Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 190 (2021) 678–683

Procedia Computer Science

www.elsevier.com/locate/procedia

2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: Eleventh Annual Meeting of the BICA Society

Application of text mining technologies in Russian language for solving the problems of primary financial monitoring

V.Yu. Radygin[a], D.Yu. Kupriyanov[a], R.A. Bessonov[a], M.N. Ivanov[b], I.V. Osliakova[c]

[a]*National Research Nuclear University "MEPHI", 31 Kashirskoe Shosse, Moscow, 115409, Russian Federation*
[b]*Financial University under the Government of the Russian Federation, 49 Leningradsky Prospekt, Moscow, 125993, Russian Federation*
[c]*MIREA - Russian Technological University, 78 Vernadsky Avenue, Moscow, 119454, Russian Federation*

Abstract

The paper deals with the issues of text mining based on a vector classification model. The term frequency and inverse document frequency function (TF IDF) is used as a measure to evaluate the selection of terms. The text preprocessing stage performs tokenization and normalization of the text. The algorithms of stemming and lemmatization of the Russian-language text are used during normalization. A set of programs has been created that makes it possible to analyze the selected subject area, create a thesaurus for thematic search for violations of the public procurement Federal Law of the Russian Federation, as well as to implement functionality for automated search and analysis of documents of arbitration courts within the framework of this law.

Keywords: text mining; natural language processing; vector space model; text tokenization; text normalization; stemming; lemmatization; thesaurus; search of documents

1. Introduction

The rapid accumulation of digital information in the world makes it not only possible, but also necessary to use data mining technologies in many areas of human activity. The use of such technologies in the field of economic security, combating money laundering and terrorist financing is becoming especially relevant. The huge data arrays arising in this area can be represented both by ordinary data typical for classical DBMS (for example, bank transactions, without taking into account the purpose of payment), and by large corpus of text data. The last category includes the financial statements of organizations, banks of arbitration, civil and criminal proceedings, public procurement systems, and even text information on the purpose of bank payments. Tasks of the first type require

10.1016/j.procs.2021.06.078

classical Data Mining technologies to solve, while the latter require more complex approaches related to the Text Mining category.

The importance of using Text Mining technologies in the tasks of economics and economic security is confirmed by the practical research papers of many authors. For example, in the research of N.N. Sakhare [1] shows the application of Text Mining technologies for the problems of stock market analysis. The importance of using Text Mining technologies for national security issues is shown in the work of S. Savas [2]. It is also indirectly confirmed by the research of A. Zanasi [3], which proposes a method for detecting crimes based on the analysis of texts of social networks. A large number of research papers can also be found directly in the area of combating money laundering. For example, Y. Qifeng [4] in his work shows the solutions for combating money laundering based on Text Mining technologies used in Union Bank. F.R. Federici [5] in his work considers a similar issue more broadly at the level of the American and European initiatives.

At the same time, unfortunately, today not all tasks arising in the work of government bodies and ensuring issues of economic security, intelligence and combating money laundering and terrorist financing are automated to the proper extent. For example, the task of analyzing the history of arbitration, civil and criminal cases remains largely "manual" labor of service employees. Automation of these processes through the use of Text Mining technologies will greatly improve the efficiency of the activities of such government bodies.

2. Task definition

In this paper, using the example of one small task required to support the activities of the Federal Service for Financial Monitoring of the Russian Federation, the possibility of effective use of Data Mining and Text Mining technologies in organizing the work of economic security services is shown. In particular, the paper considers the problem of automating the search for materials of arbitration office work in accordance with a highly specialized area of legislation – control over the procurement of goods, works and services to meet state and municipal needs.

If violations in this area are detected, the service employees need to analyze the arbitration paperwork as part of tracking cash flows, analyzing payment chains from the budget to the subcontractor, confirming or refuting the facts of opening special accounts in authorized banks according to contracts, studying the details and evaluating the facts when resolving various economic disputes. Despite the fact that at present in the Russian Federation there is already a card index of arbitration cases in the form of an information system that provides data on office work, it is often difficult to find relevant information on issues of interest in it. Therefore, this paper proposes a software solution for assessing the relevance of a text corpus of legal proceedings on a given search topic.

In accordance with the task, software was created for searching and analyzing documents of arbitration courts within the framework of the Federal Law No. 44 of April 5, 2013 "On the contract system in the field of procurement of goods, works, services to meet state and municipal needs", which was subsequently proposed for approbation by the employees of Rosfinmonitoring as part of its efforts to detect violations of the Federal Law.

The software solution is built on a micromodular architecture and includes the following components: module for preparing a structured dataset based on electronic documents written in natural language; a module for constructing a thesaurus based on Federal Law No. 44 of April 5, 2013; module for searching, importing and storing materials of arbitration proceedings; interaction module, which includes elements of graphical interfaces for filling out search forms and displaying results.

3. Methods and approaches used to implement a software solution

The software product is based on the module for preparing a structured data set based on electronic documents written in natural language. The module uses the vector model of the semantic texts classification (Vector Space Model, VSM). This model allows you to represent a documents corpus in the form of vectors belonging to the same vector space. Each document corresponds with a vector that expresses its meaning. Such a view allows you to compare words, search for similar words among them, carry out classification, clustering and much more. Thus, the vector model is the basis for solving such problems of information retrieval as document search on demand, classification and clustering of documents.

In general, text consists of not only words but also other elements such as punctuation, numbers, and special symbols. All of the above text elements are called terms. The weights of all terms found in the entire corpus of documents within the framework of this document are the vector that represents this document in vector space:

$$d_i = (w_{i1}, w_{i2}, w_{i3}, \dots, w_{in}), \quad (1)$$

where d_i is the vector identifying the *i-th* document in the vector space, w_{ij} is the weight of the *j-th* term in the *i-th* document, and *n* is the total count of documents in the presented corpus. It should be noted that the vector describing the document will contain information about terms that may not occur in this document, but are present in the entire set of documents. Thus, the resulting set of vectors representing all documents of the set under study has the same size.

One of the questions of the vector models description of is the question of choosing the evaluation of terms (weights) in a text document. Among the main generally used estimates are: boolean weight, which is non-zero if the term occurs in the document and zero otherwise; weight based on the term frequency (TF): with such an estimate, the weight is determined as a function of the count of occurrences of the term in the document; weight based on the term frequency and the inverse document frequency (TF-IDF): with such an assessment, the weight is determined as the product of a function of the term counts in a document and a function of the inverse to the count of documents where this term occurs.

V. Hashemzadeh [6] shows the possibility of using an algorithm for processing text documents based on a vector model of semantic classification with an assessment by the TF-IDF criterion, and also presents experimental results of processing texts in eight different languages, including European, Turkic, Arabic and Persian. This work allows us to conclude that the TF-IDF measure is the most versatile, which determined its choice for the implementation of the software product considered in this article.

According to the measure of the TF-IDF vector, the weight of w_{ij} is calculated by the formulas:

$$w_{ij} = tf_{ij} \times idf_j, \quad tf_{ij} = \frac{\theta_{ij}}{\sum_{i=1}^{N} \theta_{ij}}, \quad idf_j = log\,\frac{N}{C_j}, \quad (2)$$

where *N* is the total count of documents, θ_{ij} is the count of the term occurrences with index *j* in the document with index *i*, and C_j is the count of documents containing text with this term.

The text presented in the usual human form, may comprise a terms set that are not informative (noise) for data analysis. Among them, one can distinguish, for example, punctuation marks or various variations of word forms. This data does not allow extracting any features for constructing meaningful models. To highlight meaningful terms in the text, they usually resort to the stages of text preprocessing. As a rule, the stages of text data tokenization and normalization are distinguished among them. The tokenization stage is based on the process of text dividing into separate words or other lexemes (word formation or combinations). These units, as a rule, are called tokens. As a result of the tokenization stage, the source text is converted into a list of words (tokens). The text tokenization stage includes: converting the words of the raw text to lower case; replacing punctuation marks with spaces or blank lines; declaration of words or lexemes as separate tokens. At the stage of text preprocessing, it is useful to use additional actions to remove the so-called stop-words. Stop-words are words that are found in a large volume of texts and do not carry a special semantic load. Usually they include interjections, prepositions, particles, some pronouns, adjectives and other parts of speech, for example: interjections: oh, uh, wow, well, really; pronouns: I, we, mine, yours, you; introductory constructions: let's say, suppose, in general, for example; generalizations and imprecise definitions: everything, about, somewhere and others.

When using machine learning to analyze text data, these words tend to add a lot of noise. Therefore, it is most often recommended to clean the preprocessed text of such irrelevant words. In the work of T. Jiang [7], the process of identifying significant terms based on the modernized TF-IDF assessment is shown. Based on the experience of this author, it can be noted that, in general, the tokenization stage does not represent any complex (in terms of logic and algorithms) or laborious (in terms of implementation) operations. However, taking into account the specifics of the Russian language, at this stage, some nuances can lead to the loss of meaning and affect the accuracy of further machine learning. For example: reducing acronyms to lower case can be confused with expressing emotions;

replacing all punctuation marks with spaces may lose the original phrases; some proper names can occur in pairs, and during tokenization they will be split into two separate words.

At the second stage (normalization of the text) the words are reduced to the canonical form of the word. For nouns, the canonical form of the word is the singular and nominative, for adjectives, the masculine, singular and nominative.

For vector analysis, the transformation of normalization does not lead to losses, since the contextual word form rarely has additional useful information and the word meaning is determined precisely by the canonical, initial form. At the same time, reducing words to their initial form allows you to reduce the total array of words in the document, and thus reduce the number of features. As shown in the research of R.-C. Chen [8], highlighting the most important features and reducing their total number can improve the classification accuracy and reduce the processing time of a large amount of data.

The stemming and lemmatization methods are used to normalize the text. Stemming is a crude heuristic process of finding the stem (root) of a word. This method is based on truncating the non-root portions of the word. There are many algorithms for the implementation of stemming - stemmers. In the work of N.A. Razmi [9] shows how the stemming process can be used to isolate the roots of significant terms found in a given set of articles, and also provides a way to visualize the most significant terms, which allows us to talk about the effectiveness of this approach.

In turn, lemmatization is a more subtle process, which is carried out on the basis of morphological analysis using a specialized dictionary in order to eventually bring a word to its canonical form – a lemma. The process includes determining the part of speech of a word and applying different normalization rules for each part of speech according to the rules for the formation of word forms. E.M. Buchanan [10] studies a practical application of the lemmatization process. The basis of this process is the use of special dictionaries. Among the grammatical dictionaries of the Russian language most often used for morphological analysis, one can single out the dictionaries of A.A. Zaliznyak, which contains more than 170 thousand words and more than 5 million word forms and the dictionary of A.N. Tikhonov, which has a modern design and graphological structure of the formation of word forms. It should be noted that the process of lemmatization is much less productive than the process of stemming, but at the same time it is more accurate, provided that there is a word form in the dictionary.

The process of feature extraction based on the grammatical definition of the canonical forms of terms is somewhat inflexible. It is often very difficult to understand in advance which grammatical template will most effectively identify meaningful terms and phrases in the text. Therefore, in natural language processing, so-called N-grams are mainly used. An N-gram is a sequence of words that form collocations – stable phrases that have signs of a syntactically and semantically integral unit, in which the choice of one of the components is carried out according to meaning, and the choice of the second depends on the choice of the first. A sequence of two words is often called a bigram, a sequence of three words is called a trigram, and a sequence of four or more words is called an N-gram, where N is replaced by the count of words forming a collocation. The success of this approach for processing text in natural language is shown in the work of A. Borjali [11]. Analysis of the results of this author allows us to conclude that the use of N-grams in the text documents analysis allows you to expand the attribute space by adding phrases to the tokens set.

4. Practical implementation of the software product

When choosing software for the implementation of the complex modules, an analysis was made of the capabilities of modern programming languages and their auxiliary libraries. Among the programming languages, the scripting languages Ruby and Python were considered, as well as the object-oriented programming language Java. The listed programming languages are multiplatform and are well suited both for creating web applications with a client-server architecture and for writing desktop programs. Despite the authors' wide experience in building software products based on the Ruby language [12], this language was recognized as not satisfying the set requirements, since it does not have high-quality libraries for processing texts written in national languages, including Russian. In contrast, the Python and Java languages have a high-quality implementation of the NLTK (Natural Language Tool-Kit) library, which allows us to talk about their applicability for solving the assigned tasks. At the same time, among these two languages, the choice was made in favor of the Python language, since it has an

advantage in the speed of software development compared to the Java language. The authors conclusions are confirmed by the works of many domestic and foreign developers. For example, K. Arun [13] shows that the Python language in conjunction with its libraries is suitable both for text preprocessing and for the implementation of machine learning algorithms for multilingual text processing. It should be noted that it is also necessary to pay attention to modern protection methods while implementing software. So, Sh.G. Magomedov [14] proposes to expand the currently developing SIEM technology (Security information and event management) with a user behavior analysis unit and shows that such a solution can be implemented for most digital platforms such as banking, medical, educational and other platforms.

A complete set of software development tools includes: Python programming language; Scikit – Learn library, an extension for the SciPy library (Scientific Python); library NLTK (Natural Language Tool-Kit), a package of tools for natural language processing, the Gensim library, which allows you to implement semantic text modeling without a teacher (word2vec); the SpaCy library, which allows you to process texts in natural language, in particular, to perform text preprocessing in preparation for training; library Pymorphy2 – a morphological analyzer that allows you to perform lemmatization and obtain derived word forms from a known canonical form according to the given grammatical characteristics of word formation based on the OpenCorpora dictionary (for forms that are not in the dictionary, it allows you to build word formation hypotheses).

The analysis of the subject area for the construction of the Thesaurus module was based on the texts of the Federal Law of April 5, 2013 No. 44, as well as articles and materials of arbitration proceedings related to breaking it. When analyzing these documents, a frequency analysis of the words (terms) found in the documents was carried out. The results of the analysis made it possible to identify not only the main corpus of words (Fig. 1a), but also potential stop-words). In addition, analysis and feature extraction were performed using N-grams. Based on the vector analysis of the text with the TF-IDF measure, an alternative array of keywords was obtained (Fig. 1b). On the basis of the integral assessment of the results of vector analysis with the TF-IDF measure and the analysis of the results of constructing N-grams, a thesaurus was compiled for thematic search for violations of the Federal Law No. 44 dated April 5, 2013 ensuring state and municipal needs.

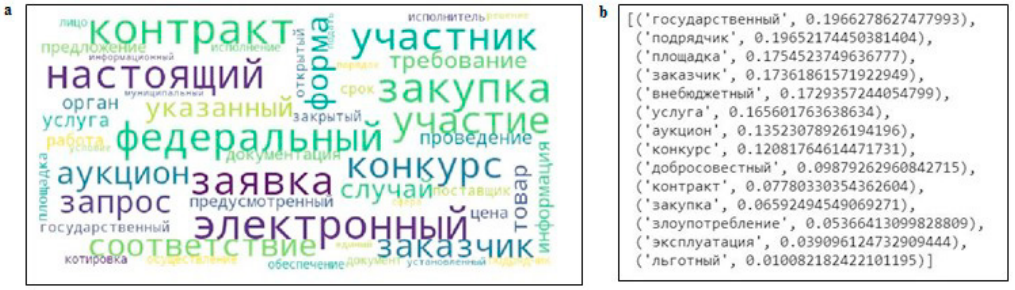

Fig. 1. (a) Visualization of the most frequently used terms extracted from the texts of Federal Law No. 44 and related materials of arbitration proceedings; (b) Rating of terms obtained by vector data analysis and TF-IDF measure.

The download of documents for training the model was carried out from the site "Electronic Justice", which contains a card index of arbitration cases and a bank of decisions of arbitration courts (https://ras.arbitr.ru). To automate this process, a search robot was created based on the Selenium WebDriver library.

It should be noted that the results obtained correlate well with the results obtained by the expert assessment method. The expert assessment was carried out by experts in economic security of the Department of Financial Monitoring of NRNU (MEPHI). At the same time, the group criterion for the formation of expertise was used when interviewing nominal groups. The final assessment of the effectiveness of the created complex was carried out by comparison with an expert search for relevant documents of arbitration office work. The sample included 400 documents, about 30% of which were recognized by the expert as relevant. The Precision's efficiency score was about 0.91, the Recall score was about 0.85 and the final F1 score was 0.87, which allows us to speak about the high efficiency of the created product. Moreover, further training of the model can improve the quality of document search in the future.

5. Conclusion

The developed software package allows for automated search and download of documents of arbitration proceedings within a given topic from the site "Electronic justice". The software package approbation showed that the use of the developed system within the framework of the activities of Rosfinmonitoring employees makes it possible to find documents, and thus identify the circumstances of transactions concluded, confirm or deny signs of dubious transactions, identify theft of budget funds and / or financial losses, thereby ensuring the economic security of the Russian Federation.

The most logical development of the program complex may be the expansion of the search functionality. As a rule, when automating the office work process, archives are created in which electronic copies of paper documents (scans) are stored. At the same time, if the information system allows you to create links between accounting objects and electronic copies of documents, then the organization of the processing textual data, as well as the classification of graphic objects contained in these copies, becomes important when searching in such information systems. The issues of archiving and searching documents in archives, taking into account the features of compression and noise reduction, which is used to improve the image quality in scanned documents, are discussed in the work of S.K. Kartsov [15].

The Text Mining technologies proposed in the work can be used to solve a wide range of problems of economic security, countering terrorism and money laundering.

References

[1] Sakhare N.N., Imambi S.S., Kagad S., Kapadwanjwala T., Malekar H., Dalal M. Stock Market Prediction Using Sentiment Analysis // International Journal of Advanced Science and Technology Vol. 29, No. 4s, (2020), pp.1126-1133.

[2] Savas S., Topaloglu N., Data analysis through social media according to the classified crime // Turkish Journal of Electrical Engineering & Computer Sciences Vol. 27 (2019), pp. 407 – 420.

[3] Zanasi A., Artioli M. Text and video mining solutions to national security intelligence problems // WIT Transactions on Information and Communication Technologies, Vol 42, 2009, WIT Press, pp. 3-12.

[4] Qifeng Y., Bin F., Ping S. Study on Anti-Money Laundering Service System of Online Payment based on Union-Bank mode // 2007 International Conference on Wireless Communications, Networking and Mobile Computing, Shanghai, 2007, pp. 4991-4994.

[5] Federici F.R. Money laundering, terrorist financing and how to contrast them: data and text mining in business intelligence solutions // WIT Transactions on Information and Communication Technologies, Vol 38, WIT Press, pp. 315-324.

[6] Hashemzadeh, B., Abdolrazzagh-Nezhad, M. Improving keyword extraction in multilingual texts // International Journal of Electrical and Computer Engineering, 2020, 10(6), pp. 5909-5916

[7] Jiang, T., Jia, L., Wan, M.C., Meng, J.H. The Text modeling method of Tibetan text combining Word2vec and improved TF-IDF // Journal of Physics: Conference Series, 2020, 1601(4), 042007

[8] Chen, R.-C., Dewi, C., Huang, S.-W., Caraka, R.E. Selecting critical features for data classification based on machine learning methods // Journal of Big Data, 2020, 7(1), 52

[9] Razmi, N.A., Zamri, M.Z., Ghazalli, S.S.S., Seman, N. Visualizing stemming techniques on online news articles text analytics // Bulletin of Electrical Engineering and Informatics, 2021, 10(1), pp. 365-365

[10] Buchanan, E.M., De Deyne, S., Montefinese, M. A practical primer on processing semantic property norm data // Cognitive Processing, 2020, 21(4), pp. 587-599

[11] Borjali, A., Magnéli, M., Shin, D., (...), Muratoglu, O.K., Varadarajan, K.M. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation // Computers in Biology and Medicine, 2021, 129, 104140

[12] Radygin, V.Y., Lukyanova, N.V., Kupriyanov, D.Yu. LMS in university for in-class education: Synergy of free software, competitive approach and social networks technology // Proceedings of the International Scientific-Practical Conference Information Technologies in Education of the XXI Century, (ITE-XXI). Melville, NY: AIP Publishing, 2017, 020015.

[13] Arun, K., Srinagesh, A. Multi-lingual Twitter sentiment analysis using machine learning // International Journal of Electrical and Computer Engineering, 2020, 10(6), pp. 5992-6000.

[14] Magomedov S.G., Kolyasnikov P.V., Nikulchev E.V. Development of technology for controlling access to digital portals and platforms based on estimates of user reaction time built into the interface // Russian Technological Journal. 2020, 8(6) pp. 34-46.

[15] Kartsov, S.K., Kupriyanov, D.Y., Polyakov, Y.A., Zykov, A.N. Non-local Means Denoising Algorithm Based on Local Binary Patterns // Intelligent Systems Reference Library, 2020 Vol. 182, Q2 pp. 153-164.