Enhanced LSA Method with Ukraine Language Support

Nataliia Kunanets[a], Yurii Oliinyk[b], Dmytro Myhal[b], Khrystyna Shunevych[a], Antonii Rzheuskyi[a] and Yuriy Shcherbyna[c]

[a] *Lviv Polytechnic National University, 12 Bandera street, Lviv, 79013, Ukraine*
[b] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Prosp. Peremohy, Kyiv, 03056, Ukraine*
[c] *Ivan Franko National University of Lviv, University Street 1, Lviv, 79000, Ukraine*

Abstract

The use of semantic models is relevant in automated learning systems, in solving certain tasks, such as: extracting knowledge from texts, information retrieval, abstracting, checking the correctness of vocabulary terms and definitions, automatic generation of associative links in hypertext databases, etc. No less important is the development of new tools and instruments to automate semantic analysis. Such methods of analysis allow you to collect basic information about a particular topic, focus and mood of the texts, which will further simplify the automated work with them, such as cataloging, search and comparison. The objectives of this study are: development of the LSA method with support for processing Ukrainian-language texts, justification of the choice of technologies for the implementation of methods and tools of semantic analysis, study of the effectiveness of the developed method and software.

Keywords 1

Semantic Analysis, Thematic Modeling, Ukrainian Language, Latent Semantic Analysis

1. Introduction

The flow of information in the modern information society is constantly growing and needs to be processed, including semantic and sentiment analysis. There are quite a few means of semantic analysis of texts, which in the general case are defined and directly depend on the task and language. Increasingly popular are services-tools of network thinking, which are used for text analysis in workflows, as well as on the World Wide Web. One such service is InfraNodus - a network thinking tool that detects relationships and patterns in the body of texts [1]. The InfraNodus approach improves existing textual content retrieval techniques that use approaches such as latent semantic analysis or LSA, pLSA, Pachinko distribution, latent Dirichlet or LDA allocation, relational theme models, the word2vec algorithm, and its lda2vec extension. Another system that provides effective text analysis is WordWanderer. It implements a visualization technique that increases the capabilities of the usual tag cloud to the navigation interface for text. This tool supports functionality from "contextual representation", which represents all words that occur next to the selected word, to "comparative representation", which arranges words based on the degree of their association, as well as meaning and semantic relationship. WordTree, or word tree as you can interpret the name of this system, implements a new technique of visualization and retrieval of information in text documents. The word

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: nek.lviv@gmail.com (N. Kunanets); oliyura@gmail.com (Y. Oliinyk); dimamygal0708@gmail.com (D. Myhal); krishirak@gmail.com (K. Shunevych); antonii.v.rzheuskyi@lpnu.ua (A. Rzheuskyi); yshcherbyna@yahoo.com (Y. Shcherbyna)
ORCID: 0000-0003-3007-2462 (N. Kunanets); 0000-0002-7408-4927 (Y. Oliinyk); 0000-0001-7787-4011 (D. Myhal); 0000-0002-2282-4575 (K. Shunevych); 0000-0001-8711-4163 (A. Rzheuskyi); 0000-0002-4942-2787 (Y. Shcherbyna)

tree is a graphical version of the traditional keyword-incontext (KWIC) method and provides the generation of quick queries and research results of a specific text context.

2. Related works

Sentiment analysis is characterized as contextual manipulation of text that recognizes subjective information in the source material and then extracts it, as well as helping many companies to understand the social mood of their product or service. Almost all professional and non-technical companies use sentiment analysis to reliably interpret customer contributions and wishes regarding the conformity of a product or service. Sentimental research is used by leading companies[1].

Sentiment analysis is considered as a technique of determining and extracting human feelings with the help of unstructured text and is carried out with the help of natural language processing and machine learning. Machine learning is a great way to help with learning and learning data sets from social networks. This is the most common strategy used in mood analysis. With the help of machine learning you can use lexical methods, as well as methods based on rules. Sentiment analysis is the best tool for determining a positive or negative assessment. The various difficulties in evaluating opinions are that the public does not always formulate feelings in the same way, which means several statements in the mode of giving assessments and comments, and others in the form of phrases that do not convey any proper thinking[2].

User opinions in product reviews or other textual data are crucial for manufacturers, sellers and suppliers of goods and services. Therefore, the analysis of moods and the extraction of opinions have become important areas of research. In the extraction of user reviews, approaches based on topic modeling and the Hidden Dirichlet Distribution (LDA) are important techniques used to extract aspects of the product in terms of mood analysis. However, the LDA cannot be directly applied to user reviews and other short texts due to the problem of data sparseness and the lack of co-occurrence models. Baris Ozyurt and M. Ali Akcayol propose a new method of mood analysis based on aspects, the LDA Sentence Segment (SS-LDA)[3].

Modern approaches to the analysis of moods at the level of attributes, based on deep learning, have achieved promising results; however, these methods create strict preconditions, require manually marked training data and predefined attributes by experts, and classify only sentiment in introduced categories that have limited implications. Yi Han and Mohsen Moghaddam develop a rule-based methodology for the mass extraction and analysis of user expressions from the reviews available on social media and e-commerce platforms. The methodology further improves the current uncontrolled approaches to the analysis of moods at the attribute level, ensuring the effective identification and display of mood expressions of individual users on their respective attributes[4].

There are various methods of sentiment analysis, but recently the methods of embedding words are widely used in the classification of sentiments. Currently, Word2Vec and GloVe are some of the most accurate and used word embedding methods that can convert words into meaningful vectors. However, these methods ignore information about the sentiment of the texts and require a large body of texts to learn and form accurate vectors. As a result, due to the small size of some corpora, the researcher often has to use pre-trained embedding of words that have been trained in other large text corpora[5].

There are some methods that use a social context, but they have only built a global model of sentiment analysis without being able to extract personalized expressions. Some personalized methods have been proposed to address this problem, but they suffer from data scarcity and inefficiency. Based on personalized mood analysis methods, we use information about the social context and capture variable and expressive user expressions at the community level to address these issues. Xiaomei Zou, JingYang, Wei Zhang, Hongyu Han offer a common approach to analyzing the sentiment of microblogging. The authors constructed two classifiers. One is the global microblogging sentiment analysis model, which can use sentiments common to all users. One of them is a specific model of joint analysis of the sentiment of microblogging, which can extract sentiments under the influence of user personalities. In addition, the authurs extract knowledge about the similarity of the community and use it to improve the learning process for a community-specific mood model[6].

Although there are not many software tools for supporting texts in the Ukrainian language today, some of them have already proven themselves. Some of them can be found on the website lang-uk [7], which is open to join the community of specialists in the field of computer word processing (programmers, linguists, researchers). This community is built on common principles and supports existing and the development of new projects for the collection of Ukrainian corpora and other textual data. The main directions of their work are collection and publication of corpora and other sets of text data in Ukrainian, creation on the basis of these data of models for solving applied problems of processing of Ukrainian texts and implementation of these models in a set of publicly available microservices. Among their active projects are collecting texts for the BrUK corpus, collecting the maximum possible corpus of Ukrainian texts from various sources, building models of vector representation of words based on this corpus, creating an annotated corpus for the task of constructing connections between entities, means of text tone analysis [8, 9].

Researchers [10] M. Romanyshyn, A. Romanyuk and M. Lobur analyzed the existing approaches to the use of sentiment analysis and tried to determine the most convenient for the Ukrainian language. The authors advocate the benefits of a rule-based approach but require manual rules to be created. Such rules are usually narrowed down to a specific topic. One of the advantages of the method is the accuracy of the results.

An important result that allowed [11] localized the SentiStrength program to the Ukrainian language based on the use of a corpus approach for processing a large number of texts and automatically obtaining information on their positive or negative orientation. The results of the study show that the software product localized to Ukrainian texts gives quite valuable and reliable results (accuracy 0.70). Researchers have developed a sentimental dictionary on political topics [12]. A comparative analysis of the linguistic annotation of texts, performed manually and semi-automatically using the UAM Corpus Tool, shows in favor of the latter.

The procedure of adaptation of the VADER sentiment analyzer to the Ukrainian language offered by researchers shows good results [13].

An interesting project is the morphosyntactic analyzer of the Ukrainian language [14], which allows for automated morphological analysis of tokens. This model is trained on the gold standard and has an accuracy of 91.6% morphological features and 81.7% synthetic connections. An API has also been developed that receives text in UTF-8 and a size of no more than 1 megabyte, and outputs its full parsing (division into words, sentences, morphos and parsing). Another variant of the morphoanalyzer is pymorphy2 [15], written in Python and placed in an open github repository. It allows you to bring a word to a normal form, such as "people -> man ", or" walked -> walk ". It is also possible to put the word in the desired form, such as plural, change the case of the word, etc. Pymorphy2 also allows you to get grammatical information about the word - its number, genus, case, part of speech, etc. [16]. When pymorphy2 OpenCorpora dictionary is used, and hypotheses are built for unfamiliar words.The library has a fairly high speed, the speed is from several thousand words per second to more than 100 thousand words per second depending on the operation, interpreter and installed packages.

3. Natural language topic modeling methods

Any linguistic research becomes much more effective if you rely on the body of texts and tools that support the possibility of processing texts in the Ukrainian language. An important factor in such research is the general regionally annotated corpus of the Ukrainian language (GRAC) [17], which is a large, structured and representative collection of texts written in the Ukrainian language and allows you to create separate subcorpora based on the corpus, search for words and phrases of different models and grammatical forms. search, form a sample, sort its elements by various parameters and obtain the necessary statistics. Because the general regionally annotated corpus is a morphologically distinct corpus, it allows you to create queries by word form, lemma, or tag, and their various combinations.

Thematic modeling is a way to build a model of a body of texts that reflects the transition from a set of documents, a set of words in documents to a set of topics that characterize the content of these documents. The most popular methods of thematic modeling can be divided into two main groups - algebraic and probabilistic. Algebraic models include the standard VSM vector text model and LSA

latent semantic analysis, and the most popular probabilistic models are PLSA latent semantic analysis and LID Dirichlet latent placement.

Latent-semantic analysis is a method of processing information in natural language, which helps to identify the relationship between the collection of documents and the terms found in them, comparing some factors (topics) of documents and terms, in particular to identify latent relationships between phenomena or objects under study, as well as in the classification of documents to extract context-dependent meanings of lexical units through statistical processing of large corpora of texts. The properties extracted from the documents clearly characterize high-level concepts and topics in order to give an accurate idea of their content. In each article on a particular topic, most words are used quite often, and each word can occur in several topics. In this case, words can have several meanings, because their inherent ambiguity, the same word can be associated with several concepts, and are interpreted differently in different subject areas.

Ambiguity often complicates the correct definition of similarities and relationships in the body of texts. For example, different documents discuss the same topic, but use different terms, which can lead to incorrect recognition by the text system, they will be interpreted as concepts that are not related to each other. However, the corresponding words of both documents may be present in several similar documents. This situation means that such words are semantically related, such documents are potentially related, despite the fact that they differ in the words in which they express the meaning of the text.

This method facilitates the search for documents in collections and terms in texts as in a hidden (latent) property space, in which the dimensions correspond to high-level concepts. In collections, each document is presented as a weighted combination of components, and each term may have different connections to other concepts. This system of relationships is very similar to that of the Principal component analysis (PCA) method, which is used to map vector space with possible relationships between dimensions. The principal components method solves a number of problems:

1. Search for hidden patterns in phenomena and documents.
2. Analysisphenomena characterized by numerical parameters, rather insignificant at the initial stage.The number of major components identified in the study will contain more information than the initially measured features.
3. Identification of the main features that best characterize the main component, as well as stochastic relationships between them.
4. Forecasting the levels of the studied phenomena on the basis of the regression equation obtained on the basis of the information of the main component.

The advantages of this method of forecasting in contrast to the classical regression analysis can be explained by the fact that the model must include the maximum possible number of factors, which are often characterized by significant correlation. The forecast for such variables is usually accurate. Therefore, there is a need to replace the original interconnected variables with a set of uncorrelated parameters. This problem is solved using the method of principal components, which allows you to form the characteristics built on the basis of primary measurements.

When analyzing texts, this method allows you to implement practical tasks, including:

- analysis of causal relationships between components and establishing their stochastic relationship with the main components
- selection of generalizing indicators
- ranking the results of observations of the main components
- classification of objects of observation
- formation of the list of tonality of the information.

In order to determine the frequency list of phrases of the model, it is enough to formulate a query, the result of which we obtain three lists of a representative sample of uses of the studied phrases, sorted by decreasing frequency of phrases. Also at this stage we calculate the weight of each word. To do this, divide the absolute frequency of each word by the total number of words and multiply by 100. The simplest tools allow, say, to determine the tonal load of the text by the number of words with a certain sentimental meaning. For example, the absolute frequency of the word "good" is 8, and the total number of words in the list is 4750, the weight of the word is equal to: 8/4750 * 100 = 0.17%. To simplify the process of calculating the weight of each of the words, we use Excel, writing the formula

and applying it to all tables with reference lists of phrases. Similarly, we calculate the absolute frequency of use of words using the function of intermediate results, we obtain a list in which the digital equivalent of the frequency of their use.

A convenient tool for sentimental text analysis is SentiStrength, which is lexicon-based and uses additional (non-lexical) linguistic information and rules to determine mood in short informal texts in English, and available applications for Ukrainian. The software product allows you to assess the strength of the tone for each message on two scales simultaneously: -1 to -5 and 1 to 5. The use of two scales is based on research by psychologists that people perceive positive and negative judgments in parallel [18]. Both positive and negative moods can also be contained within one sentence [19]. SentiStrength can also output results based on binary classification (positive / negative) as well as based on three classes (positive / negative / neutral). Due to the fact that the program is based on the use of a dictionary of emotionally colored words, it can also be customized to a specific subject area.

The basis of the algorithm is a list of emotionally colored words, in which each word in the list is assigned a certain value that corresponds to the intensity of emotion (from 2 to 5). In certain words, the variable part is denoted by the symbol asterisk (*), for example, after the word ador * in the text, the program will find such words as: adorable, adored, adoring, etc. The only word in the dictionary that has both positive and negative evaluations is the word miss meaning "to miss someone" (it is often used to express both sadness and love).

The analyzer has a built-in learning algorithm in order to optimize the meaning of emotions in a dictionary marked by a person.

There is also an algorithm for correcting spelling mistakes. The program automatically removes letters that are repeated more than twice (for example, *helllo* will replace the program with *hello*), also removes duplication, if such a connection is not typical of English (for example - *niice*, the program replaces with *nice*), the program removes duplication in words if after correction the word becomes normative (for example, *nnice* will be corrected on *nice*, but not *hoop* on hop, or *baaz* on *baz*). The program uses a list of words that can increase or decrease the tone value. The value of each word increases by 1 or 2 units (for example, with the adverbs *very* or *extremely*) or decreases by 1 unit (for example, with the word *some*). There is a list of words that express objections. They change the tonality values to the opposite. Recognition of cases when the presence of negative words does not change the tone is not built-in, because, in the trial data set, such cases were rare. If there are emotionally colored words with errors in the sentence, they are corrected and the value of the key is changed by 1 unit.

3.1. Vector space model

This model is used to solve many problems of rapid analysis of documents, as well as to compile tables of search, classification and clustering, and serves as a basis for many other algorithms. In this model, the document is considered as an unregulated set of terms - words and additional elements that make up the text. Terms can be both words and their combinations, the so-called n-grams, documents - ideally: sets of thematically homogeneous texts, or just any three-dimensional text arbitrarily divided into pieces, such as paragraphs. First of all, all documents are pre-processed. First of all, it is the removal of all punctuation marks and special characters. Also, it is necessary to exclude the so-called stop words or noise words - these are words that do not carry a semantic load, so their usefulness and role in the search is not significant. For example, all prepositions, suffixes, adjectives, exclamations, numbers must be excluded. A list of noise words for the Ukrainian language can be found in the repository of the Ukrainian search engine analyzer Lucene [20].

3.2. Latent semantic analysis

LDA - Latent Dirichlet Allocation - a generating model used in information retrieval and allows you to explain the results of observations using some implicit (latent) groups. This model is an extension of another, similar in properties to the PLSA model, and eliminates its main shortcomings by using the Dirichlet distribution, resulting in a set of topics is more specific and clear. This model avoids many disadvantages of its previous version of PLSA, such as:

- "retraining" occurs when the model is too complex, namely one that has too many parameters relative to the number of observations;
- lack of regularity in the generation of documents from a set of received topics, which significantly improves the final sample.

pLSA is a statistical method for analyzing the correlations of two types of data. In general, this method is a development of latent-semantic analysis, but unlike its predecessor, which was essentially an algorithm for constructing a vector representation with subsequent reduction of its dimension, probabilistic latent-semantic analysis is based on mixed decomposition and use of probabilistic model, which allows better and more clearly identify possible topics of documents. PLSA, in turn, uses the probabilistic method instead of SVD. The basic idea is to find a plausible model with hidden themes that can generate the data observed in our term-document matrix. In particular, you need a model P (D, W), such that for any document d and words w, P(d, w) corresponds to this record in the therm-document matrix [15]. Assume that each document consists of a set of topics and each topic consists of a set of words, then PLSA adds a probability factor to the assumptions:

- topic z is present in document d with probability P(z | d);
- for a given topic z, the word w is taken from z with probability P(w | z).

4. Enhanced LSA method with Ukraine language support

A corpus of Ukrainian-language texts is submitted to the input of the model, each of which must have a title and content to be analyzed. The body should contain a sufficient number of documents on various topics so that the model can highlight several different topics. The case should not contain duplicate documents. It is also necessary to pass two limiting parameters [21].

The first parameter is the maximum number of topics that need to be output. Because the model will not be able to determine whether it is accurate enough to distribute the case on a certain number of topics, the user must set the upper limit. This does not mean that the model will identify as many topics as defined in the constraint, as some topics may include others.

The second is the maximum number of words that should contain all topics in general. This parameter is necessary to visualize the result in the form of a graph, because too complex structures can be difficult to perceive.

At the exit of the model we get a number of topics that were selected from the body of texts. Each topic is a set of terms that are sorted by relevance to that topic in descending order.

To perform calculations in our study, the method of latent-semantic analysis was chosen, which is supplemented by our work, which provides support for the processing of Ukrainian-language texts. The possibilities of semantic analysis of the method were increased due to the use of the NER model. Named Object Recognition (NER) technology involves searching for and classifying objects in the text by a predefined category, including people names, organization names, locations, time parameters, quantities, monetary equivalents, and so on.

The developed grammar-based system provides better accuracy and speed and uses a free open source library, which has greatly simplified advanced natural language processing using the Python programming language.

The spaCy module from the python library gives the model the ability to recognize a wide range of named objects, add arbitrary classes to the object recognition model, and teach the model to update it with new learned examples. The model is configured to memorize normalized terms and TF-IDF matrix of each analyzed case, which helps to increase the speed of the method.

The main advantages of this method can be considered the high quality of definition of existing topics if the body of texts is large enough, as well as the ability to find non-obvious semantic relationships between words. The disadvantage of this algorithm is the high computational complexity and low speed, which requires recalculation of all metrics for the entire body in the case of adding a new document. However, by optimizing the preservation of calculations, you can significantly increase the speed of the system, so this disadvantage should not interfere with its operation. Also, the method has high requirements for the body of texts, which consists of many different documents on the subject.

4.1. An example of application of the developed method

The input data is a corpus of Ukrainian-language texts, consisting of four short documents that contain common words to clearly show the results (Fig. 1).

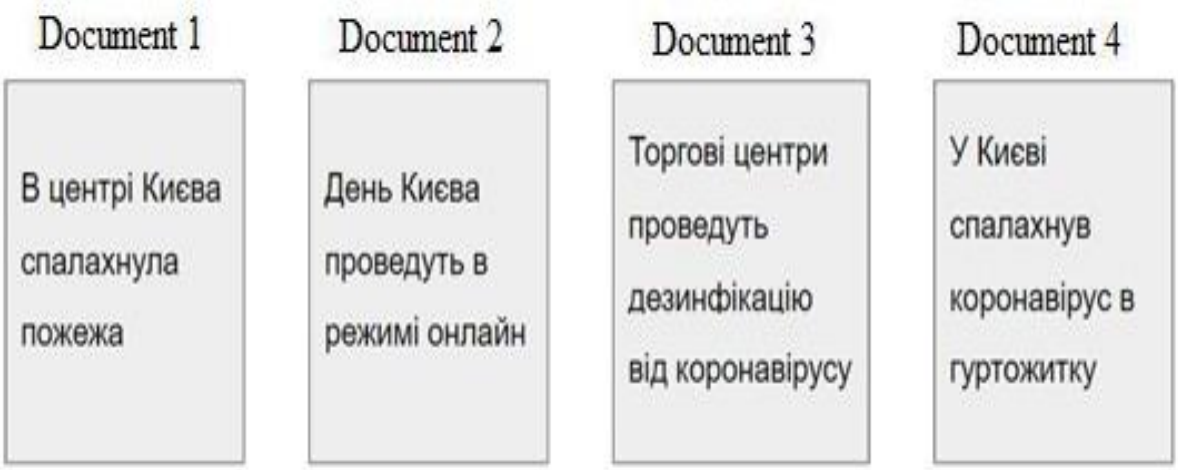

Figure 1: Input Document

After bringing all the words to the normal form, the part of speech of this term is determined using the pymorphy2 morphoanalyzer (Fig. 2).

Figure 2: Example of term vectors for each document

As a matrix *A* used TF-IDF matrix, the matrix *A**, containing only *k* the first linearly independent components, reflects the basic structure of the various dependencies present in the original matrix (Fig. 3).

-0.0252581	0.568787	-0.584267	-0.179613
-0.0775253	0.355595	0.344476	0.66604
-0.541819	-0.0851051	-0.0404995	-0.0743223
-0.116261	0.155475	0.0155916	-0.14681
-0.0513918	0.462191	-0.119896	0.243214
-0.541819	-0.0851048	-0.0404995	-0.0743224
-0.0142237	0.29369	0.702724	-0.480114
-0.309672	0.135245	0.151988	0.295859
-0.541819	-0.0851048	-0.0404995	-0.0743224
-0.0197407	0.431238	0.0592284	-0.329863

U

1.09468	0	0	0
0	0.813891	0	0
0	0	0.711116	0
0	0	0	0.613816

S

-0.0460827	-0.0259509	-0.141442	-0.988533
0.771551	0.398387	0.48236	-0.115443
-0.59772	0.718905	0.352408	-0.0414319
-0.212877	-0.569028	0.789387	-0.0880862

V^T

Figure 3: Singular schedule components for selected input data

As the matrix A was used TF-IDF matrix, the matrix A^*, containing only k the first linearly independent components, reflects the basic structure of the various dependencies present in the original matrix (Fig. 4).

Figure 4: The result of the distribution of terms on the topics and the work of the NER-analyzer for the selected input data

5. Enhanced LSA method software

Node.js technology and JavaScript programming language were used to develop the software that implements the method. The program code is available at the link [22,23].

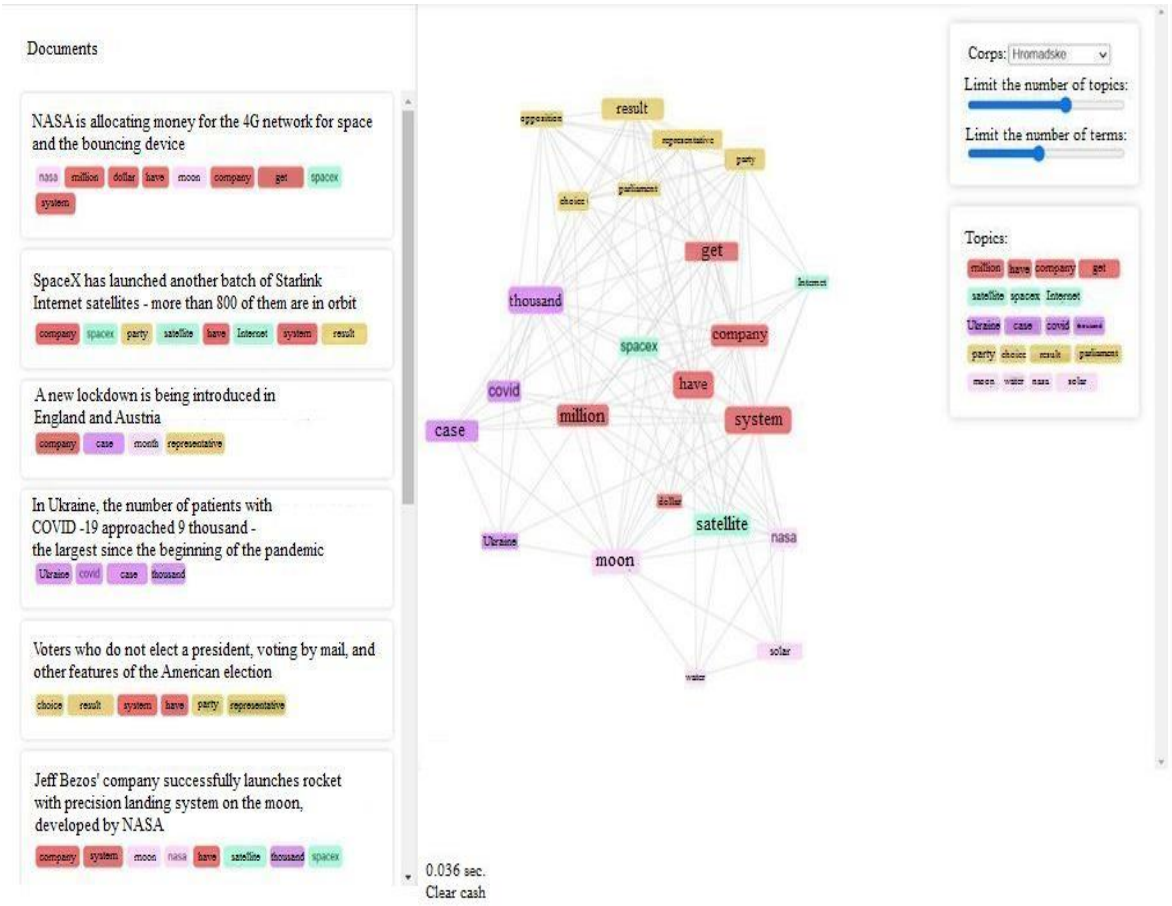

Figure 5: Graphical interface of the developed web application

During the debugging of the system, it is loaded with a set of corpora of texts on various topics. The main panel of the system interface allows the user to configure the input data for the model (Fig. 5).

On it the user by means of the drop-down list can choose the case of texts of the necessary subject, for carrying out the analysis.

Also, this panel contains two sliders, with which the user can set limits on the number of topics and the number of terms that will be displayed as a result of the analysis in the system interface.

By default, the first case loaded into the system is selected, the slide limit on the number of terms is set to the maximum value - 50 terms, and the limit on the number of topics - 40% of the number of documents in this case.

The most important part of the graphical interface is the graph of terms, obtained as a result of semantic analysis of the corpus of Ukrainian-language texts (Fig. 6). Nodes in this column are terms, and edges are an indicator of the presence of two terms in one document.

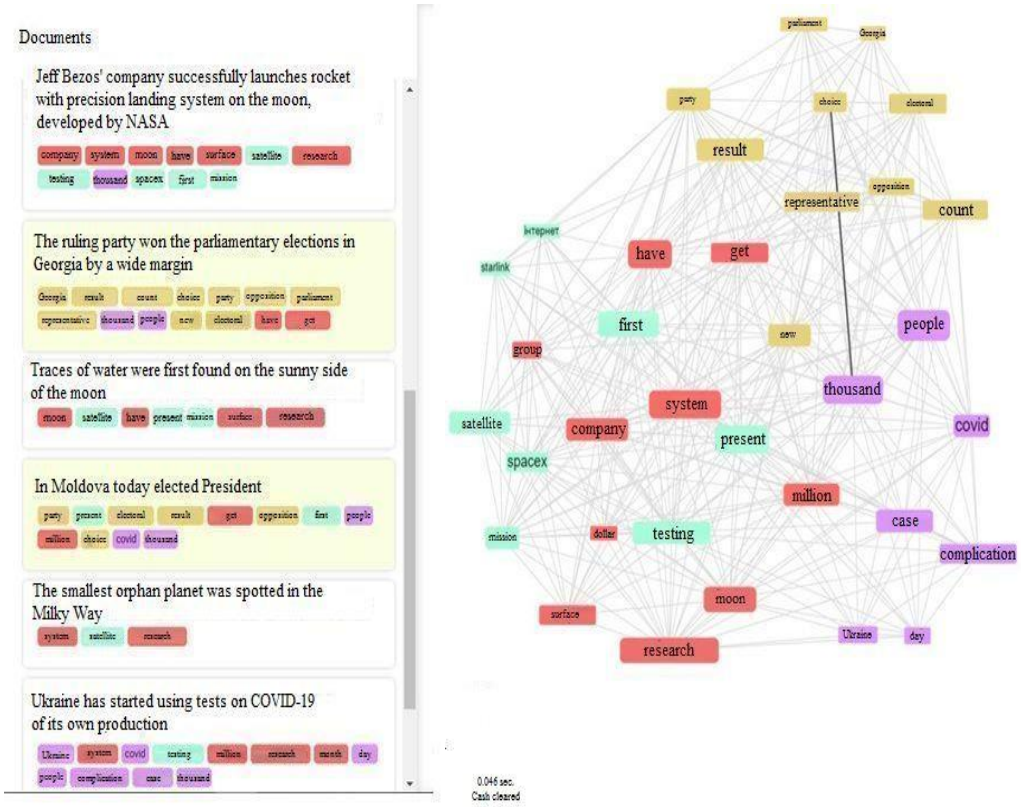

Figure 6: Selection of documents combining defined terms

TF-IDF matrix calculations take more and more time with the increasing number of documents in the corpora.

Also, it only needs to be recalculated if new documents were added or the content of any existing document was changed. Since our user web interface does not allow any of mentioned above actions, we can create a compressed cache file that contains a TF-IDF matrix for certain corpora.

Moreover, all normalized terms that were extracted from the corpora into its "bag of words" can be stored as well in a separate file.

With the help of that, we can dramatically increase the system performance and spend less time on calculations.

6. The efficiency of the developed method and discussion

To study the efficiency of the model, it was decided to measure the speed, taking into account the number of documents in the case.

This technique was chosen because the disadvantage of the method of latent-semantic analysis is the high computational complexity and low speed, which requires recalculation of all values for the whole body in the case of adding a new document.

Thanks to the proposed optimization of saving calculations, the speed of the system has significantly increased.

The diagram of Fig. 7 shows a graph of the dependence of the speed of the model on the number of documents in the body of Ukrainian texts. Speed is measured in seconds.

The blue line shows the performance of the model with computational caching enabled, and the red color shows no caching.

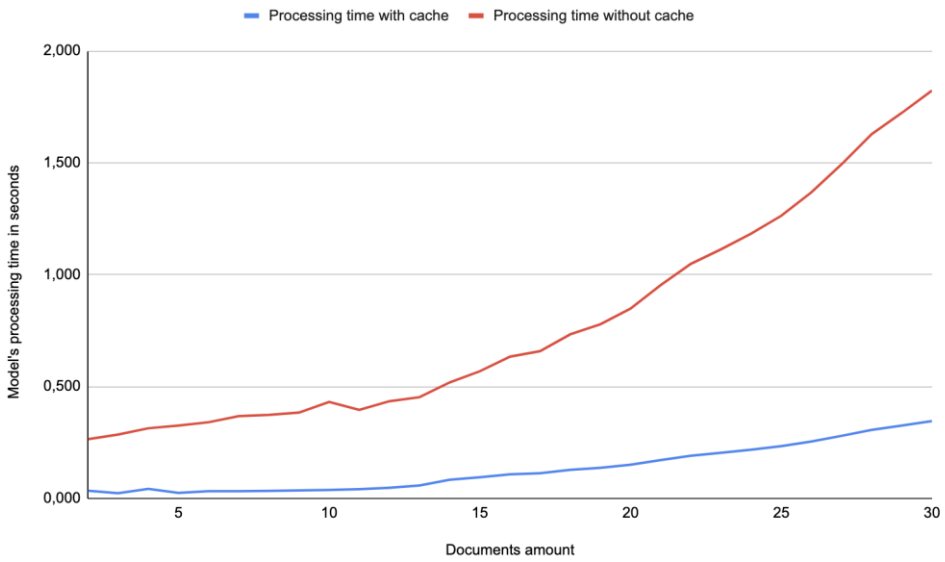

Figure 7: Selection of documents combining selected terms

For a case containing only 2 documents, the speed of the model without the cache is 264 milliseconds, compared to the speed of the model with the cache - 34 milliseconds (ratio ~ 1/7). If you choose a case that stores 30 documents, the speed of the model without a cache is 345 milliseconds, compared to the speed of the model with a cache - 1824 milliseconds (ratio ~ 1/6). Thus, we can conclude that due to the improvement of the method it was possible to increase the speed of the system by about 6 times.Discussion. Several issues are currently under discussion. Firstly, there is no full NER model for the Ukrainian language at this time, and the existing ones contain only a few thousand objects. Therefore, it is necessary to develop a three-dimensional NER model. Secondly, the rather popular NodeJS technology is still widely used for Big Data solutions due to architectural features. Unlike existing clustered systems (Apache Spark, Apache Hadoop), NodeJS solutions are more like a grid system, so they need other implementation approaches.

7. Conclusions

A modified method of semantic analysis of the text has been developed, which enables the analysis of Ukrainian-language content by applying the method of latent-semantic analysis and the Pymorphy2 morphoanalyzer. Semantic capabilities have also been enhanced through the use of the NER model.The software implementation of the system was performed using Node.js technology. The python language was also used to interact with the pymorphy2 morphoanalyzer. To increase the speed of the method, the model remembers the normalized terms and TF-IDF matrices of each analyzed case. The study showed that due to the optimization of calculations significantly increases the speed of the system, about 6 times, and the speed of analysis of the body of 12 documents is reduced from one second to 0.1 seconds.In further development it is necessary to implement the choice of different algorithms for thematic modeling, as well as to conduct a more detailed analysis of unique terms. Parallel data processing can be provided for increasing the speed of volume text corpuses.

8. References

[1] P. Chitra, T.S. Karthik, S. Nithya, J. Jacinth Poornima, J. Srinivas Rao, Makarand Upadhyaya, K. Jayaram Kumar, R. Geethamani, T.C. Manjunath, Sentiment analysis of product feedback using natural language processing, Materials Today (2021). URL: https://www.sciencedirect.com/science/article/pii/S2214785320407795

[2] K. Jindal, R Aron, A systematic study of sentiment analysis for social media data, Materials Today (2021). URL: https://www.sciencedirect.com/science/article/pii/S2214785321000705
[3] B. Ozyurt, M. Ali Akcayol, A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA, Expert Systems with Applications. URL: https://www.sciencedirect.com/science/article/pii/S0957417420309519
[4] Y. Han, M. Moghaddam, Analysis of sentiment expressions for user-centered design, Expert Systems with Applications. URL: https://www.sciencedirect.com/science/article/pii/S0957417421000452
[5] S. Rezaeinia, R. Rahmani, A. Ghodsi, H. Veisi, Sentiment analysis based on improved pre-trained word embeddings, Expert Systems with Applications 117 (2019) 139–14.
[6] X. Zou, J. Yang, W.Zhang, H. Han, Collaborative community-specific microblog sentiment analysis via multi-task learning, Expert Systems with Applications. URL: https://www.sciencedirect.com/science/article/pii/S095741742031015
[7] Lang-uk projects. URL: www.lang.org.ua
[8] N. Kunanets, Y. Oliinyk, D. Kobylynskyi, A. Rzheuskyi, K. Shunevich and V. Tomashevskyi, The model "information gatekeepers" for sentiment analysis of text data. CEUR Workshop Proceedings 2387 (2019) 164–177.
[9] N. Kunanets, Y. Oliinyk, A. Rzheuskyi, O. Artemenko, Sentiment analysis of user responses of tourist services, in: International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), 2019, pp. 502–506. doi: 10.1109/PICST47496.2019.9061470.
[10] M. Lobur, A. Romaniuk, M. Romanyshyn Defining an approach for deep sentiment analysis of reviews in Ukrainian. Bulletin of the National University "Lviv Polytechnic" 747 (2012) 124-130.
[11] M. Dilai, O. Levchenko, Attitudes toward feminism in Ukraine: a sentiment analysis of tweets. Advances in Intelligent Systems and Computing (AISC) 871 (2018) 119–131.
[12] M. Dilai, Y. Onukevych, I. Dilai, Sentiment analysis of the US and Ukrainian presidential speeches. URL: http://ena.lp.edu.ua:8080/handle/ntb/42572.
[13] U. Kryva, M. Dilai, Automatic Detection of Sentiment and Theme of English and Ukrainian Song Lyrics, in: 14th International Conference on Computer Sciences and Information Technologies, Lviv, 2019. pp. 20-23.
[14] Morphosyntax analyzer. URL: www.mova.institute/analyzer
[15] M. Korobov, Morphological analyzer and generator for Russian and Ukrainian languages. International Conference on Analysis of Images, Social Networks and Texts. Springer, 2015.
[16] D. Myhal, Y. Oliinyk, Modern tools and methods of semantic analysis of Ukrainian texts, in: Proceedings of the 10th International scientific and practical conference “Topical issues of the development of modern science”, Publishing House “ACCENT”, Sofia,2020, pp. 544-557.
[17] Regionally annotated corpus of the Ukrainian language. URL: http://uacorpus.org/
[18] SentiStrength. URL: http://sentistrength.wlv.ac.uk
[19] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social Web. Journal of the American Society for Information Science and Technology. New York: John Wiley & Sons 63 (1) (2012) 163-173.
[20] Lucene_uk. URL: www.github.com/arysin/lucene_uk
[21] Thorsten Brants, Francine Chen, Ioannis Tsochantaridis, Topic-based document segmentation with probabilistic latent semantic analysis, in: Proceedings of the eleventh international conference on Information and knowledge management (CIKM '02). Association for Computing Machinery, New York, NY, USA, 2002, pp. 211–218.
[22] D. Myhal, Y. Oliinyk, Semantic text analysis methods and tools, in: Proceedings of the Information systems and management technologies, Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, 2020, pp. 97-101.
[23] Enhanced LSA method with Ukraine Language Support. URL: https://github.com/DmytroMyhal/Dissertation