IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 1, February, 2023, pp. 165-170
Journal homepage: http://iieta.org/journals/ria

Conceptual Model for Automatic Proofreading of Technical Documents

Zhanna S. Ixebayeva[1*], Kabylda Jetpisov[1], Aigul B. Medeshova[2], Akmaral Kh. Kassymova[3]

[1] Department of Information Systems, L.N. Gumilyov Eurasian National University, 2 Satpayev Str., Astana 010008, Republic of Kazakhstan
[2] Department of Computer Science, Makhambet Utemisov West Kazakhstan University, 162 Dostyk Str., Uralsk 090000, Republic of Kazakhstan
[3] Higher School of Information Technology, Zhangir Khan West Kazakhstan Agrarian Technical University, 51 Zhangir Khan Str., Uralsk 090009, Republic of Kazakhstan

Corresponding Author Email: zhanna.ixebayeva@aol.com

https://doi.org/10.18280/ria.370120

Received: 11 November 2022
Accepted: 5 January 2023

Keywords:
text, structure, analysis, method, semantic search

ABSTRACT

This study deals with a set of issues related to the development of a conceptual model for automatic proofreading of technical documentation. The purpose of this study is to investigate the prospects for creating the software for automatic proofreading of text documents with an assessment of the prospects for its subsequent implementation in various areas of scientific cognition and in activities of various educational institutions. The methodological approach is a combination of a systematic study of modern algorithms for checking technical documents with an analysis of the prospects for building a concept for creating an optimal model for automatic document proofreading. The main results of this study should be the definition of the main areas for the development of issues for the creation of the concept under consideration and identification of the constituent elements of the conceptual model for automatic proofreading of technical documentation, which is important from the standpoint of ensuring the proper level of quality of functioning of such a system. The prospects for further research in this area are determined by the relevance of the stated topic conditioned by the urgent need to develop and implement an effective system for verifying technical documents as soon as possible.

1. INTRODUCTION

The development of universal tools for analysing texts of literary and business texts is one of the most difficult tasks in the field of data mining. A popular line of development of algorithms for processing text documents today is considered to be the use of machine learning methods that allow solving problems of natural language processing. Machine learning algorithms commonly used for natural language processing are Porter's algorithm, BERT (Bidirectional Encoder Representations from Transformers), CBOW (Continuous Bag of Words), Word2vec, GloVe, etc. Porter's algorithm does not use word stem bases but works by successively applying a series of suffix and suffix pruning rules. BERT is based on a transformer, for pre-training natural language processing. The logic of the CBOW architecture is very simple: predict a word depending on the context in which the word is found. Word2Vec includes a set of algorithms for calculating vector representations of words, assuming that words used in similar contexts are semantically close. The algorithm minimizes the difference between the product of vectors of words and the logarithm of the probability of their joint occurrence using stochastic gradient descent. For example, it turns out to link together different satellites of one planet or the postal code of a city with its name [1].

The problem of the optimal search for information is one of the key problems in the field of computer science. The development of most software products sooner or later leads to the implementation of mechanisms for adding, saving, and receiving information for its subsequent processing. The main solution to this problem is a variety of database systems that perfectly performs these tasks at the software level. However, if the system functionality requires working directly with a user query, which often consists of several criteria, then the processing of the results obtained falls entirely on the shoulders of the user [2]. In such a situation, the problem of creating a conceptual model for automatic proofreading of technical documents that optimally meets the needs of users becomes particularly relevant.

In the 21st century, the progress of IT is evident, which determines the possibility of choosing the optimal conditions for creating the concept of an automatic text processing system of any degree of complexity and orientation. Of particular importance in this context is the use of computer text corpora designed to solve the problem of automating this process. A computer corpus is an array of natural texts of a modern language (both written and spoken), presented in a digital medium and properly ordered for use for scientific or practical purposes [3]. The use of text corpora allows identifying the lexical and grammatical compatibility of words, their frequency, patterns of word usage, and use them as a source of additional information about the subject area and the use of the term [4]. Today, such electronic text arrays as the National Corpus of the Russian Language (NCRL), the British National Corpus (BNC), the Bank of English, and the American National Corpus (ANC) are widely known in professional

circles, and their use is advisable when working with general-purpose texts [5].

Notably, ontologically controlled information systems are currently one of the most important branches of the development of intelligent information systems, which is an extremely important aspect from the standpoint of the prospects for creating an automatic system for checking technical documents. The application of ontologically managed information systems based on the automation of text editing ensures the implementation of such processes as structuring and systematization of information, integration of distributed information models and systems based on the use of semantic properties, aggregation of various information resources, visualization of necessary information and transformation of the knowledge search process into modern access technology to the chosen field of research. The features of their construction are closely related to the development of the theory and fundamentals of designing automatic text proofreading models and to the development of the theoretical foundations and design methodology, including the formal approach, fundamental principles and mechanisms, the generalised architecture and structure of the system, the formal model and methodology of designing domain ontology, the formal model of knowledge representation, generalised algorithms of knowledge processing procedures, etc. [6]. In this context, the complex formulation of specific tasks should increase the importance of automatic text-checking programmes and facilitate the search for ways to solve them.

The existing methods of data mining allow processing large arrays of text documents (more than 1 million texts) to identify various parameters of the documents included in this array, as well as the patterns that characterise their totality. Since these algorithms involve the extraction of a wide range of diverse characteristics from texts (which is often a complex task in itself, the solution of which involves the use of complex and not always fast-acting algorithms), there is a need to store the extracted characteristics (along with the documents themselves) in the reference and information collection of the created software system [7]. At the same time, it should be noted that there are various methods of checking the quality of text in automatic mode, for example, two methods of analysing text in natural language: linguistic analysis, which is based on extracting the meaning of the text by its semantic structure, and statistical analysis, based on extracting the meaning of the text by the frequency distribution of words in the text [8]. Thus, the issues of creating the concept of automatic proofreading of technical text documents are relevant today in the scientific and educational environment and require their early resolution, as it will allow processing large arrays of text documents.

2. LITERATURE REVIEW

A review of the literature devoted to the establishment of the concept of automatic proofreading of technical documents indicates the variety of approaches of researchers to the options for solving the problem. In particular, the group of authors represented by Barakhnin et al. [9], in a joint study of various aspects of designing the structure of a software system for processing a corpus of text documents, note that one of the most difficult and pressing problems is the development of a universal set of tools for analysing texts in literary and business texts. As noted in by Bolshakova et al. [10], when recognising business text words, the most important factor is familiarity with the text (its topic, structure and most frequent words), keywords and topic elements are recognised relatively well, the end of the text is predictable and well recognised. For a literary text, a large "support" falls on the initial (preamble) and middle (plot development) compositional fragments and is differently correlated with the components of the communicative and semantic division: with the theme for the preamble, with the dialogue (especially the keywords or rhyme) for the middle fragment. Thus, when talking about text structures and analysis procedures, it is necessary to take into account various types of contexts, in particular, the functional style, compositional structure and rhetorical coherence of the text.

For its part, Volkovsky and Kovylin [11] draw attention to the fact that the current approach to finding information in multiple documents offers a solution to the ranking problem for a user query. This principle (the disadvantages of which are described above) is the basis of many popular web search engines. An alternative approach to solving the search problem is to obtain semantically significant text relative to the user query. At the same time, Alekseeva [12] studying the use of computer programmes for processing a specialised text corpus, pointed out the importance of functional standards for the operation of word processing programmes. According to the researchers, functional standards cover the area of developing functional requirements for the following processes: configuration management; supply of spare parts (initial and additional); maintenance, repair, and overhaul; modification and revision (information updates) of operational monitoring and fault reporting.

Notably, Kaibasova [13] highlighted certain aspects of the automatic determination of the parameter of uniqueness of texts by individual programmes in her study. The author notes that when solving linguistic problems of text processing, various approaches and methods of converting text information into sets of numerical data that will be used to extract knowledge are possible. When comparing texts and identifying matches in them, automatic linguistic analysis tools are needed. Thus, the researcher draws attention to the fact that the development of the concept of automatic text proofreading involves the use of linguistic analysis tools, the work of which is also subject to an automatic algorithm. Alternatively, a cluster method is proposed, which is a multidimensional statistical procedure that collects data containing information about a sample of objects, and then arranges the objects into relatively homogeneous groups [14]. However, the problem of creating software for automatic proofreading of text documents with an assessment of the prospects for its further implementation in the scientific field still remains unresolved. That is why the present study is an attempt to fill this gap.

3. MATERIALS AND METHODS

The subjects of the study are the programmes and principles of checking technical and other texts that have developed to date, adopted in the existing system of science and education, and the prospects for forming a universal model of automatic quality control of technical texts based on the developed software. The main research method is a system analysis of modern algorithms for checking technical documentation. With the help of this method, the prospects of building the concept of creating an optimal model of automated document

processing were determined. For a thorough study, scientific developments were selected within the framework of the design of the software structure of the technical document processing system in the context of finding optimal solutions for creating the concept of automatic proofreading of technical documents.

Some of the materials studied within the framework of the stated topic were borrowed from foreign publications devoted to the issues in hand. To facilitate the perception of information and create the highest quality, detailed and objective picture of the study, all materials cited in this study were translated into English. In general, the stated combination of materials and methods of performing this research meets the set tasks and contributes to the highest quality and objective disclosure of the issues raised in the subject of this study, since it allows forming a complete picture of the current situation with the quality software for checking technical texts in the modern scientific environment and the education system, as well as to assess the prospects for further search for opportunities to form a concept for improving the quality of automatic proofreading of technical documents.

A systematic study of the available scientific developments in the field of designing the structure of a software system for processing text documents allows forming a clear idea of the current situation in the field of assessing the text quality in documents and the possibilities for changing it for the better. At the same time, the analysis of the existing prospects for the creation of an optimal model for automatic document proofreading allows one to get an idea of the real possibilities of creating the considered conceptual model, taking into account all the features of the situation with development of software for text processing. The use of the chosen combination of materials and methods for the study of the issues submitted for consideration in the future can give qualitative results in the context of the further evolution of methods for checking technical documents, which is important in various aspects of scientific and educational spheres.

4. RESULTS

Today, the use of machine learning methods is widely used as a popular area for the development of algorithms for processing text corpora. This allows solving numerous problems of natural language processing, which is extremely important when working with the system for automatic proofreading of technical documents. Scientific developments in this area are caused by such factors as the specific structure of literary and business texts, as well as the lack of fully formed systems for analysing large volumes of text in Russian. The concept of building a complete model of automatic proofreading of technical documents involves the construction of a semantic network of the text as its main programme model.

Syntactic analysis, as a rule, is the first stage in the bulk of automatic text document processing systems. This stage involves the selection of individual sentences and words of the analysed text. The volume of words is reduced by stemming and clearing the text from official speech patterns. To this end, the Porter algorithm (defining and replacing unique stemmes) is applied, which cuts off the endings and then calculates the Levenshtein distance (embedding matrix) for the final results (Figure 1).

To date, the processing of text documents of various levels of complexity and genre orientation is one of the most dynamically developing branches of modern digital technologies. In recent decades, the main area of improving algorithms for processing text document corpora was the use of machine learning methods. The current approaches to automatic text analysis can be presented in the form of the following diagram (Figure 2).

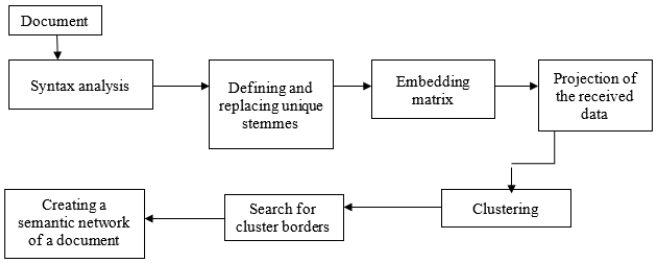

Figure 1. Building a semantic network of the text as its main programme model

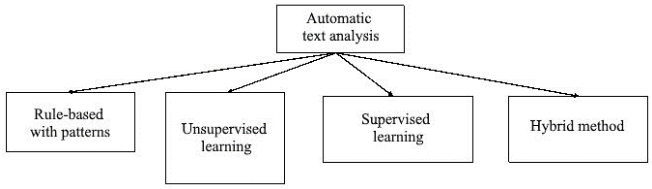

Figure 2. Basic approaches to automatic text analysis in the context of the templates and methods used

A rule-based approach prescribes a sequence in modelling the conceptual modelling constructs, and the action to be taken at each stage. Supervised learning is a machine learning approach that's defined by its use of labelled datasets. These datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Unsupervised learning uses machine learning algorithms to analyse and cluster unlabelled data sets. These algorithms discover hidden patterns in data without the need for human intervention. The hybrid method is based on a combination of various approaches (e.g., unsupervised learning & supervised learning).

Figure 2 schematically shows the main approaches to solving the issues of automatic text proofreading. In addition, there are two secondary methods that are not included in this diagram:

– the method of graph-theoretic models, which involves the breakdown of the body of the text into words, each of which has its own weight, which is important in determining the text tonality;

– the method of using models that include deep neural networks, an example of this is the recently popular BERT (Bidirectional Encoder Representations from Transformers) algorithm.

A variety of algorithms for checking text documents can be implemented by creating special software. At the same time, the structure of the created software network should be focused on end-users and other software systems. At the same time, this software must have the following functionality:

1. Provide free access to text corpora.
2. Provide the ability to automatically process text corpora from an existing database.
3. Send the received results to the database in a timely manner.

4. Implement flexible planning for the implementation of data processing tasks.

5. Provide high-quality processing of static characteristics and their representation in the specified form.

6. Improvement of algorithms for analysing text corpora.

In this context, it is essential to present the text analysis process as a series of sequential steps performed in order to determine the specific characteristics of the text in combination with the functionality of the automatic proofreading system of technical documents. The main elements of these two components are shown in Table 1.

Table 1. The sequence of steps of automatic text analysis and the parameters of the functionality of the automatic proofreading system

Stages of automatic text analysis	Functionality of the automatic proofreading system
Generation of text corpora	Ensuring reliable storage of large volumes of texts
Semantic analysis	Providing quick access to databases
Semantic analysis (definition of the genre)	Ensuring the establishment of a high-quality database proofreading structure
Final processing of the results	Providing the ability to store the final results

The creation of a conceptual model for automatic proofreading of technical documents involves the consistent development and implementation of the structure of a software system that allows qualitatively solving large-scale tasks of storing text corpora with a total volume of millions of units, while it should be possible to batch process texts online in the volume of up to several thousand documents. Such a system has been successfully implemented in the project of monitoring Russian-language mass media in Kazakhstan and is successfully functioning at the moment. In addition, the concept of automatic proofreading of technical documents involves the development of a system structure in which each element will be defined specific functions. The accuracy and reliability of the results of the automatic check are determined by the synchronicity of the interaction of all elements of the system in a given sequence. After all, without interaction between elements, the function of each individual element cannot be fully realized. In this context, the following types of structural elements of the system should be distinguished, expressed in the form of a certain semantic structure and are crucial in terms of ensuring the accuracy of the results of automatic proofreading of technical documents specifically and texts of various genres and styles in general (Figure 3).

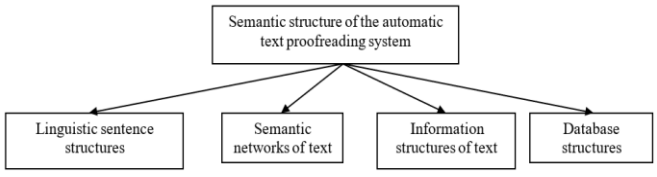

Figure 3. Semantic structure of the automatic text proofreading system

The linguistic structures of text sentences define a local understanding that is limited to the specific sentences of the text structure. The basis of such a semantic-syntactic representation is the syntactic tree of the sentence, which has "semantic" nodes or "semantic" connections. Such structures provide detailed analysis, reflected in the form of a tree – the syntactic and semantic expression of a sequence of words and phrases in a sentence. If there are dictionary entries for all the words of the sentence, provided that the structure is correct – in terms of the rules of the input language – the software builds a consistent syntactic structure, first surface, then underlying. Semantic networks of text are a global structure, blurry for understanding. In this network, a different deep semantic component is implemented, the "essence of the text" is the result of the translation of the semantic and syntactic structures (representations) of all sentences of the text into the language of more "elementary" units. Such networks reflect the semantic structure of the text being checked with the language forms and morphological components used in it.

The information structures of the text fix its generalised understanding in accordance with the selected terminology classifiers, subject heading lists, thesauri, etc. They are widely used in information search engines. They contain a large amount of data, based on which the texts are analysed for compliance with certain pre-defined criteria. The high demand and large-scale development of various information search engines is an undoubted advantage for evaluating the effectiveness of this automatic understanding approach. Database structures represent a selective, mediated special understanding that largely takes into account the extralinguistic expression, the representation of the real part of the present. Such structures are widely used in artificial intelligence systems, because they display the whole text and do not respond to the division into separate sentences. These structures usually contain the actual results in digital, graphical, and verbal terms.

The most problematic issue that is effectively eliminated when modelling text research by an automatic system is the coherence of the text and the integrity of its structure. It is the coherence of the text and its thematic focus that determines the final effectiveness of the proofreading system and the accuracy of the final results. In order to resolve this issue qualitatively, it is necessary to address the problems of referentiality. Among the means of ensuring coherence in the texts of technical documents, the following should be noted:

– repeated concepts used in the same lexical expression;

– repeated concepts used in various lexical expressions;

– pronouns and pronominal terms related to the means of expressing the meanings of repeated concepts;

– terms denoting stable, generalised logical and compositional connections and constructions located between different levels of text components;

– connective words and phrases.

Thus, the conceptual model of automatic proofreading of technical documents involves the development and implementation of a complex system of interrelated elements, each of which performs its own functions and, in interaction with other elements of the system, affects the results. This concept implies the need to develop and implement special software that takes into account all the features of the verification system and the characteristics of the specific technical documents that is subject to verification.

5. DISCUSSION

The discussion of the general issues of developing a conceptual model for automatic proofreading of technical

documents emphasises the importance of the issues under consideration and the diversity of opinions of researchers on various aspects. Thus, the team of authors represented by Bolshakova et al. [10], in the study of automatic word processing and computational linguistics, note the importance of the language factor as a determining element of any text processing. According to the researchers, language is a sign system, meaning that the main element of such a system is a sign. The sign serves as a means of reflecting a particular element of reality. Due to the presence of this sign in the language, this element not only gets representation in the system of knowledge about the world inherent in the native speaker, but also the opportunity to transfer this knowledge to another. Knowledge becomes communicable. In addition, the researchers note that the information is contained in the text (not in the language), but the text is constructed and analysed using the language. Hence, it is easy to assume that texts of significantly different types impose their own requirements on the language used. First of all, this refers to texts that differ in the degree and type of information load: texts of different functional styles [15].

For its part, Parkhomenko et al. [16] exploring the issues of comparing text clustering methods, suggest that this method is effective in various types of text search: information, research, and also in detecting spam. At the same time, cluster-based navigation is an effective solution to the problem of finding the necessary results of automatic proofreading of technical documents. The main advantages of text clustering over other methods are that it groups a set of texts in such a way that texts in one group are more similar to each other than to texts in other clusters. In turn, researchers [17] in a joint study of a wide range of issues of thematic modelling of texts in natural language, note that, as a rule, the number of topics found in documents is less than the number of different words in the entire set. Therefore, hidden variables – topics – allow representing a document as a vector in the space of hidden (latent) topics instead of representation in the word space. As a result, the document has fewer components, which allows processing it faster and more efficiently. Thus, thematic modelling is closely related to another class of problems known as data dimensionality reduction. In addition, the found topics can be used for semantic analysis of texts.

According to the researchers, thematic modelling is a way to build a model of the location of text documents, in which the subject of the material and its complexity are automatically determined, along with the features of building the semantic core. In addition, the transition from the space of terms to the space of found topics helps to resolve synonymy and polysemy of terms, and more effectively solve problems such as thematic search, classification, summarisation and annotation of collections of documents and news streams [18]. In a joint study [19], assess modern approaches to creating a system for automatic processing of technical documents. In their opinion, the existing approaches in the field of automatic processing of text corpora are still far from ideal: translation of proper names, incorrect sentence structure, lack of grammatical connections, etc. The undeniable advantage of automatic proofreading is the speed and comparative, relatively manual procedure, cheapness of text processing. However, the risk of gross thematic errors increases if a highly specialised type of proofreading is required, when both highly qualified check and excellent knowledge in a particular field are necessary. At the same time, the researchers note that the system of automatic proofreading of text allows optimising this process and minimise the cost of processing text information [20]. This indicates the presence of different aspects of the assessment of texts in the works of even the same researcher.

In turn, the research team represented by Turdakov et al. [21], cite the comparative advantages of text processing with the help of Texterra project, which seems important in the context of this study. The researchers note that unlike many existing text processing and analysis projects, the main priorities in the Texterra project were the use of automatic methods and high data processing speed while maintaining the highest possible quality of text analysis. As a result of the project, the technology was created, which is successfully implemented in several commercial projects with Russian partners and partners from other countries, as well as in the own services of the V. P. Ivannikov Institute of System Programming of the Russian Academy of Sciences (ISP RAS). Thus, the researchers emphasise the relevance of the automatic text proofreading system and its practical effectiveness. In turn, the researchers [9], in a joint study of the design of a software system for processing text document corpora, note the need to introduce several types of storages into such a system, which provide the ability to quickly access databases, system components, and a subsystem for generating analytical reports. This fact highlights the complexity of building a system for automatic document proofreading and the need for a clear sequence of operations when working with it.

The subject under consideration is reflected by studies of the issues of automatic processing of information arrays. In particular, Haramundanis [22] notes that the use of automatic systems for checking technical documents reduces the processing time of documents and has a positive effect on their final quality. In turn, Grady [23] draws attention to the fact that automatic systems for checking text documents would soon replace manual proofreading systems due to the variety of their advantages. Thus, the discussion of the issues presented for consideration in the works of modern researchers emphasises the relevance of the subject matter and the variety of scientific approaches to its coverage.

The concept of automatic proofreading of texts of various complexity and genre orientation in general, and technical documents, in particular, is considered as a fundamental way to work with large text arrays in limited time intervals. Among the advantages of automatic proofreading of technical documents, there are significant cost and time savings compared to manual proofreading, as well as the possibility of involving a limited number of specialists in the maintenance of these systems. In addition, this conceptual model involves the use of a number of key subsystems designed to ensure the achievement of the necessary quality indicators of document proofreading, such as: linguistic structures of sentences of the text being checked, its semantic networks, information structures, and database structures. Automatic proofreading of technical documents involves the construction of a semantic network of the text as its main programme model and the cyclical quality checks of texts in a given unit of time.

6. CONCLUSIONS

The automatic system for checking technical documents requires the development of special software that takes into account all the technical features of the system and guarantees high-quality work for a long period of time without loss of quality and saves the results of checks in special databases. In

addition, in the work of an automatic proofreading systems based on special software, it is important to take into account the lexical features of documents, tables, graphs, drawings, diagrams, etc., presented in it and the specifics of the presentation of digital information. In general, the automatic proofreading system under consideration should take into account all of the above factors. Thus, the conceptual model of automatic proofreading of technical documents involves the creation of a complex system for checking the quality of text materials using a variety of components that perform various functions, and in a complex that provides specific results based on the checks performed.

Further study of this issue would help to add factual material to expand the existing understanding of the capabilities of automatic proofreading systems and create prerequisites for the introduction of real programmes for automatic proofreading of technical documents in practical use.

REFERENCES

[1] Seligmar, E., Schubert, T., Achutha, M.V., Kumar, K. (2015). Formal verification. An essential toolkit for modern VLSI design. Morgan Kaufmann, Burlington.

[2] Silva, E.A., Valentin, E., Carvalho, J.R.H., Baretto, R.S. (2021). A survey of model driven engineering in robotics. Journal of Computer Languages, 62: 101021. https://doi.org/10.1016/j.cola.2020.101021

[3] Lee, H.S., McNamara, D., Bracey, Z.B., Wilson, C., Osborne, J., Haudek, K.C., Liu, O.L., Pallant, A., Gerard, L., Linn, M.C., Sherin, B. (2019). Computerized text analysis: Assessment and research potentials for promoting learning. Computer-Supported Collaborative Learning Conference, 2: 743-750. https://doi.dx.org/10.22318/cscl2019.743

[4] Régis-Gianas, Y., Jeannerod, N., Treinen, R. (2018). Morbig: A static parser for POSIX shell. SLE 2018: Proceedings of the 11th ACM SIGPLAN International Conference on Software Language Engineering, pp. 29-41. https://doi.org/10.1145/3276604.3276615

[5] Thomas, S.W., Adams, B., Hassan, A.E., Blostein, D. (2014). Studying software evolution using topic models. Science of Computer Programming, 80: 457-479. https://doi.org/10.1016/j.scico.2012.08.003

[6] Aizstrauts, A., Ginters, E., Baltruks, M., Gusev, M. (2015). Architecture for distributed simulation environment. Procedia Computer Science, 43(C): 18-25. https://doi.org/10.1016/j.procs.2014.12.004

[7] Ginters, E. (2019). Augmented reality use for cycling quality improvement. Procedia Computer Science, 149: 167-176. https://doi.org/10.1016/j.procs.2019.01.120

[8] Zaytsev, V. (2015). Grammar zoo: A corpus of experimental grammar ware. Science of Computer Programming, 98: 28-51. https://doi.org/10.1016/j.scico.2014.07.010

[9] Barakhnin, V.B., Kozhemyakina, O., Mukhamediev, R., Borzilova, Y., Yakunin, K. (2019). Designing the structure of a software system for processing text document corpora. Business-Informatics, 13(4): 60-72. https://doi.org/10.17323/1998-0663.2019.4.60.72

[10] Bolshakova, E.I., Klyshinsky, E.S., Lande, D.V., Noskov, A.A., Peskova, O.V., Yagunova, E.V. (2011). Automatic processing of texts in natural language and computational linguistics. MIEM, Moscow.

[11] Volkovsky, O.S., Kovylin, E.R. (2019). Computer system of intellectual semantic search with the text generation using. Bulletin of the Kherson National Technical University, 1(3): 238-244.

[12] Alekseeva, N.I. (2012). The use of concordance programs in teaching linguists on the example of processing a specialized corpus of texts. Teacher of the XXI Century, 4: 196-200.

[13] Kaibasova, D.Z. (2020). Extraction of statistical data to determine the uniqueness of documents based on the analysis of the content of the curriculum of disciplines. The Scientific Heritage, 44: 57-63.

[14] Zhang, Y., Zhang, Y., Zhang, R. (2020). Text information classification method based on secondly fuzzy clustering algorithm. Journal of Intelligent and Fuzzy Systems, 38(6): 7743-7754. https://doi.org/10.3233/JIFS-179844

[15] Taraban, R., Pittman, J., Nalabandian, T., Yang, W.F.Z., Marcy, W.M., Gunturu, S.M. (2019). Creating and testing specialized dictionaries for text analysis. East European Journal of Psycholinguistics, 6(1): 65-75. https://doi.org/10.29038/eejpl.2019.6.1.rta

[16] Parkhomenko, P.A., Grigoriev, A.A., Astrakhantsev, N.A. (2017). Review and experimental comparison of text clustering methods. Proceedings of the Institute for System Programming of the Russian Academy of Sciences, 29: 161-200.

[17] Korshunov, A.D., Gomzin, A.B. (2012). Thematic modelling of texts in natural language. Proceedings of the Institute for System Programming of the Russian Academy of Sciences, 7: 215-243.

[18] Zaman, F., Shardlow, M., Hassan, S.U., Aljohani, N.R., Nawaz, R. (2020). HTSS: A novel hybrid text summarisation and simplification architecture. Information Processing & Management, 57(6): 102351. https://doi.org/10.1016/j.ipm.2020.102351

[19] Litvinov, V.V., Moiseenko, O.P. (2014). Automated system for processing dynamic collections of multilingual text documents on sea and river affairs. Mathematical Machines and Systems, 2: 59-64.

[20] Perez-Rodriguez, G., Perez-Perez, M., Fdez-Riverola, F., Lourenco, A. (2019). Online visibility of software-related web sites: The case of biomedical text mining tools. Information Processing & Management, 56(3): 565-583. https://doi.org/10.1016/j.ipm.2018.11.011

[21] Turdakov, D.S., Astrakhantsev, N.V., Nedumov, Ya.I., Sysoev, A.V., Andrianov, I.B., Mayorov, V.A., Fedorenko, D.M., Korshunov, A.I., Kuznetsov, S.V. (2014). Texterra: Infrastructure for text analysis. Proceedings of the Institute for System Programming of the Russian Academy of Sciences, 8: 421-437.

[22] Haramundanis, K. (2014). The art of technical documentation. Digital Press, Clifton. https://doi.org/10.1016/C2013-0-06781-3

[23] Grady, J. (2016). System Verification. Proving the design solution satisfies the requirements. Academic Press, Oxford.