DOI 10.31891/2307-5732-2022-307-2-25-32
УДК 004.048

NATALIYA BOYKO
Lviv Polytechnic National University
https://orcid.org/0000-0002-6962-9363
e-mail: Nataliya.i.boyko@lpnu.ua
OLEKSANDR PETROVSKYI
Lviv Polytechnic National University
https://orcid.org/0000-0002-5729-544X
e-mail: oleksandr.petrovskyi.knm.2018@lpnu.ua

METHODS OF CLASSIFICATION OF MACHINE LEARNING FOR CONSTRUCTION OF MATHEMATICAL MODELS ON MULTIMODAL DATA

This article is dedicated to topic modeling as an unsupervised machine learning technique. It is analyzed how it seems possible to determine the topics of documents in order to categorize them further with the help of topic modeling methods. Such methods as latent semantic analysis, probabilistic latent semantic analysis and latent Dirichlet allocation are considered. An approach that allows the construction of effective topic models of text document collections in Ukrainian and other synthetic languages based on peculiarities of this linguistic language type is proposed, and its main stages are described. The proposed approach consists of a custom input data preprocessing pipeline, which covers file loading, text extraction, removal of improper symbols, tokenization, removal of stop-words, stemming of each token and a newly introduced model pruning stage, which makes any of the modern topic modeling methods applicable for synthetic language topic modeling. The approach was implemented in Python programming language and used to obtain the topic model of the collection of Ukrainian-language scientific publications on civic identity and related topics. An expert in political psychology, who studies the phenomenon of civic identity, was involved in the research for the topic model quality evaluation. As a result of expert evaluation of the topics singled out during the modeling, it was proposed to clarify the formulation of cluster names based on the semantics of the sets of words that form them. In general, according to the expert, the topics singled out represent the concept of the civic identity of an individual and will allow researchers to simplify the work with literature sources on this issue when used to categorize documents. This demonstrates the efficiency of the proposed approach.

Keywords: topic modeling, natural language processing, text preprocessing, latent Dirichlet allocation, latent semantic analysis, pachinko allocation, synthetic language.

БОЙКО Н. І., ПЕТРОВСЬКИЙ О. С.
Національний університет “Львівська політехніка”

МЕТОДИ КЛАСИФІКАЦІЇ МАШИННОГО НАВЧАННЯ ДЛЯ ПОБУДОВИ МАТЕМАТИЧНИХ МОДЕЛЕЙ НА МУЛЬТИМОДАЛЬНИХ ДАНИХ

Стаття присвячена тематичному моделюванню як техніці машинного навчання без вчителя. Аналізується можливість визначення тем текстових документів методами тематичного моделювання з метою їх подальшої категоризації. Розглядаються такі методи, як латентно-семантичний аналіз, ймовірнісний латентно-семантичний аналіз та латентне розміщення Діріхле. Запропоновано підхід, який робить можливим ефективну побудову тематичних моделей колекцій текстових документів українською та іншими синтетичними мовами, заснований на особливостях мов цього лінгвістичного типу, та описано його головні етапи. Авторський підхід полягає у особливому конвеєрі попередньої обробки вхідних даних, що охоплює завантаження файлів, видобування тексту, видалення зайвих символів, токенізацію, видалення стоп-слів, стеммінг кожного токену, і нововведений етап прунінгу, що разом дозволяє застосовувати будь-які сучасні методи тематичного моделювання для колекцій документів синтетичними мовами. Описаний підхід був реалізований мовою Python і використаний для побудови тематичної моделі колекції україномовних наукових публікацій з проблематики громадянської ідентичності та суміжних тем. Експерт з політичної психології, який вивчає феномен громадянської ідентичності, був залучений до дослідження за темою оцінки якості моделі. У результаті експертної оцінки виділених під час побудови моделі тем було запропоновано уточнити формулювання назв кластерів на основі семантики наборів слів, що їх утворюють. Загалом, на думку експерта, виділені теми відображають поняття громадянської ідентичності особистості та дозволять дослідникам спростити роботу з літературними джерелами з цього питання при категоризації документів. Це свідчить про ефективність запропонованого підходу.

Ключові слова: тематичне моделювання, обробка природніх мов, попередня обробка тексту, латентне розміщення Дірихле, латентно-семантичний аналіз, розміщення пачінко, синтетична мова.

Introduction

The study's relevance lies in the fact that in the conditions of growing informatization of modern society, knowledge is continuously produced in a single information space, which leads to a rapid increase in the amount of data, mostly poorly structured or unstructured. This fact dramatically complicates its processing and usage. Also, modern human faces problems related to duplication, inconsistency and distortion of information due to the absence of an implemented verification mechanism and difficulties in categorization and systematization of data [1].

However, the main problem remains that computers cannot understand the information they operate, as people understand it, distinguish what is primary and what is secondary, identify topics, ideas, and relevant concepts. Thus, the efficiency and quality of the search leave much to be desired. To some extent, this problem can be solved with the help of topic modeling as a means of building a model, which makes it possible to determine the topics of documents for their further categorization. The topic model provides short descriptions of documents and words that can be used to efficiently process extensive collections of documents while maintaining meaningful

statistical relationships beneficial for basic tasks such as classification, novelty detection, generalization, and similarity and relevance of judgment. The urgency of the problems of search and systematization of information in the modern world led to the choice of topic modeling as the theme of this research.

The aim of the research is theoretical substantiation, development and empirical verification of the approach to constructing effective topic models of a collection of text documents in Ukrainian and other synthetic languages.

The object of the research is topic modeling as a way to build a model of a collection of text documents.

Objectives of the study:

- theoretical analysis of literature sources on the issue of topic modeling;
- development of the approach for the construction of effective topic models for the Ukrainian and other synthetic languages;
- software product development using the LDA method of topic modeling;
- construction of a topic model of the collection of Ukrainian-language scientific publications on the issue of civic identity and related topics;
- assessment of the obtained model quality.

Literature review

Topic modeling belongs to unsupervised machine learning techniques. Its purpose is to create a model based on a collection (corpus) of text documents, with which it is possible to determine what topic the particular document belongs to and which words are associated with a particular topic.

Topic modeling can be considered a simultaneous clustering of documents and words on a single set of clusters, called topics. In terms of cluster analysis, a topic is the result of bi-clustering, i.e., simultaneous clustering of words and documents according to their semantic proximity. Fuzzy clustering is usually performed, so one document can simultaneously belong to several topics to varying degrees. Thus, a concise semantic description of a word or document is a probabilistic distribution on a set of topics. The process of finding these distributions is called topic modeling [2; 3].

As a rule, the number of topics in documents is less than the number of its words. Therefore, hidden (latent) variables in the form of topics allow representing the document as a vector in the space of encountered clusters instead of words, and as a result, the document has fewer components, which allows faster and more efficient processing. Thus, topic modeling is closely related to another class of problems known as dimensionality reduction [4].

In general, topic modeling makes it possible to get brief descriptions of the elements of the corpus, which allow effective processing of extensive document collections while maintaining meaningful statistical relationships that are useful for basic tasks such as classification, novelty detection, generalizations, as well as the similarity and relevance of the judgment.

Because of this, topic models are used to identify trends in scientific publications or news streams, classify and categorize image documents and video streams, retrieve information (including multilingual), tag web pages, detect spam, etc. [3].

One of the main applications of topic models is information retrieval. Search engines represent documents as word frequency vectors. The search of documents via short queries is implemented by finding the vectors in which query words often meet. The topic model allows the exact mechanism to be used for searching documents of similar subjects from the whole document or a long text fragment. In this case, the documents are represented as vectors of frequencies of topics, not individual words. Topic frequency vectors can also contain terms, authors, years of publication, institutes, conferences, journals, sites, etc., allowing to specify any object or a set of objects as a query to find objects of the same or different type having similar thematics [5].

Topic models can be also used for automatic image annotation [6], propaganda identification [7], meeting topic detection [8], topic-based evaluation for conversational bots [9], multi-grain SMS spam filtering [10], online reputation monitoring [11], user profiling from drug reviews [12], intelligent interruption systems [13], exploring molecular data sets [14], exploratory literature review for management research [15].

Materials and methods

Latent semantic analysis (LSA), described and patented in 1988 by S. Deerwester, S. Dumais, G. Furnas, R. Harshman, T. Landauer, K. Lochbaum, L. Streeter, was one of the first methods of topic modeling. The main idea of LSA is to divide the input term-document matrix into two separate matrices: document-topic and topic-term. Also, it is assumed that:

1. There is a finite set of topics, and discrete distribution of topics generates the collection, terms (words) and documents; terms and documents are explicit variables, and topics are hidden (i.e., latent).

2. The probability distribution of terms depends only on the topic, not on the document.

3. It is enough to know which terms occur in which documents to identify the subject, and neither the order of terms in documents (the "bag of words" hypothesis) nor the order of documents in the collection (the "bag of documents" hypothesis) is essential. In other words, it is assumed that the subject of the document can be determined even after a random permutation of words in it, although for a human such text loses its meaning [16].

LSA was first used to automatically index texts, detect the semantic structure of the text, and retrieve artificial, computer-generated documents. Also, this method was later used quite successfully to represent

knowledge bases and build cognitive models. However, the method of latent semantic analysis has a fundamental drawback: its probabilistic model does not correspond to reality since one of the necessary conditions for using the method is the normal distribution of words and documents, while in reality, the Poisson distribution is observed. Therefore, latent semantic analysis cannot be considered a method suitable for building a reliable topic model of a collection of text documents.

In 1999, Thomas Hofmann introduced the probabilistic latent semantic analysis (pLSA) method, which was based on using the Expectation-Maximization algorithm, an iterative method for determining the degree of similarity of parameters in probabilistic models depending on number of hidden variables. In comparison with the standard latent semantic analysis, which originates from linear algebra and aims to reduce the dimensionality of the input data, usually by singular value decomposition (SVD), the probabilistic latent semantic analysis is based on a mixed decomposition based on the model of hidden classes. This makes the pLSA method much better than LSA in the context of topic modeling of a text document collection. However, further research revealed significant shortcomings, namely: tendency to overfit and impossibility to expand the collection of documents.

The next stage in the development of topic modeling was latent Dirichlet allocation (LDA). The method was first proposed in 2000 by J.K. Pritchard, M. Stevens, and P. Donnelly in the context of populational genetics, but three years later, the work of David Blay, Andrew In, and Michael Jordan was published on the use of latent Dirichlet allocation for topic model construction [17].

In LDA, as in pLSA, each document is considered a set of different topics, but the essential difference between the methods is that in the latent Dirichlet allocation it is assumed that the distribution of topics is a priori the distribution of Dirichlet. Because of that, a much more correct set of topics is obtained. Also, with the help of Bayesian regularization, the main shortcomings of probabilistic latent semantic analysis are eliminated [18].

The latent Dirichlet allocation is still considered the primary topic modeling method since its further development is reduced mainly to LDA improvement. Hundreds of modifications of the LDA model are known, which take into account various specific features of text collections, such as the correlated topic model (CTM) [19], pachinko allocation model (PAM) [20], etc.

Proposed approach and its implementation

Latent Dirichlet allocation is one of the most common methods of topic modeling used by data scientists. Therefore, it was used to create a topic model of the text document collection in this study. However, proposed approach is applicable in combination with any modern topic modelling method.

The collection of documents, the topic model of which was created, consisted of 2926 files. These are mainly scientific works in the Ukrainian language, devoted to civic identity and related concepts such as civic values and culture, social and organizational identity, archetypes, game theory, trust, historical memory and the phenomenon of identity in general. Number of files according to their extension: .pdf – 1388, .docx – 1134, .doc – 344, .rtf – 52, .djvu – 8.

The Python programming language was used to build a topic model. Many libraries were created to solve problems of different spectra due to its popularity among analysts, data scientists, machine learning specialists and the cohesion and activity of the developer community. For example, there are instrumentalities for natural language processing, stemming (extraction of word bases in synthetic languages, such as Ukrainian, so that the words like "ліс" (forest), "лісу" (of forest) and "лісний" (forest) are not regarded as three different terms) and topic modeling itself.

The software product is divided into two logical parts. The first part is a data processing module that contains the logic of its loading, pre-processing, deleting punctuation, numbers, hyperlinks, stop-words, post-processing, pruning, etc. Each document is processed in the form of a pipeline, which consists of such stages:

- file loading;
- text extraction;
- removal of improper symbols;
- tokenization (splitting text into words);
- removal of stop-word;
- stemming of each token;
- pruning;

Firstly, the program determines the names of all documents in the folder and extracts their textual information with the help of the textract library. The latter provides tools for processing files of various formats, starting from doc and pdf documents and ending with inscriptions on png and jpeg images. Because textract methods make it possible to read the text as a sequence of bytes, we also need to translate each document processing result into UTF-8 format for the correct display of Cyrillic characters. Also at this stage, all characters are translated to lowercase.

The next step is the removal characters that do not belong to the Ukrainian alphabet, such as various punctuation, diacritics, mathematical symbols, and so on. This is realized by means of a special regular expression which is passed to the substitution function to replace matches with an empty string.

Regular expressions are a powerful tool for operating text. They are used for high-speed search by arbitrary template, text transformation, form validation, description of parsers' grammar for programming languages, etc. In our case, we use the regular expression “[^абвгґдеєжзиіїйклмнопрстуфхцчшщьюя\s]” - a negated (complementary) class with all lowercase letters of the Ukrainian alphabet and space (\s stands for space symbol).

All the characters of the text that are not in square brackets match it, and when the search engine finds such a character in the text of the document, it will replace it with an empty string or, in other words, delete it.

On the one hand, this step makes it possible to get rid of dates, page numbers, formulas, punctuation marks, hyperlinks, foreign words and other parts of the text that do not have a thematic coloring and will only clutter the input data for training the model, which significantly facilitates data processing. However, as a result of deleting all non-Ukrainian characters and punctuation of the text, invalid terms may appear (for example, "K.Marx" will become "kmarx"). Because of this, more thorough filtering of the dictionary and bags of words will be required at the pruning stage. After deleting extra characters, the text line of the entire document is broken into separate words (so-called tokens).

Next, each word is checked for belonging to a set of stop-words. A database of stop-words formed from 6 different sources found on the GitHub platform was used. However, since natural language processing is a relatively young branch of Ukrainian data science, at the moment, there is no complete collection of stop-words for the Ukrainian language, so the database was also extended by hand. Its total size is 1995 words; it contains the most commonly used words (various prepositions, conjunctions, parts of phrases, etc.) that do not have a semantic meaning but are only used to form sentences. Some of the words: “так” (yes, such, as), “може” (can, maybe), “аби” (to), “тепер” (now)...

At the next stage, the base words are extracted during so-called stemming. This is a crucial part of any work with natural languages, especially with synthetic ones (as Ukrainian). In such languages, words are inflected, agglutinated, combined with prefixes and suffixes, turned from one part of speech into another without losing their thematic meaning. Without stemming, the words “біг” (run), “бігун” (runner), “бігуни” (runners), “забіг” (run), “біжучий” (running) that refer to the topic of running and have a common base word, will be considered as five different, not interconnected words. After base word extraction, the dimensionality of the general dictionary will be reduced dramatically, which significantly facilitates the topic model construction. However, like other NLP tools, there is no known software for Ukrainian language stemming that would cope with its task perfectly. Uk_Stemmer turned out to be the best available at the time of writing this work, so it was used to stem the tokens. The stemmer does not always extract base words correctly, sometimes leaving suffixes or prefixes, so even after additional manual filtering, invalid terms, such as “україн” (Ukraine) and “українськ” (Ukrainian), can sometimes be found in bags of words. This drawback makes the accuracy of the topic model worse, but the only way to solve this problem is to use a more efficient stemmer.

As soon as the data has been extracted, cleaned, tokenized, filtered and stemmed, it is ready for use in the LDA algorithm. However, experiments on building a topic model based on such data revealed problems related to the fact that there were still many words in the dictionary that were not useful for topic modeling, and some did not correspond to reality. In the initial experiments, the resulting topics contained many surnames and names of authors of scientific works, cities where works were published, words with spelling errors or incorrect terms (as in the example of "K.Marx"). Therefore, it was decided to introduce an additional step of filtering the dictionary - pruning (Fig. 1).

To handle this problem, the topic modeling results were visualized using the pyLDAvis library, which made it possible to see the words that form particular topics and their importance coefficient. Those that turned out to be invalid were "blacklisted" - added to the pruning file. Therefore, after stemming, each term is additionally compared with the list of invalid words and is not added to bags of words in case of belonging to the list.

The set of invalid terms includes 270 words, most of which are proper nouns, abbreviations and words with spelling mistakes.

The second part of the software product is responsible for building a topic model itself.

Fig. 1. Invalid terms

```
[(0,
  '0.007*"особистост" + 0.006*"професійн" + 0.006*"соціальн" + 0.005*"людин" + '
  '0.005*"суб\xd1\x94кт" + 0.004*"процес" + 0.004*"життєв" + 0.004*"розвитк" + '
  '0.004*"житт" + 0.004*"ідентичност"'),
 (1,
  '0.008*"людин" + 0.007*"суб\xd1\x94кт" + 0.006*"соціальн" + 0.005*"розвитк" + '
  '0.004*"особистост" + 0.004*"професійн" + 0.004*"житт" + 0.004*"процес" + '
  '0.004*"особистісн" + 0.004*"громадянськ"'),
 (2,
  '0.008*"професійн" + 0.006*"людин" + 0.006*"розвитк" + 0.006*"особистост" + '
  '0.004*"соціальн" + 0.004*"житт" + 0.004*"особистісн" + 0.004*"досвід" + '
  '0.004*"життєв" + 0.004*"діяльност"'),
 (3,
  '0.006*"людин" + 0.006*"розвитк" + 0.005*"соціальн" + 0.005*"професійн" + '
  '0.005*"особистост" + 0.004*"громадянськ" + 0.004*"суб\xd1\x94кт" + 0.004*"досвід" '
  '+ 0.004*"ідентичност" + 0.003*"особистісн"'),
 (4,
  '0.008*"людин" + 0.007*"суб\xd1\x94кт" + 0.005*"особистост" + 0.005*"житт" + '
  '0.005*"життєв" + 0.005*"соціальн" + 0.005*"розвитк" + 0.005*"процес" + '
  '0.004*"досвід" + 0.004*"особистісн"')
]
```

Fig. 2. Topics of the model

Thanks to the tools provided by the gensim library for building statistical generation models and processing natural languages, Python has a high-level functionality for transforming data to the format required for LDA to work. The general dictionary of the collection is formed using the corpora.Dictionary() method of the library to which the collection of documents is transferred. The bags of words required by the LDA are formed by matching the resulting dictionary to each document in the collection. The vector of all bags of words is the corpus in terms of the Latent Dirichlet allocation. Next, an object of the gensim.models.LdaMulticore class is initialized to which the created dictionary, the bags of words of each document, and the expected number of topics *T* are passed. The parameters α and β are determined automatically in accordance with a given number of topics *T*. Using the LdaMulticore class instead of LdaModel speeds up the construction of a topic model, because the calculation of the expected value and correction of model parameters that occur during the iterative training algorithm are parallelized. At the output we get *T* topics in such form (Fig. 2):

The number marked in violet is the index of the topic, starting with 0, and yellow is the vectors of terms

with a certain weighting factor that make up the topic.

In order to get a more informative representation of the results of topic modeling of a text document collection, it is necessary to use visualization tools. As mentioned in the previous section, each topic can be displayed in the form of a word cloud, as this is sufficient to understand clustering results. However, because the model was in the testing phase, such visualization techniques were used instead: histogram, illustrating the distribution of term usage within the topic; histogram, illustrating the distribution of term usage throughout the whole corpus; intertopic distance map.

The pyLDAvis library was used to visualize the results of document collection topic modeling.

Experiments

In the first experiments, as noted above, the shortcomings of the existing mechanism of stop-word filtering, stemming, and mistake handling were revealed, due to which the obtained topic model did not correspond to reality.

After the stop-word database became almost 20 times bigger and an additional data processing step (pruning) was introduced, which should not allow invalid terms in the general dictionary and corpus, these problems were solved.

Topic modeling of a collection of 30 documents randomly selected from a dataset with parameter T=15, gave the following results (Fig. 3):

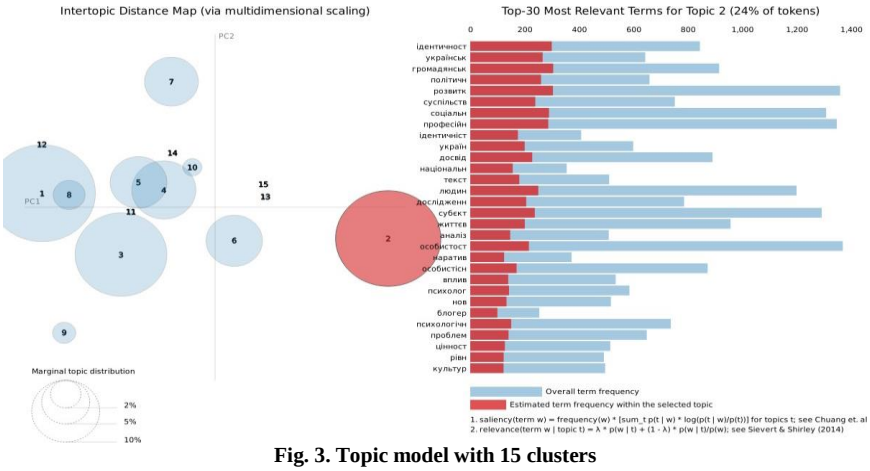

Fig. 3. Topic model with 15 clusters

The intertopic distance map showed that the number of clusters was too large for construction of an effective model of text document collection. The topic №2 (“civic identity”), №9 (“person”), №7 (“professional activity”), №3 (“social roles”) and №1 (“psychology”) are clearly expressed.

The *second* topic ("civic identity") includes: "громадянськ" (civic), "ідентичност" (identity), "українськ" (Ukrainian), "політичн" (political), "суспільств" (society), "національн" (national), "соціальн" (social), "розвитк" (development)

The *ninth* topic ("human") includes: "людин" (human), "субєкт" (subject), "груп" (group), "особ" (individual), "соціальн" (social), "особист" (person), "проблем" (problem), "ідентичност" (identity).

The *seventh* topic ("professional activity") includes: "професійн" (professional), "діяльност" (activity), "особистісн" (personality), "досвід" (experience), "працівник" (worker), "досліджуван" (studied), "адаптаці" (adaptation), "середовищ" (environment).

The *third* topic ("social roles") includes: "рольов" (role), "рол" (role), "розвитк" (development), "ціннісн" (value), "особистост" (personality), "субєктніст" (subjectiveness), "психодрам" (psychodrama), "життєв" (life).

The *first* topic ("psychology") includes: "субєкт" (subject), "соціальнопсихологічн" (socio-psychological), "рольов" (role), "учинк" (act), "особистост" (personality), "вибор" (choice), "психолог" (psychology), "факторн" (factor), "конфлікт" (conflict).

The remaining topics were not explicit enough for their possible interpretation and use for the categorization of documents. In addition, most of them overlapped partially (topics №4 and №5) or completely (topics №1 and №8).

Thus, it was decided to reduce the number of topics to seven because this number of topics would let the model be easily interpreted. At the same time, the topics will not be too general, which would deprive any meaning of clustering (Fig. 4).

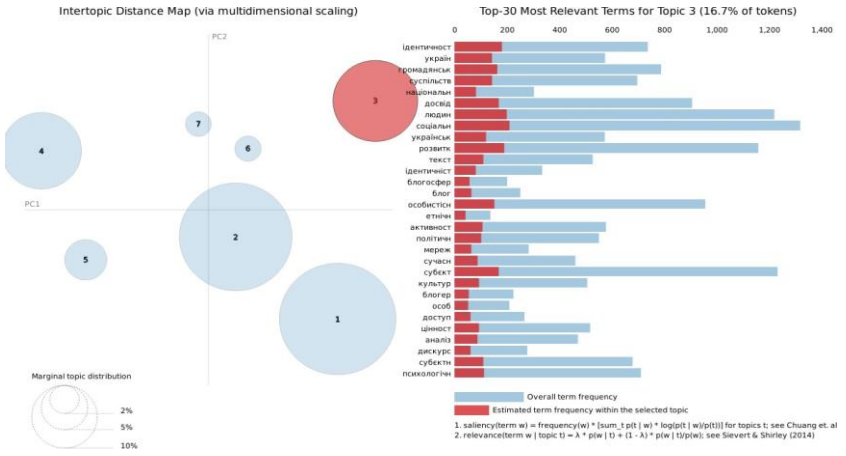

Fig. 4. Topic model with 7 clusters

The most expressed topics were №4 (“professional activity”), №3 (“identity”), №1 (“social roles”) and №2 (“personality”). Topic №6 has less clearly defined boundaries, as it contains terms associated with volunteering, psychology, culture and other types of human social activity. Although topics №7 and №5 do not overlap and are not even close on the intertopic distance map, they turned out to be very close in general thematics: both topics

focus on professional activity but are closer to topics №3 and №1 respectively.

Topics 1-4 together cover 92, 7% of all terms in the general vocabulary. This fact indicates that from the point of view of the classification task, the given number of topics is still too large, and at a value of T=4, it would be possible to unambiguously categorize the collection of documents, obtaining more or less “pure” topics. However, one of the advantages of unsupervised machine learning methods is the absence of predefined rules, which provides an opportunity to identify hidden relations and build new hypotheses on the clusterization results basis.

Therefore, topics №7 and №5 were left as a context extension of topic №4 by topics №3 and №1 respectively for further analysis of the topic model with the involvement of an expert in the subject area.

Discussion

I. Petrovska, an expert in political psychology, researching the phenomenon of civic identity for six years, has more than 25 scientific publications on this issue, was involved in evaluating the constructed topic model of the collection of text documents.

As a result of expert evaluation of the topics singled out during the modeling, it was proposed to clarify the formulation of cluster names based on the semantics of the sets of words that form them.

In particular, the topic №1, formerly called "social roles", was renamed "social role of the citizen" (Fig. 5).

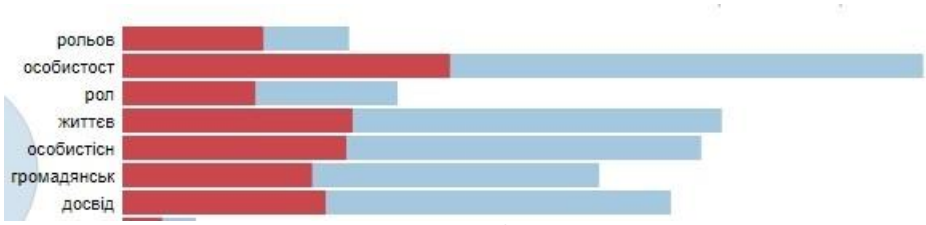

Fig. 5. "Social role of the citizen"

If citizenship means a certain formal status, rights and responsibilities, then in this respect, a person gets a specific role - the role of a citizen that they must play throughout life. In this case, it is possible to interpret civic identity at the individual level in the context of modern role theories of social psychology (E. Bern, E. Goffman, etc.). The individual may, to varying degrees, identify themself with the role of a citizen, "know their role" good or bad, but in each case, the role of the citizen becomes part of their role repertoire [21].

Given the concepts that form the topic №2 (Fig. 6), its name was changed from "personality" to "subjectness of personality".

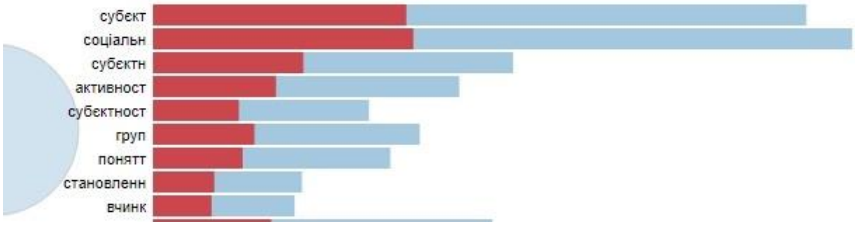

Fig. 6. "Subjectness of personality"

According to the expert, it is the individual's subjectivity that has a significant impact on the formation of civic identity. Considering subjectivity as an integral property of personality, which is manifested in the ability to independence, activity, initiativeness, responsibility, self-determination, self-regulation and self-improvement, to a conscious and active attitude to the world and oneself in it, as an initiative-creative principle of personality, which helps to set goals and outline life plans, choose life strategies, create conditions for personal development. The subjectiveness of an individual is an essential category for achieving self-identity, building a holistic image of "I as a citizen", civic self-determination [22].

The analysis of the terms that form the topic №3 gave grounds for correcting the name of the topic from "identity" to "social identity" (Fig. 7).

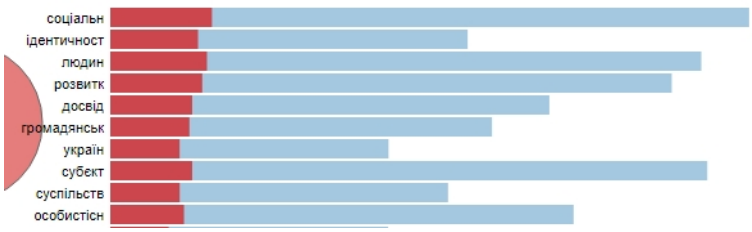

Fig. 7. "Social identity"

Civic identity is an element of the system of an individual's social identities. A crucial indicator of mature

civic identity is its inclusion in the individual's system of social identities (the presence of civic identity in the subjective hierarchy of other personal identities). Otherwise, when civic identity is not an element of the system, not included in the significant (emotionally and existentially) meanings, it will have the character of a superficial layer, which is not sufficiently developed and easily changes depending on the external situation [21]

In contrast to the previous three topics, which under minor adjustments in names, the topic №4 was significantly reformulated (from "professional activity" to "individualization of civic identity") taking into account the essential characteristics of the formation of civic identity (Fig.8).

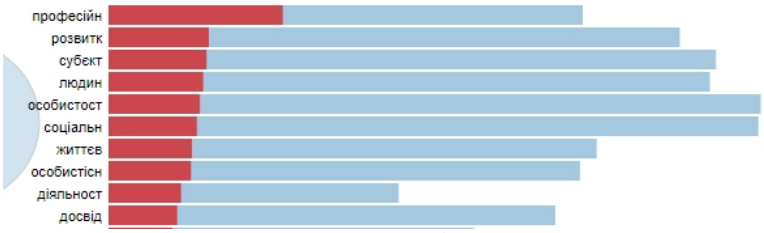

Fig. 8. "Individualization of civic identity"

Individualization of civic identity is associated with the beginning of the practical implementation of an individual's life plans in the organizational space of the state, in particular, with the beginning of professional activity and building of own career, acquisition of personal experience of activity in the organizational environment of the state. Civic maturity is achieved through the individualization of the meanings of citizenship. The personal attitude to the state and fellow citizens develops; meaning and value are given to one's citizenship. The content of civic guidelines will depend on the perception of opportunities/conditions of self-realization in the organizational environment of the state, security (stability, existence of social guarantees), social recognition. On this basis, the awareness and understanding of the content and meaning of life as a citizen in the state are formed [21].

The expert suggested removing the clusters №5 (4.1%), №6 (1.6%) and №7 (1.5%) from the topic model, as they do not have characteristics sufficiently expressed for a qualitative semantic interpretation.

In general, the obtained topic model of the text document collection received a positive expert assessment. According to the expert, the topics that were singled out represent an idea of the civic identity phenomenon and, when used to categorize documents, can simplify researchers' work with literature sources on this issue. This demonstrates the efficiency of proposed approach.

Conclusion

Topic modeling is a process of building a topic model that involves the simultaneous, usually fuzzy, clustering of words and documents by their semantic proximity. It is assumed that documents are formed due to a combination of a number of unknown topics, with each topic covering a set of words, the observation of each of which in the document with a certain probability indicates that the document belongs to this topic. Thus, the topics of the model are a set of hidden (latent) variables of known size that are subject to detection. Since the set of topics is usually much smaller than the set of all used words, topic modeling allows representing the document as a vector in the topic space instead of representation in the term space, as a result of what the document has fewer components, which allows faster and more efficient processing. The formed clusters reflect semantic interrelations. Therefore, topic models are also used to reveal trends in scientific publications or news streams, for classification and categorization of documents, images and video streams, for information retrieval, including multilingual, tagging web pages, detecting text spam, and referral systems and other applications.

The main methods of topic modeling include LSA (latent semantic analysis), pLSA (probabilistic latent semantic analysis), LDA (latent Dirichlet allocation), and various specialized adaptations of the latter, i.e., CTM (correlated topic model) or PAM (pachinko allocation model), etc. The specificity of the LSA method is that the input word-document matrix is decomposed into two separate matrices (document-topic and topic-word) during the construction of the model. At the same time, pLSA, LDA and modifications of the latter are based on generative probabilistic models and iterative improvement of their initial approximation using the Expectation-Maximization algorithm.

The method of latent Dirichlet allocation was used to build a topic model of a collection of scientific publications on the issue of civic identity and related topics, since, among all well-studied methods, its probabilistic model is, in essence, the closest to the fundamental nature of the topics as a semantic phenomenon. However, the proposed approach is applicable with any of the other modern topic modelling methods.

A software product implemented in Python consists of two components. The first is a module for working with data, which implements almost all its loading and preprocessing logic. The most important part of the first module is the data transformation and filtering pipeline, the structure of which is described in detail in the section devoted to proposed approach. The second module converts the document-word matrix into a word bag vector, forms a general dictionary of the collection, calculates the TF-IDF metric, performs topic modeling, the results of which are exported as an HTML page that can be viewed in any modern browser.

The topic model of the collection of scientific publications on the issue of civic identity and related topics was obtained with the help of the developed software. After minor optimization the number of 7 topics was reached. The most expressive topics identified during the modeling were given the following names: №1 - "social roles", №2

- "personality", №3 - "identity", №4 - "professional activity".

An expert in political psychology, who studies the phenomenon of civic identity, evaluated the topics singled out during the modeling, and proposed to clarify the formulation of cluster names based on the semantics of the sets of words that form them.

In general, the obtained topic model of the text document collection received a positive expert assessment. According to the expert, the topics that were singled out represent an idea of the civic identity phenomenon and, when used to categorize documents, can simplify researchers' work with literature sources on this issue. However, the expert suggested removing the clusters №5 (4.1%), №6 (1.6%) and №7 (1.5%) from the topic model, as they do not have characteristics sufficiently expressed for a qualitative semantic interpretation.

References

1. Tkalenko O. Intelligent technologies and artificial intelligence systems to support decision making / O. Tkalenko, A. Makarenko, O. Polonevych // Telecommunication and information technologies. – 2019. – Vol. 2. – P. 53–59.

2. Daud A. Knowledge discovery through directed probabilistic topic models: a survey / A. Daud, J. Li, L. Zhou, et al. // Front. Comput. Sci. China. – 2010. – Vol. 4. – P. 280–301.

3. Vorontsov K. Probabilistic topic modeling. 2013. URL: www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf

4. Jain A. Data Clustering: A Review / A. Jain, M. Murty, P. Flynn // ACM Computing Surveys. – 1999. – Vol. 31, No. 3. – P. 264–323.

5. Vorontsov K. Regularization, robustness and sparsity of probabilistic topic models / K. Vorontsov, A. Potapenko // Computer Research and Modeling. – 2012. – Vol. 4, No. 4. – P. 693–706.

6. Argyrou A. Topic modelling on Instagram hashtags: An alternative way to Automatic Image Annotation? / A. Argyrou, S. Giannoulakis, N. Tsapatsoulis. – 2018. URL: https://ieeexplore.ieee.org/abstract/document/8501887.

7. Kirill Y. Propaganda Identification Using Topic Modelling / Y. Kirill, I. Mihail, M. Sanzhar, M. Rustam, F. Olga, M. Ravil // Procedia Computer Science. – 2020. – Vol. 178. – P. 205–212.

8. Huang T. Automatic meeting summarization and topic detection system / T. Huang, C. Hsieh, H. Wang // Data Technologies and Applications. – 2018. – Vol. 52, No. 3. – P. 351–365.

9. Venkatesh A. On Evaluating and Comparing Open Domain Dialog Systems / A.Venkatesh, C. Khatri, A. Ram, F. Guo, F., et al. – 2018. URL: https://arxiv.org/pdf/1801.03625.pdf

10. Ma J. A Message Topic Model for Multi-Grain SMS Spam Filtering. / J. Ma, Y. Zhang, Z. Wang, K. Yu // International Journal of Technology and Human Interaction. – 2016. – Vol. 12, No. 2. – P. 83–95.

11. Spina D. Learning similarity functions for topic detection in online reputation monitoring / D. Spina, J. Gonzalo, E. Amigó // Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. – 2014. URL: https://dl.acm.org/doi/10.1145/2600428.2609621

12. Tutubalina E. Exploring convolutional neural networks and topic models for user profiling from drug reviews / E. Tutubalina, S. Nikolenko // Multimedia Tools and Applications. – 2017. https://doi.org/10.1007/s11042-017-5336-z

13. Peters N. Task Boundary Inference via Topic Modeling to Predict Interruption Timings for Human-Machine Teaming / N. Peters, G. Bradley, T. Marshall-Bradley // Advances in Intelligent Systems and Computing. – 2019. – P. 783–788.

14. Schneider N. Chemical Topic Modeling: Exploring Molecular Data Sets Using a Common Text-Mining Approach / N. Schneider, N. Fechner, G. Landrum, N. Stiefl // Journal of Chemical Information and Modeling. – 2017. – Vol. 57, No. 8. – P. 1816–1831.

15. Asmussen C. Smart literature review: a practical topic modelling approach to exploratory literature review / C. Asmussen, C. Møller // Journal of Big Data. – 2019. – Vol. 6, No. 1. https://doi.org/10.1186/s40537-019-0255-7

16. Hofmann T. Probabilistic Latent Semantic Analysis / T. Hofmann. – 1992. URL: https://www.iro.umontreal.ca/~nie/IFT6255/Hofmann-UAI99.pdf

17. Blei D. Latent Dirichlet Allocation / D. Blei, M. Jordan // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 993–1022.

18. Günther E. Word Counts and Topic Models / E. Günther, T. Quandt // Digital Journalism. – 2016. – Vol. 4, No. 1. – P. 75–88.

19. Blei D. Correlated topic models / D. Blei, J. Lafferty // Advances in neural information processing systems. – 2006. – Vol. 18. URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.958.2484&rep=rep1&type=pdf

20. Li W. Nonparametric Bayes Pachinko Allocation / W. Li, D. Blei, A. McCallum. – 2007. URL: https://arxiv.org/ftp/arxiv/papers/1206/1206.5270.pdf

21. Petrovska I. Psychological Model of Civic Identity Formation / I. Petrovska // Journal of Education Culture and Society. – 2021. – Vol. 12, No. 2. – P. 167–178.

22. Petrovska I. Civic identity development: ontogenetic aspect / I. Petrovska // Social Welfare: Interdisciplinary Approach. – 2019. – Vol. 9, No. 2. – P. 29–43.