Засобами мови програмування Python розроблена комп'ютерна система (КС) для генерації семантичного шаблону групи документів методом латентне-семантичного аналізу (ЛСА). Система утримує вісім програмних модулів, кожний з яких виконує один етап ЛСА. Унікальними є модуль контролю частотної матриці слів-документів та модуль виміру семантичної відстані між документами шаблону. Адаптація КС до змісту та структури шаблону документів здійснюється зміною складу працюючих модулів. З використанням КС досліджено вплив на результати генерації шаблонів методом ЛСА таких факторів, як: нормалізація частотної матриці, виключення один раз вживаних слів, виключення документів, не пов'язаних зі спільними словами,обрання міри відліку семантичної відстані між документами

Ключові слова: метод латентне-семантичного аналізу, комп'ютерна система, семантична відстань, семантичний шаблон, програмний модуль, засоби мови програмування Python

Средствами языка программирования Python разработана компьютерная система (КС) для генерации семантического шаблона группы документов методом латентно-семантического анализа (ЛСА). Система содержит восемь программных модулей, каждый из которых выполняет один этап ЛСА. Уникальными являются модуль контроля частотной матрицы слов-документов и модуль измерения семантического расстояния между документами шаблона. Адаптация КС к содержанию и структуре шаблонов документов осуществляется изменением набора работающих модулей. С использованием КС исследовано влияние на результаты генерации шаблонов методом ЛСА таких факторов, как: нормализация частотной матрицы, исключение использованных один раз слов, исключение документов, не связанных с общими словами, выбор меры отсчета семантического расстояния между документами

Ключевые слова: метод латентно-семантического анализа, компьютерная система, семантическое расстояние, семантический шаблон, программный модуль, средства языка программирования Python

UDC 004.62
DOI: 10.15587/1729-4061.2016.73551

DEVELOPMENT OF A COMPUTER SYSTEM FOR GENERATING SEMANTIC TEMPLATE OF A GROUP OF DOCUMENTS BY USING LATENT SEMANTIC ANALYSIS

Y. Taranenko
Doctor of Technical Sciences, Professor*
E-mail: lika15k@yandex.ua
M. Kabanova
Candidate of Philological Science, Associate Professor*
E-mail: marina.kabanova.00@bk.ru
*Department of Applied Linguistics and Methods in Foreign Language Teaching
Alfred Nobel Dnipropetrovsk University
Sicheslavska naberezhna str., 18, Dnipropetrovsk, Ukraine, 49000

1. Introduction

Semantic text analysis is one of the key problems of both the theory of artificial intelligence systems, related to natural language processing (NLP) and computational linguistics. Results of semantic analysis can be used to solve problems in the fields such as psychiatry (for diagnosing patients), political science (predicting election results), trade (analysis of demand for certain goods on the basis of the product reviews), philology (analysis of manuscripts), search engines, machine translation.

Despite a demand in almost all spheres of human life, semantic analysis is one of the most complex mathematical problems.

An important task is developing software for automatic processing of speech and text data to improve information retrieval systems with advanced features that use natural language queries.

The research of the method of latent semantic analysis (LSA) allows automating a number of text data processing cycles, including document indexing by thematic groups, plagiarism detection, forming databases of natural language queries. Therefore, software implementation, especially such that can increase the resolution of the method is an extremely urgent task of scientists and information technology developers.

2. Literature review and problem statement

There are several methods of semantic text analysis, which can be divided into two groups [1]: linguistic analysis;

© Y. Taranenko, M. Kabanova, 2016

statistical analysis. LSA, or latent semantic indexing (LSI), is one of the most effective statistical approaches. The LSA method allows retrieving context-dependent meanings of words using statistical processing of large sets of text data. It is based on the principles of analysis of major components applied to the creation of artificial neural networks. The set of all contexts, in which the word is either found, or not, imposes a lot of mutual constraints, which allow determining the similarity of semantic meanings of words and sets of words.

The drawback of the method in the implementation in computer systems is a resolution reduction because individual words are removed from the text to preserve the computation speed while increasing input data. This problem occurs in processing the results of singular value decomposition of a frequency matrix (SVD transform) when full vectors of words and documents are used for determining the semantic distance of words and documents.

Computer implementation of semantic analysis of the Arabic text is shown in [2]. Given the Arabic language features, the authors propose correlation noise filtering instead of SVD transform of the frequency matrix. However, singular value decomposition of the frequency matrix (SVD) is more versatile and better investigated.

An interesting development of the semantic analysis computer system is presented in [3]. The authors propose to hold the LSA of signatures to graphic documents, posted on the network. This shows the need to improve computer implementation of the LSA method for short names and information messages.

According to the authors, the project presented in [4] could be a conceptual breakthrough in computer implementation of LSA. It is proposed first to develop a semantic template that represents semantic features of the content of each document in a special case in the form of a document-term matrix. Hidden LSA is added to the semantic vector space graph. Then the query content can be identified by the cluster in the semantic space graph. The effectiveness of the model will be high. But first, it is necessary to work out the computer implementation of the semantic template allowing for the LSA features.

Since the implementation of the above project in search engines and systems for determining the identity of texts has significant commercial benefits, publications on this subject in printed literature are unavailable.

It is advisable to consider publications of some developers on the Internet. These publications should take into account that the non-working computer program listing was published for the aforementioned reason, but the results in the form of numerical data, as verified by the authors of the present work are real. The advantages of the Python programming language in the development of computer systems (CS) to work with text data are shown in [5]. This is primarily the Python-based Natural Language Toolkit (NLTK: nltk.org).

Developments of computer systems, known to the authors used the LSA method partially without the frequency matrix normalization and without an interface. Thus, news headlines, forming three groups with common sense were used for the input data in [6]. So-called stop words that do not reflect the content, and individual words that occur once were selected from the headlines. For this, as well as for the frequency matrix construction, word stems that have been determined by the Porter algorithm were used [7]. The author of [6] gives the results of SVD transform of the frequency matrix. The drawback is the lack of the frequency matrix normalization and the uncertainty of the practical application of the results.

A computer program for the LSA implementation in the Python programming language is given in [8]. The program contains the TF-IDF module [9], but it is used only to determine the so-called basic word, the selection method of which is not provided by the author. The downside of [8] is the lack of the frequency matrix normalization and the uncertainty of practical application of the results. The absence of any interface and links to the nltk Python library does not allow determining the program operability. Although the work [10] is aimed at practical use – determination of authorship authenticity of scientific papers by the LSA method, the work does not provide any data on the computer system that would have the means to study this method.

Even a very limited number of the given periodicals indicates that the studies of LSA should be continued with the advent of new application areas. These are, in particular, developments of computer systems (CS) to create semantic templates of document groups or clustering of short document titles by the LSA method. In addition, the use of the NLTK Python library with the graphical user interface is promising for these developments. The graphical user interface is required to identify groups of documents on a specific subject not only by numerical indices of semantic distances of words and documents, but in their graphical representation for visual observation and handling. According to the publications provided, the semantic template increases the efficiency of search engines, which is a priority at the present development stage of information technology.

3. Research goals and objectives

The goal of the paper is to study the application of the LSA method to create semantic templates of groups of documents.

To achieve this goal, the following problems were identified:

– to develop the CS with flexible structure with the graphical user interface, suitable for the study of the LSA method using the Python programming language and the NLTK library by means of stemming and tokenization;

– to examine the impact of the frequency matrix normalization on the LSA results, the use of the words found in all documents only once, the use of the cosine of the difference of angles between the vector of the group of basic words and vectors of documents to account for the semantic distance;

– to develop an algorithm for cyclic removal of documents unrelated in content from the frequency matrix and apply it in the CS.

4. Means and methods of research of LSA of text documents

4. 1. Functional diagram of LSA research

The functional diagram necessary for development (CS) is shown in Fig. 1. The diagram is divided into modules and units with the functions that are implemented by

means of the Python programming language 3.4, namely the libraries nltk, numpy, scipy. The units enable or disable the modules, which allows examining the impact of removal of individual words, the effect of the frequency matrix normalization procedure, the influence of the method for determining a measure of distance between documents starting from the basic word. Besides, in the event of the degenerate frequency matrix, consistent removal of unrelated documents is made automatically, which greatly simplifies the analysis.

For example, let us consider the operation results of the construction module of the frequency matrix of the word stem-document type. Stemming procedure is first performed according to the Porter algorithm. This is done in the Python by means of the Snowball Stemmer module. A code snippet of this procedure consists of three rows. The module is imported in the first row, the language for stemming in the stem variable is set in the second row for one of the following abbreviations of languages - 'danish', 'dutch', 'english', 'finnish', 'french', 'german', 'hungarian', 'italian','norwegian', 'porter', 'portuguese', 'romanian', 'russian', 'spanish', 'swedish'. In the third row, we get the word stem in a variable stemword for the word in the word variable:

```
from nltk.stem import SnowballStemmer
stemmer=SnowballStemmer(stem)
stemword=stemmer.stem(word)
```

The matrix has the form:

```
wikileaks {[ 1. 0,0 0,0 1. 0,0 1. 0,0 1. 0,0 ]}
арестова {[ 0,0 0,0 0,0 1. 0,0 0,0 0,0 1. 0,0 ]}
великобритан {[ 0,0 0,0 0,0 1. 0,0 0,0 0,0 1. 0,0 ]}
вручен {[ 0,0 0,0 1. 0,0 1. 0,0 0,0 0,0 1.]}
нобелевск {[ 0,0 0,0 1. 0,0 1. 0,0 0,0 0,0 1.]}
основател {[ 1. 0,0 0,0 1. 0,0 1. 0,0 1. 0,0 ]}
полиц {[ 1. 0,0 0,0 0,0 0,0 0,0 0,0 1. 0,0 ]}
прем {[ 0,0 0,0 1. 0,0 1. 0,0 0,0 0,0 1.]}
прот {[ 0,0 1. 0,0 0,0 0,0 0,0 1. 0,0 0,0 ]}
стран {[ 0,0 0,0 1. 0,0 0,0 0,0 1. 0,0 0,0 ]}
суд {[ 0,0 1. 0,0 0,0 0,0 1. 0,0 0,0 0,0 ]}
сша {[ 0,0 1. 0,0 0,0 0,0 0,0 1. 0,0 0,0 ]}
церемон {[ 0,0 0,0 1. 0,0 1. 0,0 0,0 0,0 0,0 ]}
```

Using the word stem allows you to allocate correctly the word in documents, so a change of the case does not mean a change of the word.

4. 2. Research software interface

The documents subject to the analysis are shown in the first top field of the main form (Fig. 2), stop and individual words and all stages of the frequency matrix transform in the second, the LSA results as pairs of documents with common words and the growing distance between the pairs in the third. In addition, the program builds a graphical representation of documents and words in a two-dimensional semantic space.

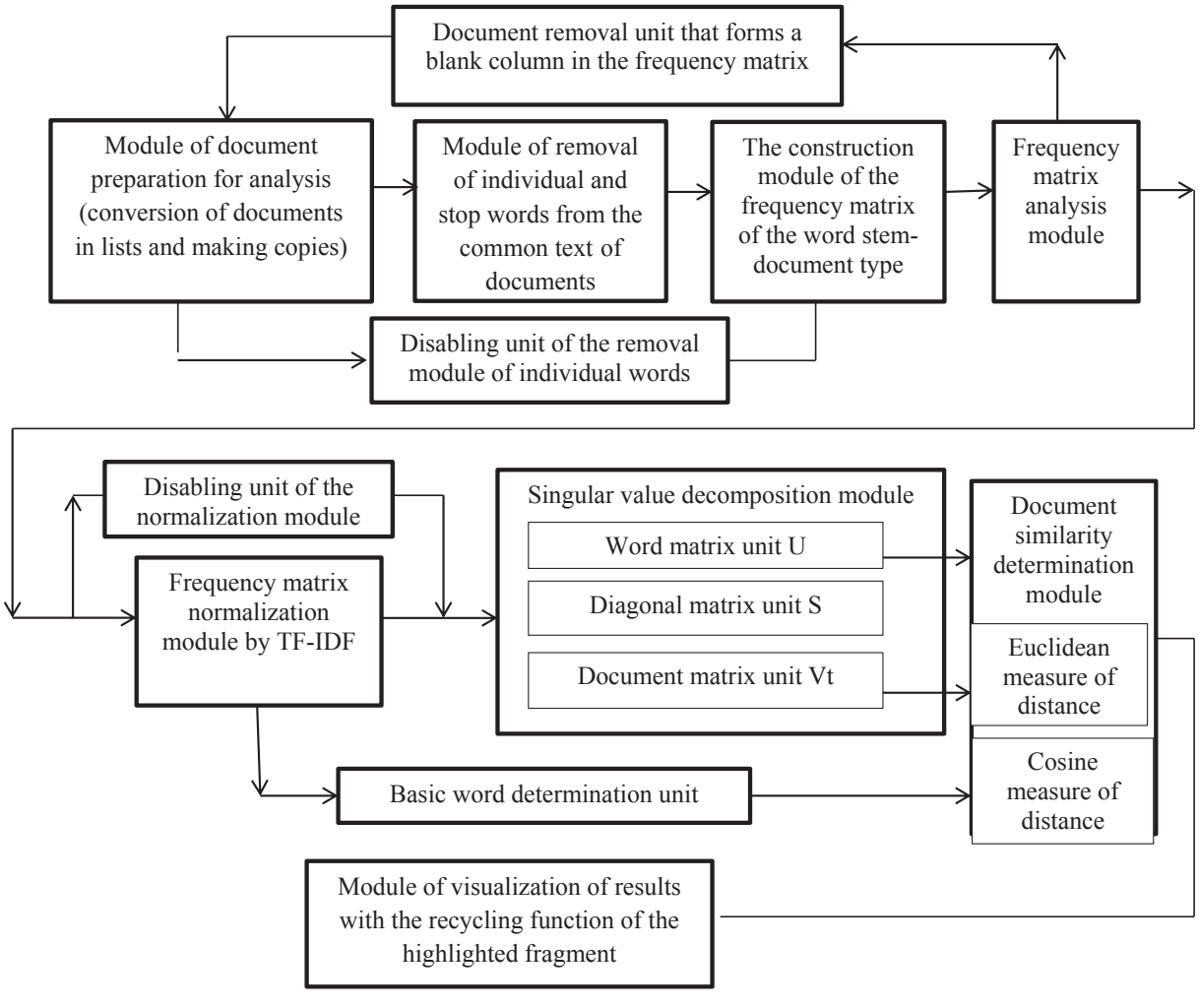

Fig. 1. Functional diagram of comparative research of LSA

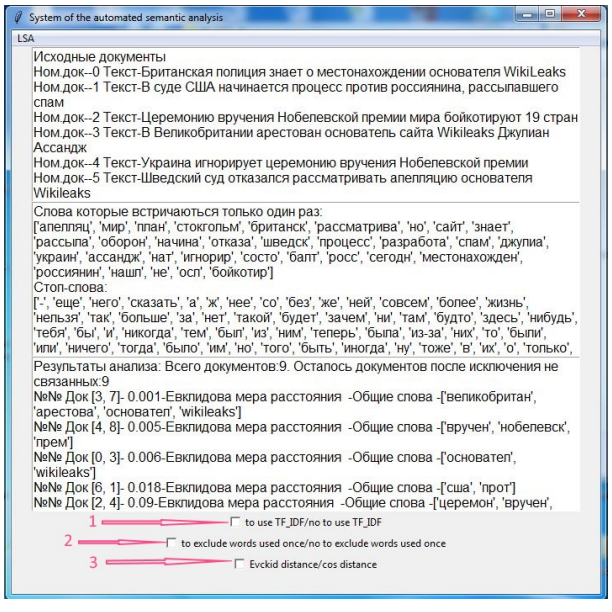

Fig. 2. The basic form of the CS interface: 1 – without the frequency matrix normalization by the TF-IDF method; 2 – without exception of individual words; 3 – change of the method of the distance computation in two-dimensional semantic space x Euclidean with cosine

The interface ensures all the functions provided by the functional diagram (Fig. 1).

5. Comparison of the LSA results with and without the frequency matrix normalization by the TF-IDF method

Comparison of the results of the impact of the frequency matrix normalization on the LSA results was carried out on the same set of documents (the set is given in [6, 8]) in two stages with the normalization module disabled and enabled (Fig. 1) and a checkbox 1 (Fig. 2). The results were included in the first and second part of Table 1 respectively.

Analysis of data (Table 1) and graphs (Fig. 3, 4) of the semantic space leads to the conclusion about the effect of the TF-IDF frequency matrix normalization on the LSA results:

– normalization allows you to better group the data according to the content – three separate groups of word stems and documents (Fig. 3, 4, Table 1);

– normalization allows you to single out the words unrelated to documents and group related documents (Fig. 4, Table 1).

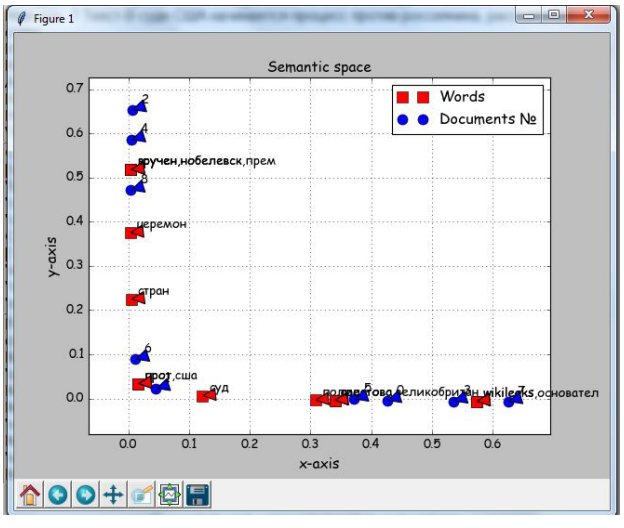

Fig. 3. Two-dimensional graph of semantic space obtained according to the proposed functional diagram of LSA without the frequency matrix normalization

Completing these comparative studies, it should be noted that the proof of their authenticity is a complete identity of the results of the first stage without the frequency matrix normalization to the results presented in [6, 8]. However, the results of the study of the effect of the frequency matrix normalization on the LSA results are not available, so the comparative analysis is useful to developers (CS).

Table 1

The results of studies of the effect of the frequency matrix normalization on the template resolution

№ p/w	№№ Document number	Euclidean measure of distance between documents	Word stems common to both documents
The semantic template of three groups of news headlines without the frequency matrix normalization. Basic word – 'против' (Fig. 3)			
1	3, 7	0.013	'великобритан', 'арестова', 'основател', 'wikileaks'
2	1, 6	0.041	'сша', 'прот'
3	4, 2	0.049	'церемон', 'вручен', 'нобелевск', 'прем'
4	0, 5	0.0 055	'основател', 'wikileaks'
5	8, 4	0.0 078	'вручен', 'нобелевск', 'прем'
6	7, 0	0.0 095	'полиц', 'основател', 'wikileaks'
7	6, 8	0.0 173	–
8	5, 1	0.0 327	'суд'
The semantic template of three groups of news headlines without the TF IDF frequency matrix normalization. Basic word – 'wikileaks' (Fig. 4)			
1	3, 7	0.0 001	'великобритан', 'арестова', 'основател', 'wikileaks'
2	4, 8	0.005	'вручен', 'нобелевск', 'прем'
3	0, 3	0.006	'основател', 'wikileaks'
4	6, 1	0.018	'сша', 'прот'
5	2, 4	0.09	'церемон', 'вручен', 'нобелевск', 'прем'
6	5, 0	0.154	'основател', 'wikileaks'
7	7, 2	0.187	–
8	1, 5	0.266	'суд'

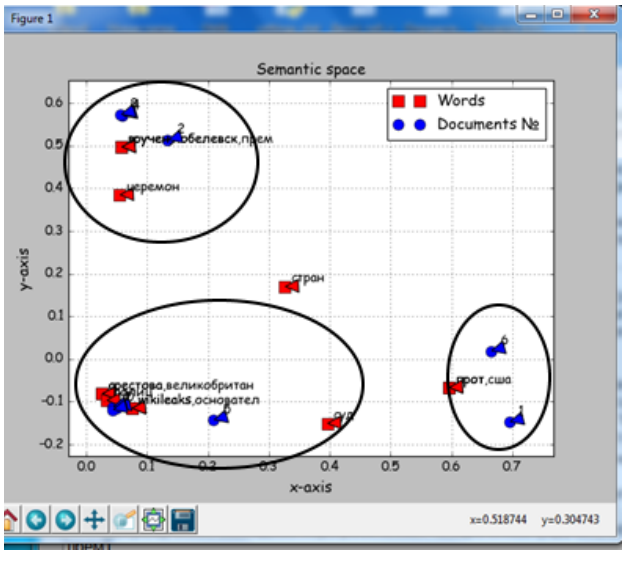

Fig. 4. Two-dimensional graph of semantic space with a clear grouping of words and documents obtained according to the functional diagram with the frequency matrix normalization

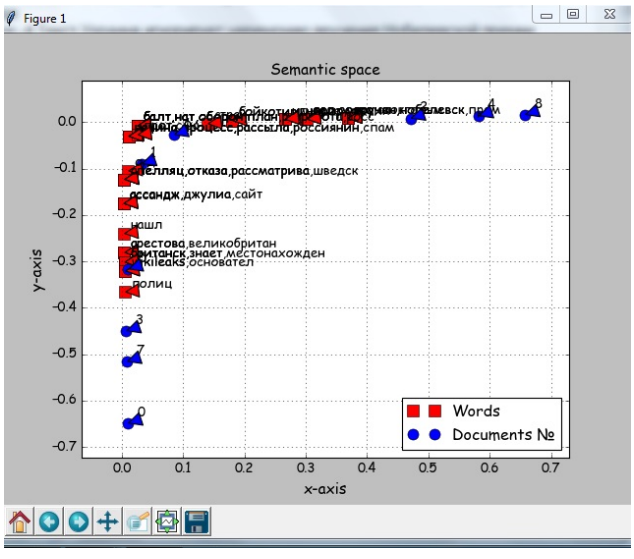

Fig. 5. Two-dimensional graph of semantic space without exception of individual words

6. Analysis of LSA results without exception of individual words

The studies were carried out in one stage on the set of documents given in [6, 8] with the disabled removal unit of individual words (Fig. 1) and marked checkbox 2 (Fig. 2). The results are listed in Table 2.

Table 2

The results of studies of keeping individual words for the template resolution

The semantic template of three groups of news headlines with the TF IDF frequency matrix normalization. Basic word – 'wikileaks' (Fig. 5)			
№ p/w	№№ Document number	Euclidean measure of distance between documents	Word stems common to both documents
1	6, 0	0.024	–
2	7, 1	0.037	–
3	3, 7	0.066	'великобритан', 'арестова', 'основател', 'wikileaks'
4	4, 8	0.067	'вручен', 'нобелевск', 'прем'
5	1, 6	0.072	'сша', 'прот'
6	2, 4	0.096	'церемон', 'вручен', 'нобелевск', 'прем'
7	5, 3	0.123	'основател', 'wikileaks'
8	0, 2	0.242	–

Based on the comparison of the data in Tables 1, 2, we can conclude that it is not advisable to keep individual words when creating semantic templates of short documents due to the emergence of pairs of documents without common words and reduced resolution of LSA.

Despite the fact that the exception of individual words does not affect the LSA results, the considered function (CS) may be useful for finding the specified words in semantic space. Indeed, the graph (Fig. 5) in this case is overloaded, but the matplotlib.py Python module provides the "magnifying glass" function, which allows examining the scaled parts of the graph.

7. Development and research of the upgraded cosine measure of distance between words and documents in two-dimensional semantic space

Let us first consider the method of determining the proposed basic word. To do this, we present the normalized matrix of the considered template. The matrix, normalized by the TF-IDF method: rows (word stems) – 13, columns (documents) – 9.

wikileaks {[0,27 0,0 0,0 0,2 0,0 0,27 0,0 0,16 0,0]} – The sum of the row 0.9

арестова {[0,0 0,0 0,0 0, 38 0,0 0,0 0,0 0, 3 0,0]} – The sum of the row 0.68

великобритан {[0,0 0,0 0,0 0, 38 0,0 0,0 0,0 0, 3 0,0]} – The sum of the row 0.68

вручен {[0,0 0,0 0, 22 0,0 0, 27 0,0 0,0 0,0 0, 37]} – The sum of the row 0.86

нобелевск {[0,0 0,0 0, 22 0,0 0, 27 0,0 0,0 0,0 0, 37]} – The sum of the row 0.86

основател {[0,27 0,0 0,0 0, 2 0,0 0, 27 0,0 0, 16 0,0]} – The sum of the row 0.9

полиц {[0,5 0,0 0,0 0,0 0,0 0,0 0,0 0,0 3 0,0]} – The sum of the row 0.8

прем {[0,0 0,0 0,22 0,0 0, 27 0,0 0,0 0,0 0, 37]} – The sum of the row 0.86

прот {[0,0 0,5 0,0 0,0 0,0 0,0 0,5 0,0 0,0]} – The sum of the row 1.0

стран {[0,0 0,0 0,3 0,0 0,0 0,0 0, 5 0,0 0,0]} – The sum of the row 0.8

суд {[0,0 0,5 0,0 0,0 0,0 0,5 0,0 0,0 0,0]} – The sum of the row 1.0

сша {[0,0 0,5 0,0 0,0 0,0 0,0 0,5 0,0 0,0]} – The sum of the row 1.0

церемон {[0,0 0,0 0, 3 0,0 0,38 0,0 0,0 0,0 0,0]} – The sum of the row 0.68

The maximum sum of the rows in the normalized matrix appears for three word stems, namely 'суд', 'сша', 'прот'. Let us choose the coordinates xw and yw of basic word stems from the first two columns of the orthogonal matrix U of the singular transform of the presented normalized matrix.

xw yw

прот{[–0.5957 0.0666]}
стран {[–0.3264 –0.1699]}
суд {[–0.396 0.1506]}
сша {[–0.5957 0.0666]}

Analysis of coordinates of basic word stems indicates that at least two stems 'сша', 'прот' of basic words have the same coordinates xw and yw. This is an important finding because set phrases better reflect the meaning of the document.

Therefore, to increase the resolution (LSA), the cosine measure should be adjusted only for one two-dimensional vector of the basic word $A(xw,yw)$ and two-dimensional vectors of documents $V_i(x_i,y_i)$.

The basic word (a group of words or phrases) is determined by the maximum sum of the corresponding row of the normalized frequency matrix (as shown above). Cosines of the angles between the basic word and each document, sorted in ascending order are determined by the ratio:

$$\cos(\alpha_i)=\frac{xw\cdot x_i+yw\cdot y_i}{\sqrt{wx\cdot wx+wy\cdot wy}\cdot\sqrt{x_i\cdot x_i+y_i\cdot y_i}}. \quad (1)$$

Then the cosine of the angle between the vectors of documents sorted in ascending order relative to the basic word vector is determined by the ratio:

$$\cos(\alpha_{i+1}-\alpha_i)=\cos(\alpha_{i+1})\cdot\cos(\alpha_i)+ +\sqrt{1-\cos(\alpha_{i+1})^2}\cdot\sqrt{1-\cos(\alpha_i)^2}. \quad (2)$$

Cosine measures of distance $K_{i+1,i}$ between pairs of documents will be determined by the ratio:

$$K_{i+1,i}=1-\cos(\alpha_{i+1}-\alpha_i). \quad (3)$$

Analysis of the data in Tables 1–3 indicates that the proposed measure of distance in the ratios (1)–(3) best represents the proximity of documents. Also, it reveals the hidden meaning (Table 3) when the words are common and the contents of documents are different.

8. Analysis of the case of the degenerate frequency word stem-document matrix

Prerequisite (LSA) is a condition of the frequency matrix dimension, namely the number of rows (word stems) should be greater than or equal to the number of columns. Due to the fact that for the above reasons it is necessary to remove individual or stop words, it may happen that at least one column of the frequency matrix is filled with zeros, that is, there will be more columns than rows. To avoid such a probability, the functional diagram (Fig. 1) provides feedback from the frequency matrix analysis module to the module of document preparation for analysis. The feature of this process is that after the column removal, the whole data processing cycle is started over. There are new individual words, the removal of which, in turn, leads to the emergence of the next column with all zeros until the basic condition is met – the number of rows should be greater than or equal to the number of columns.

We give a real example of cyclicality that occurred in our CS with the frequency matrix A.

Zero cycle A_0 (19×18) – the first 19 words of 18 documents.

The first cycle A_1 (11×14) – the second 11 words of 14 documents.

The second cycle A_2(11×11) – the third 11 words of 11 documents.

The third cycle A_3(11×10) – the fourth 11 words of 10 documents.

According to the developed algorithm, processing of the matrix is terminated at the stage of analysis of the prepared frequency matrix in the event of a mismatch between the number of words – m and documents – n, namely when n>m. The document, the column of which in the frequency matrix contains only zeros is removed from the database. The frequency matrix preparation process is started over. All the words are selected from the remaining documents, stop words that are not responsible for the contents and words that occur only once are removed. Then stemming of the remaining words is carried out.

Table 3

The results of studies of the effect of the proposed measure of distance between words and documents on the semantic template resolution

The semantic template of three groups of news headlines with the TF IDF frequency matrix normalization. Basic word – 'против'			
№ p/w	№№ Document number	Euclidean measure of distance between documents	Word stems common to both documents
1	8, 4	0.000	'вручен', 'нобелевск', 'прем'
2	7, 3	0.000	'великобритан', 'основател', 'wikileaks', 'арестова'
3	3, 0	0.000	'основател', 'wikileaks'
4	6, 1	0.001	'сша', 'прот'
5	4, 2	0.011	'церемон', 'вручен', 'нобелевск', 'прем'
6	2, 7	0.050	–
7	5, 6	0.060	–
8	0, 5	0.153	'основател', 'wikileaks'

The words are allocated according to the documents. The frequency matrix, which is subject to re-examination is constructed. The process continues until the condition m>=n is met. There may be several cycles.

The reason is that new words that are used once may appear after the removal of another document and related words. The removal of these words violates the ratio between n and m. After completion of cycles, the frequency matrix processing continues, that is normalization, singular value decomposition and analysis of results are performed. Thus, the condition m>n is met and the LSA automaticity is preserved.

9. Conclusions

1. The CS was developed by means of the Python programming language to generate a semantic template of a group of documents by the LSA method. The system contains eight software modules, each performs one stage of the LSA. The control module of the frequency word-document matrix and the measuring module of semantic distance between the template documents are unique. Adjustment of CS to the contents and structure of the document templates is performed by changing a set of modules.

2. It is proved that the frequency matrix normalization enhances the resolution of the semantic template generated by using the LSA.

3. It is proved that the removal of individual words improves the resolution of the generated semantic template and does not affect the semantic contents.

4. Application of semantic proximity of documents, the cosine of the difference of angles between the vector of a group of basic words and vectors of documents for evaluation allows increasing the resolution of the generated semantic template.

5. To ensure the continuity of the LSA, the module of the frequency matrix analysis for compliance of excess (or equality) of the number of words over the number of documents was introduced in the CS. In the event of a mismatch, the module starts over the LSA process with a new set of words and documents.

References

1. Landauer, T. K. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge [Text] / T. K. Landauer, S. T. Dumais // Psychological Review. – 1997. – Vol. 104, Issue 2. – P. 211–240. doi: 10.1037//0033-295x.104.2.211
2. Hanane, F. Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering [Text] / F. Hanane, L. Abdelmonaime, A. Said // International Journal of Data Mining & Knowledge Management Process. – 2013. – Vol. 3, Issue 1. – P. 79–95. doi: 10.5121/ijdkp.2013.3107
3. Kesorn, K. Semantic Restructuring of Natural Language Image Captions to Enhance Image Retrieval [Text] / K. Kesorn, S. Poslad // Journal of Multimedia. – 2009. – Vol. 4, Issue 5. – P. 284–297. doi: 10.4304/jmm.4.5.284-297
4. Amudaria, S. Design of Content-Oriented Information Retrieval by Semantic Analysis [Text] / S. Amudaria, S. Sasirekha // International Journal of Computer Science and Information Security. – 2011. – Vol. 9, Issue 1. – P. 92–97.
5. Wang, Z. A Python-based Interface for Wide Coverage Lexicalized Tree-adjoining Grammars [Text] / Z. Wang, H. Zhang, A. Sarkar // The Prague Bulletin of Mathematical Linguistics. – 2015. – Vol. 103, Issue 1. – P. 139–159. doi: 10.1515/pralin-2015-0008
6. Latent semantic analysis [Electronic resource]. – Available at: https://habrahabr.ru/post/110078/ (Last accessed: 30.04.2016).
7. Sheetal, A. Measuring Semantic Similarity between Words Using Web Documents [Text] / A. Sheetal, S. Sushma // International Journal of Advanced Computer Science and Applications. – 2010. – Vol. 1, Issue 4. – P. 132–154. doi: 10.14569/ijacsa.2010.010414
8. Latent semantic analysis and search on Python [Electronic resource]. – Available at: https://habrahabr.ru/post/197238/ (Last accessed: 30.04.2016).
9. Reena, K. Semantically Detecting Plagiarism for Research Papers [Text] / K. Reena, M. Preeti, V. Chavan, K. Jadhav // International Journal of Engineering Research and Applications. – 2013. – Vol. 3, Issue 3. – P. 77–80.
10. Kolyada, A. C. Authenticity of authorship of scientific publications using latent semantic analysis [Text] / A. C. Kolyada, V. D. Godunsky // Eastern-European Journal of Enterprise Technologies. – 2014. – Vol. 3, Issue 2 (69). – P. 36–40.