UDC 004.89=111

D. O. Yakymenko, *Graduate Student (Applicant),*
Ye. Yu. Kataieva, *Ph. D., Associate Professor*
Cherkasy State Technological University
Shevchenko blvd, 460, Cherkasy, 18006, Ukraine

METHODS AND MEANS OF INTELLIGENT ANALYSIS OF TEXT DOCUMENTS

The paper reviews the methods of analysis and processing of electronic documents. Methods of analysis of text documents to solve the problem of determining the thematic affinity of texts are analyzed. An overview of existing approaches to solving the classification problem is performed. The main approaches used in the task of text classification are described; the stages of the classification process are determined and the most common methods of classifying text documents are considered. The main approaches to text pre-processing, such as: lower case, root correction, stemming, lemmatization, stop word removal, normalization, are considered. Advantages and disadvantages of each approach are considered. The procedure for reducing the dimension of a set of features with a division into sub-processes: selecting features and highlighting features is considered.

Keywords: *keywords, text analysis, search, text documents, classifications.*

Introduction. The rapid development of computer technology has led to the rapid accumulation of electronic text documents. This situation will lead to the fact that existing methods of processing electronic text documents will not be able to meet the needs of users in both corporate networks and the Internet.

That is why there is a need for methods that will provide a quick and easy distribution of documents by category or keyword.

Based on this, you can identify the following major problems associated with increasing the amount of information:

- the growing volume of documents posted on the Internet is the reason for the growing difficulties in finding the necessary documents for users and their organization;
- technologies for working with text documents involve a rather large complexity of implementation, which affects the speed of data processing;
- a significant part of electronic text documents is unstructured information.

The purpose of the study is to analyze existing means of electronic texts analysis.

The research task consists in the analysis of:

1) means to work with text documents;

2) methods of processing text documents;

3) applications used to analyze electronic texts.

Presentation of the main material. The most conventional examples of intelligent text analysis include technologies for extracting factual information about the search object, fuzzy search; thematic and tonal (accurate and complete) rubrics; selection of documents; selection of key topics; construction of annotations; use of methods of intelligent text analysis to determine the directions of research of large document funds and obtain new information about the process of automatic text summarization. The most modern areas of obtaining information from texts today are:

- analytical processing of facts;
- file keeping;
- annotation of documents;
- conducting a thematic analysis of docskills (clustering and heading);
- construction and dynamic analysis of a sample of the ethical structure of texts;
- selection of key topics and information of other objects;
- study of frequency characteristics of texts [1].

Most of these methods are aimed at processing documents, each of which relates to a specific, often quite narrow topic. This statement is true mainly for small documents (such as small web pages), but not for large documents, which often relate to several topics at once. Examples of such documents are musts, scientific articles and others. We introduce a definition for this type of document.

Definition 1. A text document of any size, which simultaneously belongs to several topics, is called polythematic.

DOI: 10.24025/2306-4412.2.2022.259408

Definition 2. A text document of any size, which relates to only one specific topic, is called monothematic.

Definition 3: The subject of the text document - some subjective perception of the person, the user of the search engine, which is considered in the text of the subject area, its main content.

Definition 4: A text is a sequence of sentences, words, built according to the rules of a given language, a given sign system and forms a message that carries some useful information [2].

Polythematic documents contain multifaceted goals, thus reflecting the information needs of different users. Text repositories, which usually store such documents, are usually characterized by several topics.

To date, few methods allow to process polythematic text documents - to perform their classification or clustering. Therefore, the development of methods for solving these problems is especially important. The purpose of the classification of polythematic text documents is to assign the same text document to more than one topic. The purpose of clustering polythematic text documents is to automatically divide text documents into a priori not specified groups so that each document belongs to more than one of them [3, 4].

Classification of text documents

Document classification, often known as document categorization, is a challenge in library, information, and computer science. The assignment of a document to one or more groups or categories is the task. This can be accomplished "manually" or algorithmically. Intelligent document categorization has traditionally been the domain of library science, whereas algorithmic document classification has traditionally been the domain of informatics and computer science.

Texts, photos, music, and other media may all be categorized. Each type of document has its own set of categorization issues. Text categorization is assumed unless otherwise stated.

The tasks of automatic document classification can be divided such types: supervised document classification where some external mechanism (such as human feedback) provides information on the correct classification of documents, unsupervised document classification (also known as document clustering), semi-supervised document classification, in which sections of the documents are labelled by an external process, and unsupervised document classification, in which the classification must be performed totally without reference to external information. There are different software programmes available with varied licence models.

Documents can be categorized based on their subjects or other attributes (such as document type, author, year of publication, etc.). Only sub-subject categorization is discussed in the remainder of this article. There are two basic approaches to document subject classification: the content-based approach and the request-based approach. The categorization task is as follows: there are several papers and classifications. The aim is to discover pairs of "document, category" that are compatible with each other [5].

The general classification scheme consists of four stages:

1) pre-processing and indexing;
2) reducing the dimensionality of many features;
3) construction and training of the classifier;
4) assessment of the quality of classification.

At the stage of preliminary processing and indexing of documents, the features of the document, which are all its significant words or phrases, are formed. This stage includes tokenization, i.e. breaking the text into smaller objects, such as sentences, phrases or words; removal of functional words (semantically neutral, such as conjunctions, prepositions, articles, etc.) and morphological analysis, i.e. identification of parts of speech and stemmatization or lemmatization.

Reducing the dimensionality of a set of features is a process of giving weight to words, depending on their importance for the classification of the text, and further removing weightless terms from the set of features.

A threshold weight is set to remove terms below which terms are considered unimportant.

By reducing the dimensionality of the set of terms, you can reduce the effect of retraining - a phenomenon in which the classifier focuses on random or erroneous characteristics of educational data, and not on actually important ones [6].

We are only interested in the first and second stages. Therefore, let's consider them in more detail.

Pre-processing of text documents

Pre-processing your text simply means transforming it into a predictable and analyzable format for your work. A task is a mix of method and domain in this context. A Task may be ex-

DOI: 10.24025/2306-4412.2.2022.259408

tracting top keywords from Tweets (domain) using tfidf (method).

There are several methods for text pre-processing. Here are various techniques that you should be aware of, and we'll attempt to emphasize the significance of each.

Lowercasing – although it is sometimes forgotten, lowercasing of all text data is one of the simplest and most efficient forms of text preparation. It is applicable to the majority of text mining tasks and might be useful when your dataset is small. It also considerably improves the consistency of the predicted output [7].

Stemming – the practice of reducing word inflections (for example, troubled, troubles) to their base form is known as stemming (e.g., trouble). In this situation, the "root" might be a canonical variant of the original word rather than a true root word.

Stemming is a primitive heuristic procedure that cuts off the ends of words in the goal of appropriately changing them into their base form.

Lemmatization – on the surface, lemmatization appears to be similar to stemming in that the objective is to remove inflections and map a word to its root form. The main distinction is that lemmatization attempts to do it correctly. It doesn't merely cut things off; it truly changes the root of the term.

Stop words removal – language stop words are a group of frequently used terms. Stop words in English include "a," "the," "is," "are," and others. Stop words are used with the theory that by eliminating uninformative terms from the text, we can focus on key words instead.

Stop words removal was proved to be ineffective in classification systems, although being successful in search and topic extraction systems. However, it helps to reduce the number of features considered, keeping your models reasonably sized [8].

Normalization – text normalization is an often-overlooked pre-processing procedure. The process of converting a text into a canonical (standard) form is known as text normalization.

Text normalization has also been shown to be useful in evaluating highly unstructured clinical texts in which physicians take notes in unconventional ways. It's particularly effective for subject extraction if there are a lot of close synonyms and spelling discrepancies.

Unlike stemming and lemmatization, there is no universal method for normalizing texts. It is usually determined by the task.

Dictionary mappings (the easiest), statistical machine translation (SMT), and spelling-correction-based techniques are some typical ways to text normalization.

The optimal preparation for one job may become the greatest nightmare for another. Take note: text preparation cannot be transferred straight from job to task. Currently, there are two basic ways to text document pre-processing, known as document image generation.

1. Statistical approach. The methods of this approach, as a rule, consist in the analysis of the frequency of occurrence of words in texts in one or another of its variation and in the use of this information in the course of revealing and selecting representative signs of documents.

2. Linguistic approach. The main directions of this approach are morphological and syntactic analysis. Morphological analysis is important for Ukrainian language documents, as it allows to bring the processed features to some normal form. Such a reduction is necessary to recognize equivalent words with a common morphological basis. Parsing allows to automatically parse a text and build syntactic structures of its phrases, grouping words into classes within which they have similar syntactic behavior, these classes of words are called syntactic or grammatical categories (for example, parts of speech).

Linguistic and statistical methods complement each other because they use different approaches to information analysis, so the best results can potentially be achieved by combining both methods. However, currently linguistic methods (except for morphological analysis) are usually used only for research purposes and are not used in ready-made (commercial) information retrieval systems, as this approach is associated with much higher computational costs than statistical one. Moreover, a significant improvement in the quality of such systems in this approach compared to the statistical approach hasn’t been found, which is often due to the large noise component introduced by the errors of the automated parser. Therefore, the most common today is the combined method of pre-processing of text documents, which combines statistical approach and morphological analysis. This is a compromise between the complexity of calculations and the quality of results [9].

Since the linguistic approach to pre-processing of text documents is currently practically not used, as it requires large computational resources, we consider static methods that do not

DOI: 10.24025/2306-4412.2.2022.259408

require such large computing resources, and are common methods of pre-processing of text documents.

In the statistical approach, the pre-processing of text documents is to perform the following main steps:

1) formation of the space of documents signs;

2) display of images docskills in the space of their signs;

3) reduction of the initial avsharp signs of documents [10].

Word2vec is a technique for natural language processing published in 2013. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. As the name implies, word2vec represents each distinct word with a particular list of numbers called a vector. The vectors are chosen carefully to indicate the level of semantic similarity between the words represented by those vectors [11].

Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and creates a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in a vector space such that words that share common contexts in the corpus are located close to each other in the space [12].

Analysis of methods of text data processing

The method of determining the weights of the features of the document. The most common methods for determining the weights of document features are TF-IDF, a statistical indicator used to assess the importance of words in the context of a document that is part of a collection of documents or corpus. The weight (significance) of a word is proportional to the number of uses of that word in the document, and inversely proportional to the frequency of use of the word in other documents of the collection.

The TF-IDF indicator is often used to represent the documents of a collection as numerical vectors that reflect the importance of using each word from a set of words (the number of words in the set determines the dimension of the vector) in each document. Such a model is called a vector model and makes it possible to compare texts by comparing their representing vectors in a certain metric (Euclidean distance, cosine measure, Manhattan distance, Chebyshev distance, etc.), perform cluster analysis [13].

The first step in word processing is to calculate the TF-IDF weights for each word ω in each document.Φ_P^*.

TF is the ratio of the number of occurrences of the selected word to the total number of words in the document. Thus, the importance of the word within the selected document is assessed [14]:

$$tf(\omega, \Phi_P^*) = \frac{n_{\omega\Phi_P^*}}{n_{\Phi_P^*}}, \quad (1)$$

where $n_{\omega\Phi_P^*}$ is the number of occurrences of the word ω in the document;

$n_{\Phi_P^*}$ is the total number of words in the document.

IDF (Inverse Document Frequency) is the inversion of the frequency with which a word occurs in the documents of a collection. The use of IDF reduces the weight of commonly used words [14]

$$idf(\omega, A) = \frac{|\Phi_P^*|}{|\Phi_P^* \supset \omega|}, \quad (2)$$

where $|\Phi_P^*|$ is the total number of documents in a collection A;

$|\Phi_P^* \supset \omega|$ is the number of documents in which it occurs when $n_{\omega\Phi_P^*} \neq 0$.

So, the TF-IDF indicator is [14]

$$tfidf(\omega, \Phi_P^*, A) = tf(\omega, \Phi_P^*) * idf(\omega, A). \quad (3)$$

More weight of TF-IDF will be given to words with a high frequency of occurrence within the document and low frequency of use in other documents of the collection [14].

The method of assessing the proximity of text documents. Polythematic text documents can also be characterized by using document similarity features.

One of the methods to calculate relative estimates of thematic similarity includes the following steps:

- for each document, several (relatively small) documents representing its thematic environment are defined;
- constructed thematic environments are analyzed in order to form a set of key themes, which characterize the subject of the source doc-

DOI: 10.24025/2306-4412.2.2022.259408

ument in relation to other documents of the collection;

- obtained keyword sets are used to further calculate relative estimates of thematic similarity.

The need to find the thematic environment of a document is caused by the relativity of the concept of thematic proximity of documents, which is determined by the context in which the proximity of documents is assessed. For example, two documents characterizing, respectively, changes in stock quotes and changes in exchange rates, are likely to be considered thematically similar among a random set of documents, but at the same time they differ significantly within the limits of a highly specialized economic collection. Therefore, the assessment of thematic proximity is not only determined by the documents themselves, but also depends on the whole array of documents [15].

It is known that vocabulary and frequency of use of words depend on the subject. Therefore, only those words that are more specific to the subject of this document are taken into account to calculate estimates of thematic proximity. Such words are distinguished by the results of the analysis of the approximated thematic environment of this document.

Keyword selection method. Using the method of extracting keywords from the text allows to find the information you need in a short period of time.

Keyword is a word or constant expression of natural language, which is used to express some aspect of the content of the document; a word that has a significant semantic load. It can be a key when searching for information on the Internet or on a website.

There may be a synonymous relationship between keywords in terms of this search engine. The accumulation of keywords through meaningful analysis of texts or algorithmically, for example, by comparing words of a text with a fixed list of non-keywords, is an important step in choosing the source dictionary of information retrieval languages; the selected keywords are further combined into descriptors. Descriptor dictionaries provide links from keywords to corresponding descriptors. Keywords are the basis of search results. It should be remembered that a keyword can be not only a word but also a phrase [16].

Selection of keywords, extraction of the most important or characteristic fragments from one or many sources of information has become an integral part of our lives. Keyword selection tools are certainly useful, but their capabilities are limited to selecting and extracting the original snippets from the original document and combining them into a short text. The preparation of a summary aims to describe the main content of the text.

The main difference between the means of selecting keywords is that they, in fact, form a summary or set of citations from a particular material. Both types of presentation have two main purposes: to identify the most important idea of the full text and select keywords [17].

You can identify three key points that are not taken into account when selecting keywords from the text:

a) division into parts by keywords (by formatting);

b) selection of keywords (by weight);

c) formation of a summary document (in the form of statistics).

The method of comparison is the comparison of objects in order to identify common features or differences between them. The method of comparison is used in the process of generalization, when it is necessary to identify identities, coincidences and contradictions in the objects of study. Here identity is a full-fledged coincidence of all signs; coincidence is a coincidence of signs, starting from one; contradiction is when features of some objects are absent in others.

One of the methods used to identify clusters of documents that have similar properties only in some aspects, such as words or images, is *biclasterization*. The method is used for queries and indexing of full-text systems. Initial data is a matrix in which rows correspond to words and columns correspond to documents. To cluster documents, the number of word occurrences in a document, the total number of documents, and the number of documents containing a particular word are taken into account [18]. Thus, words can be clustered based on the documents in which they occur. Clusters are convenient for automatic construction of statistical thesauri, clarification of queries and automatic classification of documents, but it is impossible to perform meaningful text analysis using clusters.

Search for fuzzy duplicates involves clustering documents by the similarity of their certain characteristics, and the implementation of the algorithm consists of the following steps:

- canonization of steps – at this stage the text is cleaned of unnecessary words that do not

DOI: 10.24025/2306-4412.2.2022.259408

carry a meaning during comparison, i.e. the text is reduced to a canonical form;

- breaking the text into shingles;

- finding checksums – unique numbers, each of which corresponds to a text and the function of its calculation. Then, from the whole set of checksums (their number is equal to the number of words in the document minus (Ŵ-1, where EP is the number of words in the shingles), only those that are divisible by a certain selected premature number are selected;

- search for identical sequences - one shingle, which coincided during the selection, approximately corresponds to a predetermined number of identical parts in the full text [19].

Lexical and syntactic templates are characteristic expressions (phrases or inversions), constructions from the corresponding elements of natural language. Such templates allow to build a semantic model of the text. It is assumed that lexical relations in the document can be described with the help of templates (samples). This method uses a hierarchy of templates consisting of indicators of parts of speech and group symbols.

Thus, the analysis of the software used to detect text documents has shown the imperfection of the methods underlying modern automatic analysis systems. So, there is a need to create such software that would use the methods of deep linguistic processing. This will make it possible to compare text documents by content [20].

Analysis of word processing software

OBSERVER. This system offers an approach to using existing ontologies to access distributed and independently developed data repositories. It is assumed that there are many pre-created ontologies of subject areas, and the user does not have to "adapt" to a particular ontology. The user formulates his language query in terms of one or more ontologies, and the broker "searches" for relevant documents by translating the query into suitable ontologies and, if necessary, combining several ontologies to respond more accurately to the query [21].

TEXTANALYST. TextAnalyst is designed as a tool for analyzing the content of texts, semantic search for information, the formation of electronic archives, and provides the user with the following main features:

1) analysis of the content of the text with automatic formation of a semantic network with hyperlinks - obtaining a semantic portrait of the text in terms of basic concepts and their semantic connections;

2) analysis of the content of the text with automatic formation of the thematic tree with hyperlinks - identification of the semantic structure of the text in the form of a hierarchy of topics and subtopics;

3) semantic search, taking into account the hidden semantic connections of the query words with the words of the text;

4) automatic abstracting of the text - the formation of its semantic portrait in terms of the most informative phrases;

5) clustering of information - analysis of the distribution of text material by thematic classes;

6) automatic indexing of text with conversion to hypertext;

7) ranking of all types of information about the semantics of the text by "degrees of significance" with the possibility of varying the detail of its research [22].

ADVEGO PLAGIATUS - software product that allows to perform semantic analysis online, while finding the semantic core, and also allows to check the electronic document for uniqueness against web documents that are publicly available on the Internet.

The text of the article is analyzed by several algorithms:

- Shingles algorithm - exact matches of phrases, copypaste sources, as well as pages on which stolen texts are placed are checked.
- Algorithm of lexical matches - checks the similarity of a set of lemmas, terms and significant words, there are sources of rewriting, as well as pages that match the topic with the article being tested.
- Pseudo-unification algorithm - the presence of third-party characters and signs of text processing is checked by "enhancing" uniqueness services [23].

The main task of these systems is to search for information in large full-text arrays. The database of such systems can download any textual sources of information, including large ones: encyclopedias, directories, archives of periodicals, entire libraries of special literature, archives of corporate documents, specialized archives, such as historical, patent, court, transcripts, protocols and much more. In response to a specific request, the system issues a set of links. Next, the system must process each link and publish all relevant texts, i.e. the system must search not just documents, but the information contained in them.

DOI: 10.24025/2306-4412.2.2022.259408

DEEPDIVE is a system to extract value from dark data. Like dark matter, dark data is the great mass of data buried in text, tables, figures, and images, which lacks structure and so is essentially unprocessable by existing software. DeepDive helps bring dark data to light by creating structured data (SQL tables) from unstructured information (text documents) and integrating such data into an existing structured database. DeepDive is used to extract sophisticated relationships between entities and make inferences about facts involving those entities. DeepDive helps to process a wide variety of dark data and put the results into a database. With the data in a database, one can use a variety of standard tools that consume structured data.

DeepDive is a trained system that uses machine learning to deal with various forms of noise and imprecision. DeepDive is designed to make it easier for users to train the system through low-level feedback and rich, structured domain knowledge via rules. DeepDive wants to help experts who do not have machine learning expertise. One of DeepDive's key technical innovations is its ability to solve statistical inference problems at massive scale.

DeepDive differs from traditional systems in several ways:

- DeepDive asks the developer to think about features - not algorithms. In contrast, other machine learning systems require the developer to think about which clustering algorithm, which classification algorithm, etc. to use.
- DeepDive systems can achieve high quality: DeepDive has higher quality than human volunteers in extracting complex knowledge in scientific domains and winning performance in entity relation extraction competitions.
- DeepDive is aware that data is often noisy and imprecise: names are misspelled, natural language is ambiguous, and humans make mistakes. Taking such imprecision into account, DeepDive computes calibrated probabilities for every assertion it makes.
- DeepDive is able to use large amounts of data from a variety of sources. Applications built using DeepDive have extracted data from millions of documents, web pages, PDFs, tables, and figures.
- DeepDive allows developers to use their knowledge of a given domain to improve the quality of the results by writing simple rules that inform the inference (learning) process. DeepDive can also take into account user feedback on the correctness of the predictions to improve the predictions.
- DeepDive is able to use the data to learn "distantly". In contrast, most machine learning systems require tedious training for each prediction. In fact, many DeepDive applications, especially in early stages, need no traditional training data at all!
- DeepDive's secret is a scalable, high-performance inference and learning engine. For the past few years, we have been working to make the underlying algorithms run as fast as possible [24].

SCIKIT-LEARN (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms, including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interact with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is a NumFOCUS fiscally sponsored project.

Scikit-learn is largely written in Python, and uses NumPy extensively for high-performance linear algebra and array operations. Furthermore, some core algorithms are written in Cython to improve performance. Support vector machines are implemented by a Cython wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR. In such cases, extending these methods with Python may not be possible [25].

Research results. Methods of analysis of text documents significantly expand the possibilities of cataloging text documents. This becomes possible due to the fact that these methods make it possible to record the kinship of texts, the meaningful proximity of words, which is not expressed simply in their similar spelling. Also, fewer and fewer processing tasks require pre-encoding arrays for model training. Another important advantage is that the application of these methods to document analysis makes it scalable: it makes no fundamental difference for a machine to analyze ten documents or one million, while with manual coding the complexity of the analysis largely depends on the size of the text library to be analyzed. It also

DOI: 10.24025/2306-4412.2.2022.259408

eliminates the sampling problem in document-based studies, as automated analysis allows for continuous sampling.

Also, along with the advantages of automated document analysis, there are also disadvantages. The main one is that the performance of each technical task (for example, classification) requires training of a new model sharpened for this task. Training such models requires arrays of texts labeled by researchers. This also creates a second drawback - the models are not 100% accurate and are highly dependent on the quality of the data and the algorithm they were trained with. The third disadvantage is the technical resources required for text processing.

Discussion of results. There are many programs that perform linguistic processing of text, but none of them use mechanisms for extracting content from textual information.

The main disadvantages of the considered methods and applications for text processing are:

- the cumbersomeness of the computational process, which is associated with the need for constant optimization of the input document;
- to bring the document to the form when it can be calculated, text simplifications are performed. Therefore, the probability that important words for the text will be simplified is quite high;
- there is a serious danger that too much attention to textual simplification may ignore the context of words, regardless of where they occur;
- when using some methods for large texts, the size of the final array may be too large, which will require a large number of parameters during its further processing;
- during training, the methods do not take into account the order of words, and also use a limited number of context words at each training iteration.

In the process of researching the methods of analysis of text documents, a general procedure of text analysis is obtained, an overview of existing approaches to solving the analysis task is given, the main approaches used in the task of text analysis are described, the stages of the analysis process are defined, and the most common mathematical methods of text document analysis are considered. The revealed features of use, advantages and disadvantages of the specified methods allow to draw a conclusion about the need for further improvement of analysis algorithms based on the specified methods, which would be simple to implement, effective, have low computational costs during training and high quality of analysis in real tasks.

Based on the analysis of the obtained results, an algorithm will be developed, which, based on the methods analyzed above, should simplify the process of text analysis and allow the use of methods of computer analysis of text information to determine the attributes of a text document, based on which the analyzed text can be attributed to one or more groups, to which the specified attributes correspond.

Conclusions. The use of text information processing technologies is a promising area. The implementation of analytical processing of textual information is necessary in medicine, in business, and much more.

When processing textual information from a variety of disparate information resources, it is necessary to identify the following tasks:

1) selection of title, authors, keywords and construction of a conceptual model of the text;

2) integration into a full-text database;

3) search in full-text databases;

4) ensuring the relevance of the request;

5) reducing the amount of textual information and summarizing the text from several sources.

Search engines use an algorithm to search for duplicate text documents in order to index unique resources - that is, almost all branches of science and technology, due to the complexity of their organization, require intelligent processing of textual information. Only methods of processing textual information, the implementation of which involves all stages of analysis used in the analysis of language text, can claim to conduct a meaningful analysis of electronic text documents.

Almost all fields of knowledge, science and technology due to the complexity of their system organization, require intelligent data processing, therefore the processing of electronic documents today requires powerful mechanisms for analyzing textual information, which would ensure a complete and correct transition from natural language to mathematical models of text information processing.

As a result, it has been determined the need to develop a classification algorithm based on the mentioned approaches, such as lower case, root correction, stemming, lemmatization, stop word removal, normalization, which would be simple to implement, effective, have low computational costs during training and high

DOI: 10.24025/2306-4412.2.2022.259408

quality of classification in real tasks. Also, an approach to evaluating the thematic proximity of documents using feature space reduction is defined, and an algorithm for forming information-search attributes of documents for automatic clustering of documents is considered. The expediency of using methods of intelligent text analysis for this is considered.

In the future, an algorithm will be developed that will enable the application of methods of computer analysis of textual information, and software based on it. It will determine the attributes of a text document, based on which the analyzed text can be assigned to one or more groups, to which the specified attributes correspond.

References

[1] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Adison Wesley, Reading, MA, 1998.

[2] N. Kasyanchuk, and L. Tkachuk, "Protection of information in databases", in *Conf. VNTU of Electron. Sci. Publications, XLVIII Sci. and Tech. Conf. of the Faculty of Management and Information Security*, 2019, pp. 2419-2424 [in Ukrainian].

[3] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. *3rd ed.* Morgan Kaufmann, 2011.

[4] J. F. Luger, *Artificial Intelligence. Strategies and methods for solving complex problems. 4th ed.* Moscow: Izdat. Dom Williams, 2003.

[5] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algoritmhs*. MA, USA: Kluwer Academic Publisher Norwel, 2002.

[6] O. V. Havrylenko, Yu. O. Oliynyk, and G. V. Khanko, "Overview and analysis of text mining algorithms", *Project Management, System Analysis and Logistics*, no. 19, pp. 15-23, Kyiv, 2017 [in Ukrainian].

[7] M. Lemke, and G. Wiedemann, *Text Mining in den Sozialwissenschaften.* Springer Fachmedien Wiesbaden, 2016, pp. 397-419.

[8] I. V. Gushchin, and D. O. Sych, "Analysis of the influence of pre-processing of the text on the results of text classification", *Young Scientist*, no. 10, pp. 264-267, Kherson, 2018 [in Ukrainian].

[9] G. Salton et al., "Automatic text structuring and summarization", *Information Processing & Management*, vol. 33, no. 2, pp. 193-207, 1997.

[10] Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong, "Dynamic word embeddings for evolving semantic discovery", *WSDM 2018 Proc. 11th ACM Int. Conf. on Web Search and Data Mining.* Marina Del Rey, CA, USA, Febr. 5-9, 2018, pp. 673-681.

[11] Word2Vec Implementation. [Online]. Available: https://towardsdatascience.com/a-word2vec-implementation-using-numpy-and-python-d256cf0e5f28.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", *arXiv:1301.3781*, 2013.

[13] I. G. Oksanich, "Intellectual analysis of an array of text documents based on text mining technology", *Information Processing Systems*, pp. 139-143, Lutsk, 2013 [in Ukrainian].

[14] A. Yu. Zubrytskyi, "Intellectual system of text research and analysis", M.S. thesis, National Technical University of Ukraine "Ihor Sikors'kyy Kyiv Polytechnic Institute, Kyiv, Ukraine, 2019 [in Ukrainian].

[15] G. S. Linoff, and M. J. A. Berry, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 3rd ed.* NY, USA: Wiley Publishing inc., 2011.

[16] S. Deerwester et al., *Indexing by Latent Semantic Analysis*. Chicago, IL, USA: Graduate Library School University of Chicago, 1990.

[17] E. V. Bodyansky, and O. G. Rudenko, *Artificial Neural Networks: Architecture, Training, Application*. Kharkiv: TELETECH, 2004 [in Ukrainian].

[18] D. W. Lande, *Search for Knowledge on the Internet. Professional Work*. NY, USA: Williams, 2005.

[19] M. T. Hagan, H. B. Demuth, M. H. Beale, and O. De Jesús, *Neural Network Design*. 2014.

[20] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, vol. 60, no. 5, pp. 493-502, MCB University Press, 2004.

DOI: 10.24025/2306-4412.2.2022.259408

[21] A. Shalloway, and J. R. Trott, *Design Templates. A New Approach to Object-Oriented Analysis and Design.* NY, USA: Williams, 2002.

[22] "Library of software components of text analysis technology". [Online]. Available: https://www.analyst.ru/index.php?lang=rus &dir=content/downloads/.

[23] "Advego - content exchange №1". [Online]. Available: https://advego.com/.

[24] DeepDive [Online]. Available: http://deepdive.stanford.edu/.

[25] F. Pedregosa et al., "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

Д. О. Якименко, *аспірант (здобувач осв.-наук. ступеня доктора філософії),*
Є. Ю. Катаєва, *канд. техн. наук, доцент*
Черкаський державний технологічний університет
б-р Шевченка, 460, м. Черкаси, 18006, Україна

МЕТОДИ ТА ЗАСОБИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ТЕКСТОВИХ ДОКУМЕНТІВ

В роботі проведено огляд методів аналізу та обробки електронних документів. Проаналізовано методи аналізу текстових документів для вирішення задачі визначення тематичної спорідненості текстів.

Виконано огляд існуючих підходів до вирішення задачі класифікації. Описано основні підходи, що використовуються в задачі класифікації текстів; визначено етапи процесу класифікації та розглянуто найпоширеніші методи класифікації текстових документів. Розглянуто основні підходи до попередньої обробки тексту: Нижній регістр, Коренева корекція, Стемінг, Лематизація, Видалення стоп-слова, Нормалізація. Розглянуто переваги та недоліки кожного підходу. Розглянуто процедуру зменшення розмірності набору ознак із поділом на підпроцеси: обирання ознак та виділяння ознак. Розглянуто, в яких випадках кожен із підпроцесів є недоцільним для використання, та описано, які пошукові та фільтрові підходи і метрики є альтернативними або спорідненими для них.

Зроблено висновок щодо необхідності подальшого розроблення алгоритмів класифікації на базі зазначених методів, що були б простими в реалізації, ефективними, мали низькі обчислювальні витрати під час навчання та високу якість класифікації в реальних завданнях.

Визначено підхід до оцінки тематичної близькості документів з використанням редукції простору ознак і розглянуто алгоритм формування інформаційно-пошукових атрибутів документів для виконання автоматичної кластеризації документів. Розглянуто доцільність застосування для цього методів інтелектуального аналізу тексту.

Проаналізовано відкрите програмне забезпечення з використанням розглянутих методів.

Ключові слова: *ключові слова, аналіз тексту, пошук, текстові документи, класифікація.*

Стаття надійшла 20.05.2022
Прийнято 19.06.2022

DOI: 10.24025/2306-4412.2.2022.259408