fairly extensive range of models and algorithms, including latent semantic indexing (LSI), latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), additive regularization algorithms (ARTM). Also of note is the BERtopic model, implemented on the basis of the transformer architecture and representing a pre-trained topic modeling technology that uses BERT and c-TF-IDF to create dense clusters that allow you to interpret topics while preserving important words in the descriptions of topics [Vorontsov, 2021] [Vayansky, 2020] [Krishnan, 2023].

Transformer-based models, which are currently the flagship direction of the development of natural language processing technologies, are characterized by very high demands on computing resources, which significantly narrows the circle of potential users. This work is devoted to the consideration of less resource-intensive models that allow the use of NLP technologies to a wider range of researchers.

3 Models Description

The main idea of topic modeling is as follows:

- In the corpus of $\boldsymbol{M}$ documents, one is selected $\boldsymbol{K}$ topics that are supposed to be found in this corpus.
- The topic model of the corpus is a vectorized set of documents, each of which is presented as a combination of $\boldsymbol{K}$ topics.
- The algorithm determines weighted relationships between documents and topics, as well as between topics and words.

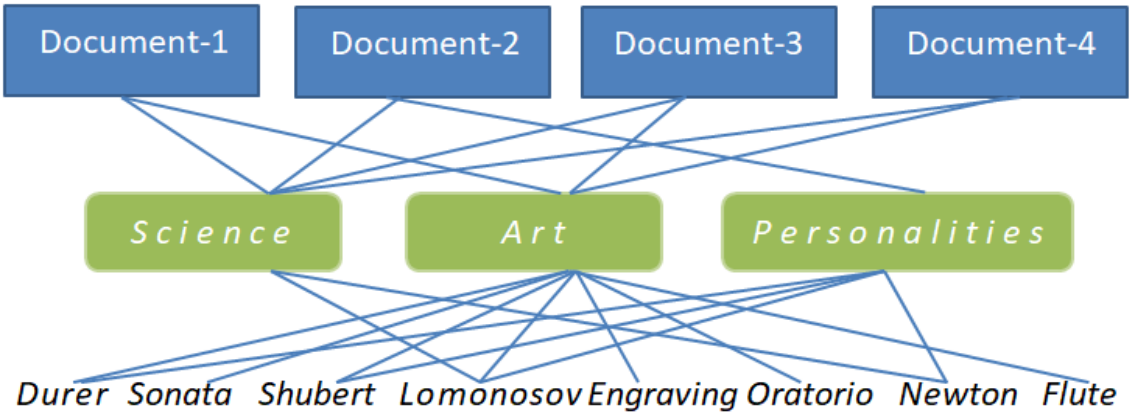

Figure 1: Illustration of the basic idea of topic modeling.

On Fig. 1 schematically shows an example of a thematic model for a corpus of 4 documents, in which 3 topics are highlighted.

For the semantic analysis of the considered collection of text documents, two of the most well–known topic modeling algorithms were used - Latent Dirichlet Allocation (LDA) and Non-negative matrix factorization (NMF) [Blei, 2012] [Krishnan, 2023].

The Latent Dirichlet Allocation (LDA) is a generative probabilistic model for highlighting topics in document collections. This model is based on the assumption that each document can be presented as a mixture of different topics, and each topic is associated with a probabilistic distribution of words.

The LDA-model Generative Process assume the following actions:

For each topic $k \in \{1, ..., K\}$:

$\Phi_k \sim \text{Dir}(\beta)$ [draw distribution over words]

For each document $m \in \{1, ..., M\}$

$\Theta_m \sim \text{Dir}(\alpha)$ [draw distribution over topics]

For each word $n \in \{1, ..., N_m\}$

$z_{mn} \sim \text{Mult}(1, \Theta_m)$ [draw topic assignment]

$x_{mn} \sim \Phi z_{mi}$ [draw word]

where α and β - hyperparameters of distributions [Liu, 2022].

Non-negative matrix factorization (NMF) is a data analysis method based on the decomposition of a large non-negative matrix into two non-negative matrices of smaller dimension. The mathematical basis of this method is the apparatus of linear algebra. By decomposing the original matrix "Words-Documents", this method allows you to identify hidden topics in a collection of documents and determine their connections with documents.

The essence of the NMF algorithm is as follows:

- there is an initial matrix **V** of dimension **m×n**, where m is the number of words and **n** is the number of documents;
- the NMF algorithm finds two matrices **W** and **H** of dimensions **m×k** and **k×n**, where **k** is the number of topics such that **V ≈ WH**;
- matrix **W** — "Words-Topics" — shows which words characterize each topic;
- matrix **H** — "Topics-Documents" — shows how much each topic is present in each document.

The NMF algorithm minimizes the difference between the original matrix **V** and the approximate matrix **WH**, taking into account the restriction on the non-negativity of all elements of the matrices **W** and **H**. This allows you to get interpretable results, since the elements of the matrices represent a quantitative assessment of the importance of words and topics in a collection of documents.

4 Dataset

The experimental data set is a collection of Russian-language texts on the history of culture, collected by the author of this work from open Internet sources. The collection includes 1300 texts presented in csv file format with a total volume of about 16 MB.

On Fig. 2 shows a diagram of the distribution of texts in the corpus under study, according to the number of characters contained in them.

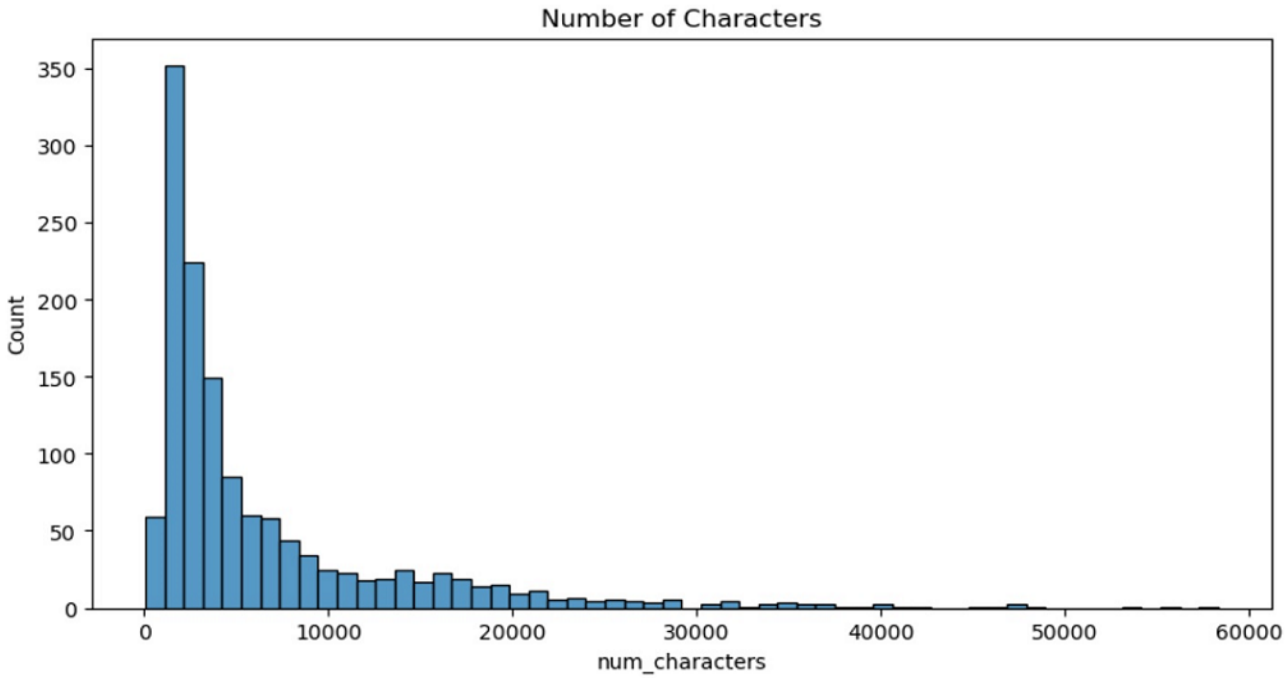

Figure 2: The diagram of the distribution of texts by the number of characters.

5 Experiments

The main purpose of the experiments was to obtain the best set of topics characterizing (reflecting) the semantics of the corpus of texts under study. Three categories of parameters served as the variable base of the experiments, which were used as:

- stop words;
- filtering words by grammatical categories, i.e. performing text analysis not for all categories of words, but only for words of selected categories;
- the number of highlighted topics.

The number of topics varied from 2 to 30. The standard Russian-language set of NLTK library stop words was used as a list of stop words, supplemented by some list of words, the expediency of excluding which was revealed during the experiments. Word filtering by grammatical categories was performed for the following variants:

- nouns only;
- nouns + verbs;
- nouns + adjectives;
- nouns + verbs + adverbials;
- nouns + adjectives + adverbials;
- all categories (i.e. without filtering).

Due to the for lack of suitable quantitative metrics for evaluating the results of topic modeling, the results were evaluated in terms of quality, i.e. from the point of view of the possibility of an adequate interpretation of the resulting sets of words characterizing the identified topics.

6 Results

The NMF model showed the best results for 15 topics with the filtering parameter "nouns + adjectives". The importance of words for the relevant texts in the document collection was taken into account based on the TF-IDF procedure. Table 1 shows the terms highlighted by this algorithm.

Table 1. NMF-model Topics.

Topics	Terms
Topic 1	проект, архитектор, архитектурный, работа, здание, дом, творческий, архитектура, строительство
Topic 2	советский, немецкий, танковый, фронт, армия, войско, дивизия, оборона, удар
Topic 3	наука, природа, научный, философский, мир, философия, понятие, закон, идея
Topic 4	керамика, изделие, керамический, сосуд, глина, гончарный, фарфор, производство, глазурь
Topic 5	педагогический, школа, учебный, образование, заведение, обучение, ребенок, университет, русский
Topic 6	храм, древний, стена, римский, пирамида, город, церковь, памятник, сооружение
Topic 7	готический, собор, романский, художественный, свод, средневековый, готика, архитектурный, неф
Topic 8	войско, армия, война, отряд, сражение, поражение, бой, союзник, римский
Topic 9	почтовый, почта, письмо, гонец, доставка, марка, перевозка, услуга, корреспонденция
Topic 10	стул, спинка, сидение, кресло, мебель, ножка, форма, стиль, мебельный
Topic 11	мачта, парус, судно, парусный, стеньга, корабль, косой, прямой, бушприт
Topic 12	сыр, блюдо, напиток, кулинарный, продукт, русский, кухня, рецепт, вкус
Topic 13	сад, парк, дворец, замок, фонтан, резиденция, парковый, водоем, аллея
Topic 14	фасад, архитектура, этаж, сооружение, здание, купол, окно, архитектурный, ордер
Topic 15	башня, мост, метр, высота, стена, строительство, турист, город, сооружение

For all sets of words identified by the algorithm as semantic features of a particular topic, an unambiguous subject interpretation is feasible, which allows each topic to be matched with its meaningful name. So, topic 1 is *"Architecture"*, topic 2 is *" Military history"*, topic 3 is *"Science"*, etc.

7 Conclusion

The use of topic modeling methods expands the possibilities of humanitarian research related to the analysis of text collections. As the results of the experiments have shown, it is advisable to use non-negative matrix factorization models as a basic tool for supporting text analysis. *NMF-algorithms* provide a fairly high quality of semantic analysis, and, unlike large language models, are characterized by very modest requirements for the hardware resources used, which makes it possible for a wide range of humanitarian researchers to use this toolkit.

References

[Neuendorf, 2017] Neuendorf K. A. The content analysis guidebook. 2Nd ed. Thousand Oaks, CA: Sage. (2017)

[Vorontsov, 2021] Vorontsov K.V. Probabilistic thematic modeling: theory, models, algorithm and the BigARTM project. / Moscow: MIPT, Federal Research Center "Informatics and Management". – 2021.

[Liu, 2022] Qun Liu, Valentin Malykh. Natural Language Processing. Lecture 11: Topic Modeling. / Huawei Noah’s Ark Lab, A course delivered at KFU, Kazan (2022)

[Blei, 2012] Blei D. M. Probabilistic topic models // Communications of the ACM. — 2012. — Vol. 55, no. 4. — Pp. 77–84.

[Krishnan, 2023] Anusuya Krishnan. Exploring the Power of Topic Modeling Techniques in Analyzing Customer Reviews: A Comparative Analysis. (2023), arXiv:2308.11520

[Vayansky, 2020] I. Vayansky and S.A. Kumar, A review of topic modeling methods, Information Systems, 94, p.101582, 2020.