С.М. Сайберт
Национальный исследовательский
Томский политехнический университет

Topic modeling and text classification

Natural language processing (NLP) is a part of machine learning and deep learning that works with texts. Topic modeling is one of NLP methods that represents a document collection. This progressive way of text processing can be used by linguists in their work. This article contains an example of topic modeling implementation. The most popular methods were applied to extract 5 topics from document collection.

Key words: topic modeling; LDA; LSA; NMF; TF-IDF; texts;

With invention of computers a lot of scientific fields have found new ways to solve their problems, studying became easier in many ways. Linguistic is not an exception. Natural language processing (NLP) is one of new possibilities that linguists can imply in their work. NLP is a combination of informatics and linguistics. It uses machine learning and neural networks to solve problems.

Topic modeling is one of the problems that NLP works with. It is a method that constructs a model of a document collection. That model represents which topic every document has. Any text can be represented as a vector of words, such vector has an excessively large dimension. Topic modeling allows to decrease the number of dimensions, it also allows synonymy and polysemy [3]. Topic models transform documents into a number of topics, those representations can be used in document classification [6].

The main purpose of topic modeling is highlighting several topics that produced a certain document collection. This task is reduced to searching an approximated product of two matrixes with less dimensionality, these are term-document and term-topic matrixes [2].

One of the first methods of topic modeling is latent semantic analysis (LSA). The main idea is to project high dimensioned vectors to a latent semantic space (approximated lower dimensional space). The projection must be linear and based on singular value decomposition (SVD) of the matrix. This method does not allow one document to have more than one topic. The probabilistic latent semantic analysis [11] (PLSA) has been developed to overcome this problem, the task is to maximize log-likelihood using EM-algorithm.

The most popular topic modeling method is Latent Dirichlet allocation (LDA) [9, С. 993]. The comparison of LDA and LSA is discussed in [5], the author notes that LDA is better than LSA in research with short texts.

The alternative of biasing models is additive regularization of topic models (ARTM). It is a general approach to the combination of topic models, in which not only log-likelihood but also other features are maximized. These

features are called regularizers. ARTM is based on PLSA because this model does not have any regularizers [14]. ARTM is used in cases when a task requires combined models and biasing models become too complicated.

Matrix factorization technics can also be valuable for topic modeling. The most suitable method is non-negative matrix factorization (NMF) [12]. The main difference of this method is restriction on resulting matrixes. They cannot contain negative values. In [15] the deep NMF was proposed, which combines deep learning and NMF. The authors of [12] in their work compared 5 different NMF algorithms and made a conclusion that Brunet-algorithm [10] is the most affective. Semantic-assisted NMF (SeaNMF) was proposed in [13]. This model uses semantic relationships between words and context for learning. The model was proposed for topic modeling of short texts (because such texts do not have enough information about mutual occurrence of words).

It can be necessary to present topics for some tasks hierarchically – from more general topics to more specific ones. In such cases hierarchy topic models can be implied [4]. In [8] LDA hierarchy was suggested (hLDA). It is based on nested Chinese restaurant process (nCRP). CRP (generates a distribution of N objects over an unlimited number of partitions), can be used in cases when the resultant number of topics is unknown and nCRP – in cases when the number of hierarchy layers is unknown.

A lot of new models are based on semantic relationships between words, that extremely depends on the document collection's language. Most of such works are done on English texts collections. Exactly semantic based models show the best results.

There are several different methods to evaluate topic model, they can be divided into two categories: expert evaluation and quantitative indicators (such as coherence and perplexity) [1, C. 5]. Expert evaluation means methods that involve human assessment of quality, for example, it can be done in tabular form by comparing top 10 words of each topic. In 2012 scientists from Stanford University developed a more complex approach «Termite» for tabular visualization of term-topic matrix.

In this work a collection of scientific text's annotations was used. The corpus contained about 80 documents on the topic of chemistry.

Standard text preparation. Texts were reduced to lower case, tokenization and lemmatization were implied; stop-words, words that appear in less than three texts, words that are shorter than four symbols were removed.

TF-IDF. Texts were transformed into digital TF-IDF vectors. The result was vector, the number of columns of which was the number of unique words in the whole corpus, the number of rows equals the number of documents, and each element means how important the word is for the document (text). Models can work with such digital vectors.

3. Topic modeling methods.

In this work four different methods are presented. These are LDA, NMF, LSA and SeaNMF. The PMI score was used to evaluate model quality (table).

Table

PMI scores for different topic models

model	n_topics	PMI score
LDA	5	1.08
NMF (forbenius)	5	0.58
NMF (kulbak-leibler)	5	0.37
LSA	5	0.61
SeaNMF	5	0.24

LDA showed the best result with the coherent score. But the model extracted one topic that contains the most popular words in the whole corpus, that means the result is not valid. The distribution of topics is represented by 5 most important words (Fig. 1).

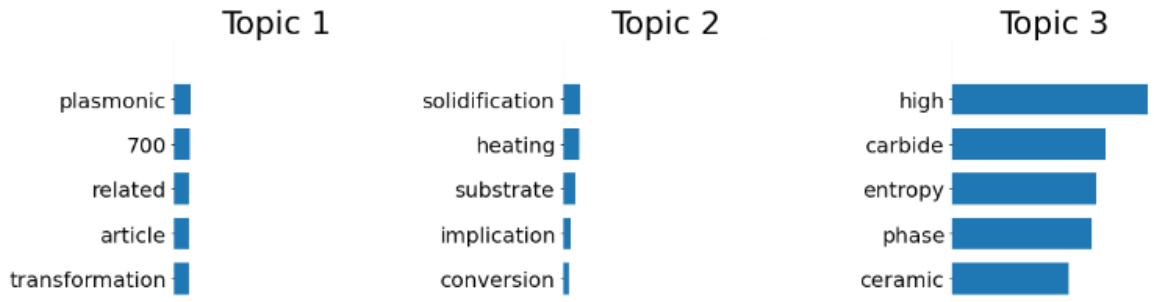

Fig. 1. LDA topic distribution

NMF (Frobenius norm was minimized). This method performed best. The words in different topics are evenly distributed (Fig. 2). The texts with dimension reduction are presented (Fig. 3), the color represents which topic each text is related to. We can see that topics perfectly separated to clusters.

In this work an example of topic modeling was shown. Just the most popular methods were used, as was shown in the introduction there are a lot more. Every case is unique, and the researcher should choose the method that suits him best.

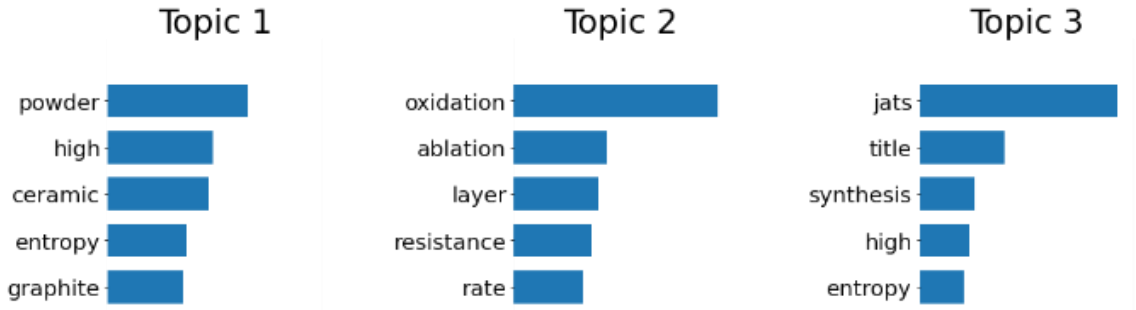

Fig. 2. NMF topic distribution

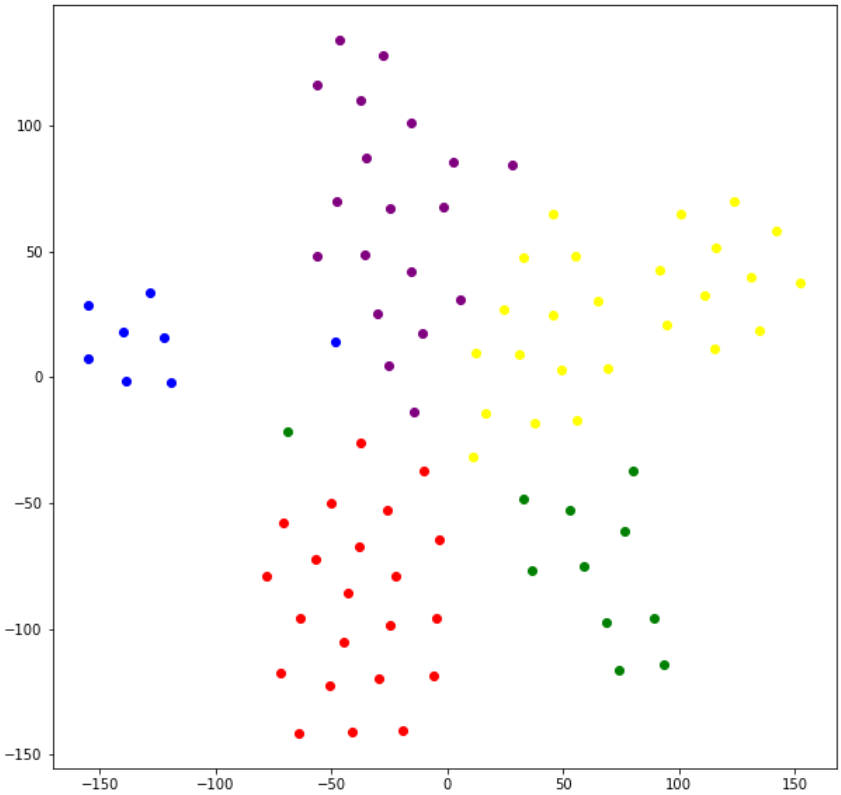

Fig. 3. Corpus in 2 dimensions

The work resulted in the division of the corpus into five topics. Further expert evaluation is required for more accurate result analysis.

Литература

1. Булатов, В.Г. Методы оценивания качества и многокритериальной оптимизации тематических моделей в библиотеке TopicNet / В.Г. Булатов // 2020. – 147 с.

2. Воронцов, К.В. Аддитивная регуляризация тематических моделей / К.В. Воронцов // Doklady Akademii Nauk. – 2013. – № 456. – С. 88–89.

3. Коршунов, А. Тематическое моделирование текстов на естественном языке / А. Коршунов, А. Гомзин // Труды ИСП РАН. – 2012. –

URL: https://cyberleninka.ru/article/n/tematicheskoe-modelirovanie-tekstov-na-estestvennom-yazyke (дата обращения: 17.12.2022). – Текст : электронный.

4. Цыганова, С.В. Построение иерархических тематических моделей коллекции документов / С.В. Цыганова, В.В. Стрижов // Прикладная информатика. – №1 (43). – 2013. – URL: https://cyberleninka.ru/article/n/postroenie-ierarhicheskih-tematicheskih-modeley-kollektsii-dokumentov (дата обращения: 19.01.2023).

5. Чижик, А.В. Исследование динамики общественного настроения в социальных сетях с использованием методов тематического моделирования / А.В. Чижик // International Journal of Open Information Technologies. – 2021. – №12. – URL: https://cyberleninka.ru/article/n/issledovanie-dinamiki-obschestvennogo-nastroeniya-v-sotsialnyh-setyah-s-ispolzovaniem-metodov-tematicheskogo-modelirovaniya (дата обращения: 22.12.2022).

6. Abdelrazek, A. Topic modeling algorithms and applications: A survey / A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat [и др.] // Information Systems. – 2023. – № 112. – URL: https://www.sciencedirect.com/science/article/abs/pii/S0306437922001090 (дата обращения: 11.03.2023)

7. Bastani, K. Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints / K. Bastani, H. Namavari, J. Shaffer // 2018. – URL: https://www.researchgate.net/publication/326505884_Latent_Dirichlet_Allocation_LDA_for_Topic_Modeling_of_the_CFPB_Consumer_Complaints (дата обращения: 11.03.2023).

8. Blei, D.M. Hierarchical Topic Models and the Nested Chinese Restaurant Process / D. Blei, T. Griffiths, M. Jordan // Advances in Neural Information Processing Systems. – 2004. – №16. – URL: https://www.researchgate.net/publication/2873720_Hierarchical_Topic_Models_and_the_Nested_Chinese_Restaurant_Process (дата обращения: 11.03.2023).

9. Blei, D.M. Latent Dirichlet allocation / D.M. Blei, A.Y. Ng, M.I. Jordan // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 993–1022.

10. Brunet, J.-P. Metagenes and molecular pattern discovery using matrix factorization / J.-P. Brunet // Proceedings of the National Academy of Sciences of the United States of America. – PNAS. – Vol. 101. – 2004. – URL: https://pubmed.ncbi.nlm.nih.gov/15016911/ (дата обращения: 11.03.2023).

11. Hofmann, T. Probabilistic Latent Semantic Analysis / T. Hofmann // UAI. – 1999. – С. 289–296.

12. Lee, D. Learning the Parts of Objects by Non-Negative Matrix Factorization / D. Lee, H. Seung // Nature. – №401 – 1999. – URL: https://www.researchgate.net/publication/12752937_Learning_the_Parts_of_Objects_by_Non-Negative_Matrix_Factorization (дата обращения: 11.03.2023).

13. Tian, S. Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations / S. Tian, K. Kang, J, Choo [и др.] // Proceedings of the 2018 World Wide Web Conference. – 2018. – URL: https://www.researchgate.net/publication/324516014_Short-Text_Topic_Modeling_via_Non-negative_Matrix_Factorization_Enriched_with_Local_Word-Context_Correlations (дата обращения: 11.03.2023).

14. Vorontsov, K. Additive Regularization for Topic Models of Text Collections / K. Vorontsov // Doklady Mathematics. – №89. – 2013. – URL: https://www.researchgate.net/publication/272616548_Additive_Regularization_for_Topic_Models_of_Text_Collections (дата обращения: 11.03.2023).

15. Wang, J. Deep NMF Topic Modeling / J. Wang, X. Zhang // ArXiv. – 2021. – № abs/2102.12998. – URL: https://www.researchgate.net/publication/349620553_Deep_NMF_Topic_Modeling (дата обращения: 11.03.2023).

Науч. рук.: Смирнова У.А., ст. преп.

А.В. Самарин, О.А. Щербатенко
Старооскольский филиал Белгородского государственного национального исследовательского университета

К вопросу о переводе английского молодежного сленга

Статья посвящена английскому сленгу молодежи, как способу их самовыражения, а также выражения их отношения к миру и окружающим людям. Используя сленг в своей речи молодые люди делают свою речь ярче, короче и экспрессивнее, что часто создает проблемы в переводе. Это объясняется лингвокультурными различиями в языковой среде русских и англичан (американцев).

Ключевые слова: перевод; сленг; молодежь; англо-американская языковая среда; студенты; школьники.

Термин «перевод» многозначен, и у него есть два терминологических значения, которые нас интересуют. Первое из них определяет мыслительную деятельность, процесс передачи содержания, выраженного на одном языке средствами другого языка. Второе называет результат этого процесса – письменный [3, с. 117].

А.В. Федоров предлагает следующее определение: «Слово «перевод» является общеизвестным и общепонятым, но и оно, как обозначение особого вида человеческой деятельности и ее результата, требует уточнения и терминологического определения. Оно обозначает следующее: