An Approach to Processing News Text Messages Based on Markeme Analysis

Alexander Sychev [1]

[1] Voronezh State University, Voronezh, Russia

sav@sc.vsu.ru

Abstract. The complexity problem of automatic filtering of messages retrieved from online media platforms and social networks is discussed. The review of approaches in document representation, feature weighting schemes and feature selection techniques is provided. In the paper an approach to the text messages processing based on the markeme analysis is suggested. Markemes identification is based on calculating the Index of Textual Markedness (InTeM). Markemes are words most important for a particular text and occur with the frequency, which is higher than that of the words of the same length. Preliminary results of the exploratory study of the proposed approach, applied to the news messages classification and clustering, are presented and discussed.

Keywords: markeme, index of textual markedness, word form, term, message, skewness coefficient, classification, clustering, feature weighting, feature selection.

1 Introduction

The problem of automatic message processing is considered usually as complex one due to the following features of the content published on online media platforms and in social networks:

- "fuzzy" subject matter of message texts and comments;
- small length of text in a message or comment;
- heterogeneity of published texts in terms of stylistics, the level of literacy of the authors, etc.;
- a large volume of published messages and comments per unit of time.

A feature of news messages and user comments is that they are short texts. Short messages usually include text documents with an average length of less than 2000-3000 characters [1].

The problem of topic classification (rubrication) of short texts in Russian has been considered in many papers, for example, in [2] and [3]. The paper [2] presents the results of research in the field of classification of short text documents and analyzes classification methods based on the analysis of the distribution of lexical descriptors

of natural language. It also describes a method for assessing the informational significance of lexical units in natural language texts. In [3] the quality estimates for several methods of thematic classification (rubrication) of news messages, using various numerical estimates of information significance as features, were experimentally obtained on the "20 news groups" data set.

In general, the text classification pipeline includes the following steps: text features selection (extraction), dimensionality reduction, application of known classification techniques or development of new ones, evaluation of the classification model [4].

The feature selection which is reducing dimensionality, removing irrelevant data, and increasing the learning accuracy, is essential to tackling problems some problems, such as the curse of dimensionality and model overfitting, caused by the high dimensionality of data [5].

For the feature selection there are different techniques, e.g. TF or TF-IDF [6], Word2Vec [7], GloVe [8], FastText [9], Contextualized Word Representations [10] and their modification. All these techniques could be related to one of the two general feature selection approaches as follows: weighted words and word embedding [4].

Word embedding techniques require a huge corpus of text data sets for training [4]. As well, this approach cannot work for words missing from these data sets.

Weighted words technique use a simplified representation of the text, usually in the form of a bag-of-words (BOW) vector model, which allows to use of fairly simple and fast algorithms for processing text documents and messages. In the BOW model, the text is represented as a set of words, usually without taking into account the grammar or the order of words sequence, but using information about the frequency of words in the text. When solving the problem of documents classification, the word frequency of occurrence is used as a decisive feature for training the classifier. Exactly the word frequency is used in well-known methods for estimating the information significance of words in the text, for example, in TF-IDF metric based methods.

Some feature selection techniques can be not efficient for specific applications, depending on the goal and data set of the application. For example, GloVe does not perform as well as TF-IDF when used for short text messages [4].

Several widely used unsupervised and supervised term weighting methods on benchmark data collections in combination with SVM and k-NN algorithms were considered in [11]. As was stated in [11] the term weighting assignment is combined to improve both recall and precision measures by a multiplication operation from two factors: term frequency factor (tf) and collection frequency factor (idf). Several different collection frequency factors, namely, the multipliers of 1, a conventional inverse collection frequency factor (idf), a probabilistic inverse collection frequency (idf-prob), a χ^2 factor, an information gain (ig) factor, a gain ratio (gr) factor, an Odds Ratio (OR) factor, and proposed by authors novel relevance frequency (rf) factor were studied in experiments. In [12] five term scoring methods for automatic term extraction on different types of text collections were evaluated to investigate the influence of three factors in the success of a term scoring method in term extraction: collection size, background collection and the importance of multi-word terms. One important conclusion from [12] is that all term scoring methods could not demonstrate

the high level of performance for collections smaller than 1,000 words due to the prevailing of the frequency criterion in all methods.

According to [13] the dimensionality reduction techniques can be organized into three groups: feature selection (FS), feature projection, and instance selection. While the first two types of methods aim to reduce the dimensionality of the feature space, the third aims to reduce the number of instances used for training.

In FS methods, the resulting feature set is a subset of the initial feature set. The feature projection results in a new group of features mapped from the original features.

FS methods are usually classified into three categories: filter, wrapper, and embedded [14]. Filter methods are executed independently of the classifier learning activity. Wrapper methods encapsulate the classifier performance to assess the relevance of features or search for the most relevant subset of features. Embedded methods include FS as part of the training process.

A relevant advantage of selecting features is in the resulting feature set which is a subset of the original features. Each resulting feature preserves the meaning of the original feature.

In [15], the InTeM index (the Index of Textual Marking of a Word Form) is used to assess the degree of subjective weight of a word form in the text. The authors in [15] assume that each word form in the text has two parameters: frequency and length. At the same time, in their opinion, the frequency of the word form is a complex subjective-objective indicator, and the length of the word form is a simple objective – linguistic one. Hence, the subjective (i.e. meaningful) weight of a word form can be obtained by subtracting the simple objective factor (i.e. the weight of the word form according to its length) from the complex subjective-objective factor (i.e. the weight of the word form according to its frequency). The resulting value - the Index of Textual Markedness of a word form (InTeM) - will indicate the degree of subjective (textual) weight of a given word form for a given text. Thus, in fact, it is proposed to calculate the following indicator to assess the informational significance of the word form t_i from a text message *m*:

$$ITM_i = WF_i - WL_i$$

where

$$WF_i = \frac{\sum_{j=1}^{N_t} f_j - \sum_{j=1}^{i} f_j}{\sum_{j=1}^{N_t} f_j}, \; i \le N_t$$

$$WL_i = \frac{\sum_{j=1}^{L_m} f_j^{(len)} - \sum_{j=1}^{l} f_j^{(len)}}{\sum_{j=1}^{L_m} f_j^{(len)}}, \; i \le N_t$$

Word forms t_i from the text should be ranked in descending order of their frequency f_i in the general list of all word forms of the text. Frequency $f_j^{(len)}$ indicates the number of occurrences for all word forms *t*, having the length *j* in the text of message *m*. N_t is the total number of different word forms in the text message *m*. L_m is defined as the maximum word form's length in the text message *m*. The length of the t_i is denoted as *l*.

Word forms with the maximum value of ITM_i are called markemes and form the set of the most significant word forms for the author of the text.

In this paper the possibility of using the markeme model of texts for standard problems of classification, clustering and thematic categorization based on the example of a collection of news text messages is considered, and preliminary results of the exploratory study are presented.

2 Dataset

For the purposes of the study, a set ***M*** containing 760 text messages on several topics was formed. The set ***M*** included messages in approximately equal proportions of four topics marked up by experts manually. The average message size was 145.6 words. The total number of unique terms in the dictionary **D**, built from lemmas, which were extracted from their message texts, was about 12 thousand units, and 600 terms, which had a total frequency of occurrence for the entire set ***M*** at least 28, were selected for the study, The maximum total frequency of occurrence in the set ***M*** for a term from the dictionary was 1281.

Figure 1 shows the frequency distribution in the ***M*** of terms from vocabulary ***D*** along the length. As you can see, long terms (more than 10-12 characters of length) are found in messages with a low frequency, which reflects the objective linguistic realities. Within the framework of the markeme approach, the excess of the frequency of occurrence in the text *T* of a specific term f_i of length *l* relative to the frequency $f_l^{(len)}$ typical for terms of length *l* gives grounds for including it in the set of markemes $\boldsymbol{MK_T}$ of a given text *T*. One should note that the calculation of ITM_i in the framework of the markeme approach does not take into account the form of the frequency distribution function over the length of word forms.

This kind of distribution can also be calculated for each text message individually.

In the study, all terms with ITM_i index value exceeded zero were identified as markemes.

Table 1 shows the number values of markemes identified from text messages and averaged by topic categories. N_{MK1} is the number of markemes identified from the global (message collection) frequency distribution of terms along the length, N_{MK2} is the number of markemes identified from the local (i.e. inside individual message) frequency distribution of terms along the length in the message.

It is obvious that those markemes that are characterized by a relatively high frequency f_M in messages and a relatively large value of the distribution asymmetry index (skewness) *Sk* over the entire set of messages ***M*** will be useful in further study.

Table 1. Average number of terms and markemes in a message

Topic	Average number of markemes		Average number of different terms in a message	Average sum of term frequencies in a message
	$< N_{MK1} >$	$< N_{MK2} >$		
Medicine	6.3	3.3	39.2	54.2
Accidents	8.8	5.1	54.0	74.1
Politics	10.8	6.2	58.6	90.0
Sports	8.1	4.2	50.3	75.3
Mean:	8.6	5.1	51.0	74.2

In this study two indicators of asymmetry for markemes were considered:

- Sk_1 is the skewness of the markeme distribution across the four topics in the ***M*** message set;
- Sk_2 is the skewness of the markeme distribution over the entire set of messages ***M*** as a whole.

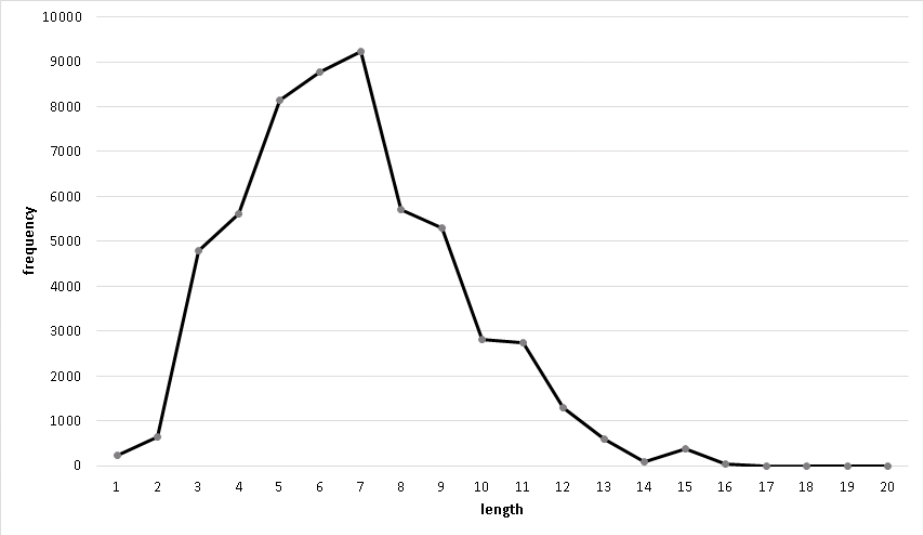

Fig.1. Frequency distribution of dictionary terms by length.

Figure 2 shows the scatter matrix for the 380 markemes selected from the ***M*** message set. The values of the Sk_1, Sk_2 parameters and the total frequency of occurrence (for the entire set ***M***) f_{Mi}, calculated for the markeme Mk_i, were used.

The $Sk_2 - Sk_2$ section of the scatter matrix shows the histogram of Sk_2 values distribution. The highest distribution density is observed near the value 10 of Sk_2 variable. The $f - Sk_2$ section proves this observation. For the Sk_1 variable values the distribution density is concentrated in the vicinity of the value 2.

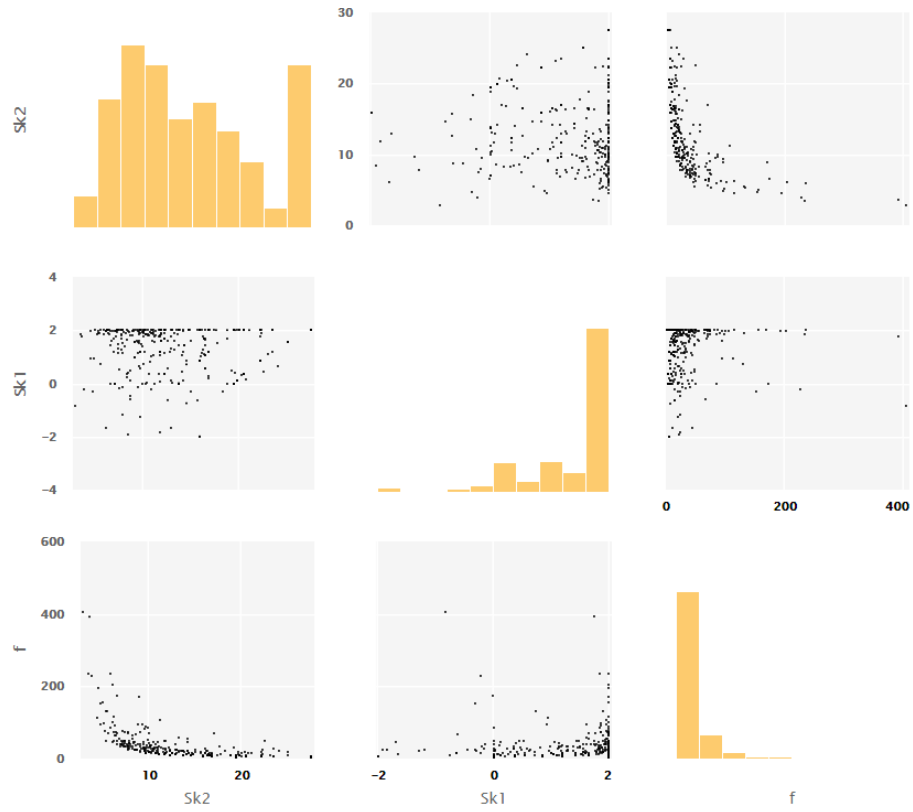

Fig.2**.** Scatter matrix for the markemes set.

One can expect that a good set of markemes will turn out from candidates that have:

– the frequency of occurrence f_{Mi} noticeably differs from the minimum values;
– the topic parameter Sk_1 tends to the limit value 2 (good topic specificity) or the Sk_2 parameter value is in the vicinity of 10 (a good indicator of the markeme specificity in ***M***).

3 Experiment

For the experiment, the filter strategy for feature selection was chosen to reduce the dimensionality of the features space. Features filtering was realized using f_{Mi}, Sk_1, and Sk_2 parameters.

Markemes Mk_i, for which the parameter values satisfied two conditions: $f_{Mi} > 6$, $Sk_1 \geq 1.9$, were selected from the total set of markemes identified from ***M***. Table 2 provides a list of the 58 markemes identified from the ***M*** set in this way.

Table 2. A list of selected markemes and sorted by topic.

Message topics			
Medicine	Accident	Politics	Sports
заболевание	авария	администрация	болельщик
здравоохранение	водитель	власть	ворота
клинический	иномарка	возглавлять	завоевать
лечение	легковушка	глава	команда
неделя	личность	государственный	первенство
пациент	мужчина	депутат	сборная
поликлиника	очевидец	должность	сезон
	пассажир	заместитель	соперник
	погибнуть	нацпроект	соревнование
	полицейский	начальник	спорт
	случиться	образование	спортсмен
	столкновение	общественный	спортсменка
	убийство	обязанность	турнир
	экспертиза	председатель	факел
		президент	футболист
		реализация	чемпионат
		руководитель	
		сельский	
		социальный	
		территория	
		чиновник	

Computational experiments on the classification and clustering of the ***M*** message set were carried out using the set of selected markemes.

3.1 Messages Classification

For the messages classification a naive Bayesian classifier, supplied with an assessment of the quality of classification by cross-validation method (10 folds), was used. The obtained estimates of the quality are given in Table 3, where rows indicate the classifier predictions for corresponding topic and columns are related to true topics in

tested data. The Accuracy value was 82%. Accuracy was calculated as ratio: (sum of correct classifier predictions) / (total number of testing examples). For comparison, Table 4 provides the estimates for the same classification, except that all the terms (600 units) from the ***D*** dictionary were used as attributes of the frequency vector of messages. The Accuracy value was 89%.

Table 3. Performance evaluation for message classification based on the frequency vector with markeme attributes.

True/Prediction	**True topic 1**	**True topic 2**	**True topic 3**	**True topic 4**	**Class precision**
Prediction topic 1	153	46	47	7	60,5%
Prediction topic 2	4	153	7	0	93,3%
Prediction topic 3	3	8	131	6	88,5%
Prediction topic 4	0	1	8	186	95,4%
Class recall	95,6%	73,6%	67,9%	93,5%	

It is noteworthy that although, in general, the markemes list representation of messages worsened the Accuracy value by about 7% , there was an improvement in recall and precision in some topics. For example, the recall of the topic 1 ("Medicine") improved significantly (with a significant decrease in the precision value), for topics 2,3 ("Incidents", "Politics") there was an improvement in the precision value (while the recall value decreased). The significant decrease of classification Accuracy (table 3) is due to decreasing in the class precision for the topic 1 and the class recall for topics 2,3. This effect can be considered as a payment for essential features space dimensionality reduction. One can see in the table 2 that the list of selected markemes-features is too short to provide the high level of class accuracy. Perhaps a more flexible scheme for selecting f_{Mi}, Sk_1 and Sk_2 parameters values could improve the situation. The dimensionality reduction of the feature space for solving the problem of message classification has happened to be more than 10 times.

Table 4. Performance evaluation for message classification based on the frequency vector with terms-attributes from the dictionary ***D***.

True/Prediction	**True topic 1**	**True topic 1**	**True topic 1**	**True topic 1**	**Class precision**
Prediction topic 1	130	8	16	2	83,3%
Prediction topic 2	9	190	10	0	90,9%
Prediction topic 3	20	10	164	5	82,4%
Prediction topic 4	1	0	3	192	98,0%
Class recall	81,3%	91,4%	85,0%	96,5%	

Table 5. Performance evaluation for the message classification based on a frequency vector with term attributes from ***D***, filtered by $Sk_1 \geq 1.9$.

True/Prediction	**True topic 1**	**True topic 2**	**True topic 3**	**True topic 4**	**Class precision**
Prediction topic 1	139	29	32	1	69.2%
Prediction topic 2	10	170	15	0	87.2%
Prediction topic 3	10	8	139	5	85.8%
Prediction topic 4	1	1	7	193	95.5%
Class recall	86.9%	81.7%	72%	97%	

For comparison purposes, there was carried out a messages classification, based on a frequency vector with attribute terms selected on the basis of $Sk_1 \geq 1.9$ filter (Sk_1 factor to some extent could be considered an analogue of the *IDF* factor in algorithms with *TF-IDF*). In fact, the boolean conversion of the Sk_1 factor was used as collection (topic) frequency factor (IDF). The total number of terms selected from the ***D*** was 152. The experiment results are given in Table 5. The value of the Accuracy was 84.3%.

3.2 Messages Clustering

For the set of markemes (given in Table 2) as attributes of message vectors K-means clustering was carried out. Table 6 summarizes the results of this experiment. As you can see from the table, the markeme set allows to accurately identify the thematic core in a set of messages for each topic, but at the same time most of the messages from the topic class subset are thematically vague. An increase in the clustering recall index can be achieved by softening the constraints (for parameters f_{Mi} and Sk_1) when selecting markems. It is worth noting that the clustering result is quite sensitive to the choice of the initial conditions for the clustering algorithm.

Table 6. Performance evaluation for message clustering based on the frequency vector with markeme attributes.

Topic	**Number of messages in topic subset of M**	**Recall**	**Precision**
1	160	36,3%	97,5%
2	210	32,9%	99,5%
3	190	32,6%	100,0%
4	200	56,5%	99,5%

The implementation of clustering with a markeme list representation of messages in topics unknown in advance situation, makes it impossible to calculate the Sk_1 pa-

rameter. In this case, one can suggest to calculate the Sk_2 parameter, which is not tied to specific topics. The topics of the clusters identified in this way could be determined by calculating the correlation coefficients between the frequency-dominant markemes within the identified clusters. Table 7 shows a fragment of the table of correlation coefficients for markeme pairs (for $Sk_1 \geq 1.9$). When calculating the correlation, the frequency of occurrence of markemes in messages (760 frequencies totally) was used as coordinates of the markeme vector.

Table 7. Markemes correlation coefficients ($Sk_1 \geq 1.9$)

Markemes pair		**Correlation coefficient**
реализация	нацпроект	0,77
клинический	здравоохранение	0,67
сельский	территория	0,60
защитник	воронежец	0,59
случай	пациент	0,54
факел	болельщик	0,51
чемпионат	спортсменка	0,51
факел	воронежец	0,51
факел	штрафной	0,51
спортсменка	завоевать	0,50
болельщик	защитник	0,49
болельщик	воронежец	0,49
штрафной	воронежец	0,48
образование	муниципальный	0,46
команда	болельщик	0,46
иномарка	водитель	0,45
защитник	штрафной	0,44
ворота	штрафной	0,43
чемпионат	завоевать	0,43
образование	нацпроект	0,43
факел	защитник	0,43
главный	оборина	0,41
убийство	мужчина	0,41
поликлиника	здравоохранение	0,41
участок	результат	0,40

4 Conclusion

The preliminary results of the study presented in this paper allow us to draw several conclusions regarding the possible use of the markeme approach in text messages classification, clustering and thematic categorization.

1. Based on the method of identifying the word form as markeme in the text, it should reflect the degree of subjective (author's) weight of this word form for a particular text. Since a lot of news text messages come from different online platforms, it is basically impossible to talk about a single authorship in the stream of news messages. In this case, the analysis of the text of a news messages is significantly different from the analysis of a large text, for example, a literary work. Of course, markeme analysis is more suitable for use as a work tool for linguistic research of texts.
2. From the point of view of the messages classification and clustering performance evaluation, both representation of a text message by a vector of markemes frequency and representation it by a vector of terms frequency based on the calculation of the TF factor give quite comparable results. Some degradation in classification accuracy can be considered as a payment for essential features space dimensionality reduction. Perhaps, a more flexible scheme for selecting f_{Mi}, Sk_1 and Sk_2 parameters values could improve the situation.
3. From the computing point of view, the markeme model of messages has the advantage: to identify the markeme from the text, it is enough to have the body of text itself only, but not the entire set of texts, as it is required, for example, when computing the TF-IDF factor. Of course, the text size is should be sufficient enough to calculate the term frequencies.
4. The markemes as a features space basis could be considered as a good choice for filter strategy in the feature selection procedure to cut the effects of curse of dimensionality and model overfitting. The choice of markemes based on the threshold values of the f_{Mi}, Sk_1 and Sk_2 parameters can be used to construct an "orthogonal" basis (in some sense) in the feature space of terms for evaluating, for example, the "blurring" degree of existing topic sections and the need to reorganize their structure. Markemes can also be used for keywords generation and annotating news messages.
5. The threshold values for f_{Mi}, Sk_1 and Sk_2 in fact are considered as tuning parameters in the feature selection procedure to improve both recall and precision measures. In this way the choice of the values mentioned above will depend on the target recall and precision levels.

Of course a relatively small collection of news texts and 4 topics are used for experiments, but the paper presents preliminary results of exploratory research. Further experiments will engage extended both the size of collection and the number of topics. More experiments and comparison with existing weighting schemes for improving document representation are expected further.

References

1. Lande D., Morozov, A., Darmokhval A.: An Approach to Identifying Duplicate Messages in News Information Streams (2006). URL http://dwl.kiev.ua/art/rdcl/rcdl2006.pdf.
2. Mbaykodzhi, E., Dral A., Sochenkov, I.: Short Text Messages Classification Method. Journal of Information Technologies and Computing Systems, issue 3, pp.93-102. (2012)
3. Zhebel V., Zharikova, S.-N., Sochenkov, I.: Feature Selection for Text Classification of a News Flows Based on Topical Importance Characteristic. Artificial Intelligence and Decision Making, issue 3, pp.52-59 (2019). (in Russian). https://doi.org/10.14357/20718594190306.
4. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes L, Brown, D.: Text classification algorithms: a survey. Inf.Switz. 10. (2019). https://doi.org/10.3390/info10040150.
5. Pintas, J., Fernandes, L., Garcia, A.: Feature Selection Methods for Text Classification: a Systematic Literature Review. Artif.Intell.Rev. (2021). https://doi.org/10.1007/s10462-021-09970-6.
6. Salton, G.; Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 1988, 24, pp. 513–523. (1988)
7. Goldberg, Y., Levy, O.: Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. arXiv 2014, arXiv:1402.3722. (2014)
8. Pennington, J., Socher, R., Manning, C.: Glove:Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; vol. 14, pp.1532–1543. (2014)
9. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv, 2016, arXiv:1607.04606. (2016)
10. Melamud, O., Goldberger, J., Dagan, I.: context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, 11–12 August 2016, pp. 51–61 (2016). https://doi.org/10.18653/v1/K16-1006.
11. Lu Y., Lan M., Su J., Tan, C.: Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 31, no. 04, pp. 721-735. (2009). https://doi: 10.1109/TPAMI.2008.110
12. Verberne, S., Sappelli, M., Hiemstra, D., & Kraaij, W.: Evaluation and Analysis of Term Scoring Methods for Term Extraction. Information Retrieval, 19(5), pp. 510-545 (2016). https://doi.org/10.1007/s10791-016-9286-2.
13. Mirończuk, M., Protasiewicz, J.: A Recent Overview of the State-of-the-art Elements of Text Classification, Expert Systems with Applications, vol. 106, 2018, pp. 36-54. (2018) https://doi.org/10.1016/j.eswa.2018.03.058.
14. Kumar, V., Minz, S.: Feature Selection: A literature Review. The Smart Computing Review, vol. 4., pp.211-229. (2014) https://doi.org/10.6029/smartcr.2014.03.007.
15. Faustov, A., Kretov, A.: The Concept of Markeme and Interim Results of Markeme Analysis of Russian Literature. Proceedings of Voronezh State University. Series: Linguistics and intercultural communication, issue 4, pp.16-32 (2017). (in Russian)