Comparative Research of Index Frequency - morphological Methods of Automatic Text Summarisation*

Alexander Osochkin
osa585848@bk.ru

Vladimir Fomin
vv_fomin@mail.ru

Xenia Piotrowska
krp62@mail.ru

Herzen State Pedagogical University of Russia
Saint Petersburg, Russian Federation

Abstract

Automatic analyses perspective and natural language processing (NLP) is being researched in the subject paper. The modified quantitative approach using a collocation algorithm is presented. This approach eliminates previously discovered issues of text processing using the vector model in the number of cases of thematic classification. The author's frequency method of extracting a set of numerical indicators taking into account the morphological features of words, as well as collocation between terms in text is proposed. The quantitative technology of automatic thematic classification using indicators which reflect morphological and parsing text features, methods of parameterization, text indexing, algorithms of artificial intelligence analysis and knowledge extraction is presented.

The efficiency and advantage of the regression decision tree method in the tasks of finding of significant frequency indexes and their logical representation are demonstrated.

The results of comparative experiments to assess the effectiveness of classification of natural language text data, using the author's, vector and set-theoretic models of text representation are stated.

Keywords: *Text-classification, machine learning, NLP, text-mining, regression decision trees.*

1 Introduction

In the process of NLP methods evolution, the separate area, studying the language based on statistical patterns with the inclusion of algorithms and models of linguistic and semantic analysis, is appeared. [Martin Jurafsky, 2017], [Kang Y. et al, 2020]. Quantitative methods use latent semantic links between text elements and enhance the capability [Johnson, 2020], [Khalezova et al, 2020], [Kashcheyeva, 2013] of the statistic text analysis approach.

The creation of different text representation models aimed at specific tasks solution has led to the dynamic development of the quantitative text analysis approach. [Ribeiro et al., 2020], [Maheshan et al, 2018], [Martin Jurafsky, 2017], [McCann et al, 2017]. A significant part of the methods is based on the quantitative approach, containing complex computational linguistic text analysis algorithms.

The researches on the quantitative approach improvement considering national, professional, linguistic language features remains relevant. The design of adaptation methods and their application, address the shortcomings and limitations while NLP digitalization is promising.

The separate field of possible NLP application in education is assistance to an actor in the solution of educational tasks in informational and communicational redundancy of electronic environment.

These are principally new opportunities for automatic interaction with text information of educational resources, identifying its content and quality characteristics.

This approach provides expert support in the search for and extraction of knowledge from a variety of information resources of the global environment, increase the criticality and practicality of thinking, the formation of new competencies, the formation of new knowledge, etc. The complexity of the knowledge extraction from text is caused by the specifics and variability of language, the human brain works, the dynamics of thinking development.

The part of tasks of substantial analysis of texts can be hardly formalized and described and requires applying of mathematical tools of uncertainty, statistical methods and artificial intelligence.

The main technology of the quantitative approach is the data representation model - Bidirectional Encoder Representations from Transformers (BERT), which has shown high efficiency in solving a wide range of tasks [Ribeiro et al., 2020], and formed the basis of digital natural language processing services. The BERT method is used in advanced technologies such as “Google AI”, “Microsoft Azure Text Analysis”, “Amazon Comprehend”, “Facebook RoBERTa AI”, etc.

The last researches in the field of NLP analysis have shown that this approach has a range of significant shortcomings.

The critical shortcomings of modern models of text processing based on the BERT method were presented at one of the largest conference Association for Computational Linguistics [Kolesnikova, 2016], [Ribeiro et al., 2020], dedicated to artificial intelligence development and processing of computer texts, took part in 2020.

The most important recommendations on improving text processing methods are defined in the problems of development of grammatical and lexical text coherence, including the stability of word combinations (collocation) in texts.

The research area is the improvement of text representation models, procedures of the digital analytical indexes formation and extraction of indicators, the development of algorithms of linguistic analysis using artificial intelligence methods.

The research aims to evaluate the effectiveness of the identification technology and clustering subject field based on the quantitative approach using collocation algorithms and regression decision trees.

2 Model of text representation

Mathematical models using for text conversion allows characteristics extracting from text data that can be represented as numeric parameters or indexes [Beel et al, 2017], [Belinkov Bisk, 2018], [Wang Zhu, 2019]. Within the framework of the quantitative approach, there were developed numerous models of text representation:

1. frequency: [Piotrowska X., 2005], [Grekhov A., 2012], [Osochkin et al, 2020],
2. frequency-morphological: [Fomin et al, 2018], [Piotrowska X., 2014],
3. vector: [Harish et al, 2012], [Mohit et al, 2018],
4. thematically-vector: [Moschitt, 2004], [Allahyari et al, 2017], [Martin Jurafsky, 2017], [McCann et al, 2017], [Moulton Jiang, 2018],
5. set-theoretic: [Marcus, 1967], [Devlin Chang, 2018], [Maheshan et al, 2018], [Belinkov Glass, 2019], [Ribeiro et al., 2020].

Despite the existence of quite a large number of approaches to text conversion, as well as their modifications, all models we suggest to divide into two types:

- vector models
- set-theoretic models.

Advanced computational models used in the computer industry of text processing will be considered and compared further:

- the vector model,
- the set-theoretic model,
- the author's set-theoretic model with collocation.

2.1 Vector model of text representation

A vector text representation model is a mathematical model where each text's object is matched with a vector that can reflect different linguistics characteristics.

The vector's coordinates can be different text elements: individual words, concepts, noun groups, sentences, semantic groups of sentences, paragraphs, so as word's semantics: meaning or fields of science. This model was proposed in the works of Salton [Salton et al, 1994] as an alternative to lexical contextless indexing.

Often vector model is called a thematic vector model because the basis of text class division is rooted in a semantic word's meaning. This meaning can characterise the field of science.

The vector model of text representation proved its worth as it could diminish the main disadvantages of frequency and theoretical models of data representation including the homonym problems, so as the semantic meaning of sentences consideration problem.

Because of the processing time problems, the vector model is mainly applied to small texts processing.

A significant disadvantage of vector model is the lack realization of structural characteristics of isolating, agglutinating, and inflecting types of languages.

We choose the "Word2Vec" library was chosen as the main tool for studying the vector model of text representation, because of:

- more than 40 languages, including Russian, support;
- no need in supervised studying;
- the using of the embedded model of replacing associative words, homonyms (Bag of Words).

2.2 Set theoretic model of text representation

Set-theoretic model supposes that every text consists from terms (word, n-grsam, sentences), possessing common characteristics and unique traits. The main concept of the subject model is the reflection of different text characteristics in relative indicators, to which mathematical methods for identification of common and unique characteristics of each sample analysed text are applied. At the heart of set-theoretic models the toolbox for frequency analysis, measures and metric proximity (Dice, Ochiai, Jaccard, Simpson etc.) is used.

As a metric of text converting into a set of numerical indicators, the interval similarity coefficient of Jaccard was taken. This coefficient is the simplest to calculate, and its values are equivalent in particular cases to other similarity metrics (Sorensen, Sokal-Snit).

The algorithm of indicators calculating based on the similarity coefficient of Jaccard is presented in detail in the work of R. Moulton [Moulton Jiang, 2018]. The generalized Jaccard coefficient of comparing the proximity of two words A and B, is calculated by the formula:

$$K = \frac{n(A \cap B))}{n(A) + n(B) - n(A \cap B))))} \tag{1}$$

Similarity indexes can be calculated for words, n-gram words, sentences, etc. The freely licensed Python library "Jaccard-index", which is able to calculate the similarity index between texts, is used to implement calculations of the Jaccard similarity index. In this paper a word was chosen as the main unit of analysis. Set-theoretic model with collocation extraction.

Following the recommendations of M. Ribeiro [Ribeiro et al., 2020], we set the task of studying the impact on improving the accuracy of text classification not only in thematic aspects but also of morphological features of words, as well as collocation between terms in the text.

In this context, we developed the FaM software, which was described in details in [Harish et al, 2012]. FaM uses the author's algorithm for text representation as a frequency-morphological set of indicators, considering collocations, and can be used to improve the accuracy of classification in NLP purpose. The mathematical model of text representation with collocation extractions considers texts only as interconnected sequences of terms. It is assumed that taking into account sustainable links in phrases and the relationship between text elements will create a more accurate model of text representation.

In order to consider a collocation, FaM calculates a number of special indicators based on the frequency of use of a sequence of words (n-grams) in the text that have certain characteristics.

To consider this FaM's feature, the author's algorithm is used. It was implemented with the help of several morphological libraries, which removes the function words and words that were not in semantic connection with the sentence members when calculating n-gram sequences. A normalized text is formed as a dataset, where each word is described as an object with its features: part of speech and morphological characteristics. For each sequence of objects and for each combination of their morphological characteristics, a separate frequency indicator is calculated. This indicator is presented as a sequence occurrence counts of objects in a normalized text, divided by the total number of objects.

Thus, the set of n-gram indicators is determined by the type of natural language and by the length of the n-sequence. The total set of extracted bigrams for the Russian language can reach more than 200 indicators.

A key factor in improving the efficiency of classification is the conversion of text into a set of numerical indicators using frequency-morphological analysis. Morphological analysis is performed by a special hybrid algorithm that uses two well-known modules for morphological analysis: Natural language processing (AOT)[Fomin et al, 2018] and http://www.solarix.ru/ [Osochkin et al, 2020]. The author's algorithm embedded in FaM allows you to use two libraries simultaneously, allowing you to get information about the analyzed word, its semantic relationship with other words in the sentence, and conduct morphological, syntactic, and frequency analysis. The algorithm aimed to find a semantic connection conducts syntactic analysis, which identifies parts of speech and functional words in a sentence and builds a syntactic tree. In the next stages, the algorithm searches for words that are syntactically related to the subject or predicate in the sentence and checks for semantic connections. The semantic relationship is checked by synthesizing a new sentence without the analyzed word, building a new syntactic tree, and analyzing nodes changes in the tree. If the context changes in the tree nodes associated with the deleted word did not occur, the normalized text in the form of a dataset does not include this word.

Using the term collocation we focused on the concept of a stable semantically interconnected binary phrase (sequence of bigrams) in the Russian academic texts.

3 Normalization and relevance of indicators

Almost all intelligent text analysis packages perform preprocessing to normalize the received data. Text preprocessing allows you to get more accurate and reliable data and a more detailed description of the features of the text.

The main procedure that allows you to significantly reduce the size of the vector space by reducing the variation of words is the lemmatization procedure. The variance reduction also has a positive effect on the vectors indexes, reducing the dimension of the vector space. The NLTK4Russian library was chosen to texts lemmatizing [Moskvina et al, 2016]. Normalization of data, received in the framework of the set-theoretic model, is carried out using the TF-IDF technology, the Scikit-learn library [Roul et al., 2017].

TF_{ij} indexes are defined as the frequency of word's use in the analysed text, regarding the total number of words in the text:

$$TF_{ij} = \frac{f_{ij}}{fi_1 + fi_2 + ... + fi_n}, i = 1,\bar{m} \qquad (2)$$

where TF_ij is the index for the j-th word in the i-th text, f_{ij} is the frequency of use of the f_j-th word in the i-th text, and f_n is the n - th word in the i-th text.

The TF-IDF method [Roul et al., 2017], [Salton et al, 1994] calculates the value of the j-th term IDF_{ij} in the i-th text as the product of the frequency of term usage in the tf_{ij} document and the normalized inverse frequency of term content in the documents.

$$IDF_{ij} = TF_{ij} * log\left(\frac{|D|}{Df_i}\right), i = 1,\bar{m}, j = 1,\bar{n} \qquad (3)$$

where D is the total number of documents in the collection. D_{fi} – the number of documents in which the term f_j occurs.

This approach allows to determine the importance of the term in the entire collection of analyzed documents. Terms with high uniqueness, which are less common in other documents, and often occur in the analyzed document, have the highest value.

4 Artificial intelligence algorithms

For tasks of parametric analysis, regression, classification, identification and knowledge extraction, NLP uses an extensive toolkit of artificial intelligence, which uses machine learning methods and algorithms (neural networks, genetic algorithms, metric algorithms, reference vectors, decision trees, etc.). Based on the study of classification methods in the previous works [Fomin et al, 2018], [Osochkin et al, 2020], [Osochkin et al, 2021] it was concluded that the use of regression decision tree algorithms for identifying the style and gender of the author of literary works was effective. The efficiency is due to obtaining higher classification accuracy when using small text bodies, compared to neural networks and the support vector method. A significant advantage of all methods of regression decision trees is the representation of results in the form of a hierarchical set of logical rules "if-then", which allows meaningful identification, interpretation, verification of classification results; it impacts assessment and value significance of each indicator. The variety of algorithms for constructing regression decision trees (Random Forest, ID3, C4.5, C5.0, CRT, CHAID, etc.) allows you to use the full potential of statistical analysis in the framework of a quantitative approach to natural language text processing. In this paper, several algorithms were used for building decision trees in the IBM SPSS data analysis package [Fomin et al, 2018].

5 Research material

The texts corpus is represented by various educational materials, divided into 10 fields of sciences (clusters): IT, History, Chemistry, Jurisprudence, Biology, Medicine, Pedagogy, Physics, Philosophy, Economics.

Text materials are represented by various types of documents, including training manuals, textbooks, lecture notes, abstracts, scientific articles, dissertations, dissertation abstracts, etc. (Table 1).

Table 1: Educational materials

Field of science	Number of author's abstracts and dissertations	Number of training manuals	Number of articles	Total
IT	200	500	300	1000
History	200	500	300	1000
Chemistry	200	500	300	1000
Jurisprudence	200	500	300	1000
Biology	200	500	300	1000
Medicine	200	500	300	1000
Pedagogy	200	500	300	1000
Physics	200	500	300	1000
Philosophy	200	500	300	1000
Economics	200	500	300	1000

6 The experiment of the field of science classification

The comparative experiments on the effectiveness of two classical methods of field of science identification and the collocation method proposed by the authors were conducted.

The task of the experiment is to classify the corpus of texts by field of sciences (without taking into account the type of document). The calculated data is extracted using the previously described text transformation models (vector, multiple, collocation).

The "exhaustive CHAID" algorithm using the Gini coefficient was chosen as the main algorithm for building the decision tree. The choice of this algorithm is due to the complexity of classification by more than 10 clusters at the same time, high accuracy [Fomin et al, 2018], [Kang Y. et al, 2020], [Yang et al, 2020] and a decrease in the tree dimension because of non-binary tree division algorithms.

The ratio of the training and test samples is 50%, without observing the proportions of the cluster dimension. The maximum tree depth is 10. Due to the small number of objects, the number of texts that can be located in the father node for division into child nodes is 2 objects. The Pierson Chi-square test is used to check the hypothesis of finding common characteristics. Since all indicators are relative, the node split significance criteria are 0.005. We studied the influence of the setting "number minimum of objects in a node" on the classification accuracy.

Table 2 shows data on the experiments which set the minimum number of objects in the node of the decision tree algorithm: from 50 to 10 with a decrease of 5.

The accuracy of identification of the field of science is reflected based on the parameters of three mathematical models of texts.

Table 2: The accuracy of the field of science identification

Minimum samples split	Vector model	Set-theoretic model	Set-theoretic model with collocation extraction
50	79,16%	70,80%	85,61%
45	79,85%	73,56%	88,15%
40	80,15%	75,10%	90,23%
35	82,56%	78,96%	91,83%
30	85,98%	80,73%	92,85%
25	88,98%	84,83%	95,15%
20	91,35%	87,25%	95,67%
15	93,15%	89,89%	97,80%
10	95,71%	92,51%	98,20%

This setting significantly affects the accuracy of the classification, due to the processing of statistical outliers and the creation of the rules for a small number of unique texts.

The text models obtained using the collocation method showed better overall classification accuracy compared with the vector and set-theoretic text representation models. The minimum difference in accuracy between the methods has been achieved with the algorithm setting to split node if at least 10 objects have fallen into it and it is 2.49% higher than the vector text representation model has and is 5.59% higher than the set-theoretic model has.

The maximum difference in accuracy is 10,08% compared with vector model and 15,13% compared with the set-theoretic model (with the algorithm setting to split node if at least 40 objects have fallen into it).

The average accuracy using collocation method has become 6,51% higher than using the vector model and 11,32% higher than using the set-theoretic text representation model. The greatest number of mistakes were made when identifying texts related to the field of "Jurisprudence". The texts of Jurisprudence were included into the field of History, Philosophy and Economics.

This mistake in identifying the field of science is due to the specifics of jurisprudence, which includes Roman law, civil law, international law, criminal law, tax code, etc. Due to this feature, some texts were included in similar clusters that use common terminology. From the classification results table, it can be also seen the tendency of increasing the accuracy of classification, with a decrease in the number of texts required to split a node into children nodes.

This tendency is due to the presence of statistical outliers, deviations in the content of texts, as well as the detailing of individual groups of educational materials.

Table 3 shows the detailed results of the classification using a mathematical model of text representation based on the set-theoretic text representation with collocation.

Table 4 shows the indicators and their values that were used by the algorithm of the exhaustive CHIAD decision tree construction method.

The results of the experiments indicate the effectiveness of the collocation method, which was achieved through the use of more complex indicators, such as bigrams of parts of speech, morphological features and syntactic characteristics.

To evaluate the significance of these indicators, we also decided to conduct additional researches with settings, where the best result was obtained (the minimum number of samples

Table 3: Classification of educational materials by ten fields of sciences

Field of science	IT	History	Chemistry	Jurisprudence	Biology	Medicine	Pedagogy	Physics	Philosophy	Economics	Acc
IT	493	0	0	1	0	0	2	1	1	2	98,60%
History	0	496	0	0	0	0	1	0	3	0	99,20%
Chemistry	0	0	491	0	5	4	0	0	0	0	98,20%
Jurisprudence	0	6	0	479	1	0	2	0	7	5	95,80%
Biology	0	3	9	0	486	2	0	0	0	0	97,20%
Medicine	0	0	1	4	2	492	0	0	1	0	98,40%
Pedagogy	1	8	0	0	0	0	487	0	1	3	97,40%
Physics	0	0	1	0	0	0	0	498	0	1	99,60%
Philosophy	0	4		0	0	0	0	0	496	0	99,20%
Economics	0	2	0	2	0	1	1	1	1	492	98.40%

Table 4: Ten most important indicators while identifying field of science

No	Indicator	Value
1	Latin symbols per sentence	7,21
2	Numerals per sentence	6,81
3	Personal pronouns per sentence	6,74
4	Noun in accusative form + verb 1-st form	6,21
5	Noun + adjective	5,61
6	Adverbial participle per sentence	4,35
7	Participles per sentence	4,25
8	Adjective + unanimated noun	4,23
9	Adjective + animated noun	4,21
10	Adverb+adverb	4,09

required node is 10). Three experiments were conducted, with the removal of one of the significant indicators from the general dataset. When removing the "Latin characters per sentence" indicator, the overall accuracy decreased by 8,68%, to 89,52%. Table 5 shows the most significant 10 recalculated indicators.

When removing the "Noun + Adjective" indicator, the overall accuracy decreased by 9,54%, to 88,66%.

Table 6 shows the 10 most significant indicators.

When removing the indicator "Adverb + Adverb", the overall accuracy decreased by 2,74% to 95,46%. Table 7 shows the 10 recalculated most significant indicators.

As a result of three experiments considering the removal of indicators with different significance, the accuracy significantly decreased, which indicates the importance of using these indicators when classifying texts by field of science.

Three indicators include bigrams, the removal of which reduced the accuracy by from 2% to 9%, and the CHAID algorithm itself replaced these indicators in the classification with other

Table 5: Significant indicators while building the tree “ Latin symbols per sentence”

No	Indicator	Value
1	Numerals per sentence	9,02
2	Adverbs per sentence	8,07
3	Noun in accusative form + verb 1-st form	7,61
4	Participles per sentence	7,51
5	Average words length	6,82
6	Average sentence length	5,51
7	Personal pronouns per sentence	5,43
8	Noun + noun	5,02
9	Adjective + animated noun	4,84
10	Adverb+Adverb	4,26

Table 6: Significant indicators while building a tree without indicator “Noun + Adjective”

No	Indicator	Value
1	Latin symbols per sentence	7,24
2	Numerals per sentence	6,85
3	Personal pronouns per sentence	6,79
4	Noun in accusative form + verb 1-st form	6,32
5	Adverbial participle per sentence	5,02
6	Participles per sentence	4,61
7	Adjective + animated noun	4,39
7	Adjective + unanimated noun	4,27
9	Adverb + adverbn	4,23
10	Noun + noun	4,02

Table 7: Significant indicators while building a tree without indicator “Adverb + Adverb”

No	Indicator	Value
1	Latin symbols per sentence	7,21
2	Numerals per sentence	6,82
3	Personal pronouns per sentence	6,78
4	Noun in accusative form + verb 1-st form	6,21
5	Adverbial participle per sentence	5,67
6	Participles per sentence	4,34
7	Adjective + animated noun	4,31
7	Adjective + unanimated noun	4,26
9	Noun + noun	4,17
10	Noun + adverb	4,15

bigrams. The extraction of indicators related to the use of parts of speech and their features from the text made it possible to significantly increase the accuracy of classification when identifying the subject field of educational material.

Experiments have shown the advantage of the model with collocation extraction in comparison with the two classical models. The obtained results indicate that the text representation model based on the set-theoretic representation of the text with collocation is effective.

Conclusions

The use of a quantitative approach with collocation extraction allows to increase the accuracy of the subject field identification. The experiment results of the subject field identification accuracy evaluation confirmed the effectiveness of the proposed modification of the set-theoretic model of text processing.

Algorithms of frequency-morphological extraction of numerical indicators and the formation of text indexes that reflect the frequency of individual parts of speech and n-gram parts of speech use can be successfully used to identify the field of science. Experiments have confirmed an increase in the total classification accuracy using collocation compared to the vector model of text representation.

Using the set-theoretic model with collocation extraction allows eliminating some of the disadvantages of the BERT data representation model that were identified earlier. In conjunction with the methods of regression decision trees, the potential of text mining can be expanded. We also plan to conduct further experiments aimed to analyse the accuracy of identifying the emotional colours of messages

Acknowledgements

The research was supported by the Ministry of Science and Higher Education of the Russian Federation (project No. FSZN-2020-0027)

References

[Allahyari et al, 2017] Allahyari M., Pouriyeh S., Assefi M., Safaei S., Trippe E. Mehdi A., (2017) A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. Computer Science, 260.

[Beel et al, 2017] Beel J., Langer S., Gipp B. (2017) TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users' Personal Document Collections. iConference Preliminary Results Papers, 1-8.

[Belinkov Bisk, 2018] Belinkov Y. Bisk Y. (2018) Synthetic and natural noise both break neural machine translation. International Conference on Learning Representations. URL: https://arxiv.org/abs/1711.02173

[Belinkov Glass, 2019] Belinkov Y. Glass J. (2019) Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics. 7, 49–72.

[Devlin Chang, 2018] Devlin J. Chang M-W (2018) Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. Google AI Language. URL: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

[Fomin et al, 2018] Fomin V., Pavlova T., Osochkin A. (2018) Eksperimenty text-minig po klassifikacii tekstov v ramkah zadach personalizacii obrazovatel'noj sredy// Informatizaciya obrazovaniya i nauki [Text-minig experiments on the classification of texts in the framework of the problems of personalization of the educational environment // Informatization of education and science]. Vol. 2 (38). 2018. Pp. 38-50 (In Rus.)

[Grekhov A., 2012] Grekhov A. (2012) Kvantitativnyy metod: poisk latentnoy informatsii. Vestnik Nizhegorodskogo universiteta im. Lobachevskogo. [Grekhov A.V. Quantitative method: searching for latent information. Vestnik of Lobachevsky State University of Nizhni Novgorod] 1 (3), 94-100. (In Rus.)

[Harish et al, 2012] Harish B., Manjunath S., Guru D. S. (2012) Text Document Classification: An Approach Based on Indexing. International Journal of Data Mining Knowledge Management Process, 1, 43-66. DOI: 10.5121/ijdkp.2012.2104

[Jadhao Agrawal, 2016] Jadhao A. Agrawal A. (2016) Text Categorization using Jaccard Coefficient for Text Messages. International Journal of Science and Research (IJSR), 5, 2046- 2050.

[Johnson, 2020] Johnson Kh. (2020) AI researchers create testing tool to find bugs in NLP from Amazon, Google, and Microsoft. VB TRANSFORM URL:https://venturebeat.com/2020/07/09/ai-researchers-create-testing-tool-to-find-bugs-in-nlp-from-amazon-google-and-microsoft/

[Kang Y. et al, 2020] Kang Y., Cai Zh., Tan Ch-W., Huang Q., Liu H. (2020) Natural language processing (NLP) in management research: A literature review. Journal of Management AnalyticsMay, 7(12), 1-34 pp. DOI: 10.1080/23270012.2020.1756939

[Kashcheyeva, 2013] Kashcheyeva A. (2013) Kvantitativnyye i kachestvennyye metody issledovaniya v prikladnoy lingvistike. Sotsial'no-ekonomicheskiye yavleniya i protsessy = Kascheyeva A.V. Quantitative and qualitative methods of research in applied linguistics. Socio-economic phenomena and processes], 3 (49), 1-8. (In Rus.)

[Khalezova et al, 2020] Khalezova N., Piotrowska, X., Terbusheva, E., Piotrovskaya, V., Neznanov, N. (2020) Cross-sectional Study of Clinical and Psycholinguistic Characteristics of Mental Disorders in HIV Infection. R. Piotrowski's Readings in Language Engineering and Applied Linguistics (PRLEAL-2019). Proceedings of the III International Conference on Language Engineering and Applied Linguistics. CEUR-WS, 2552, 161-178 URL:http://ceur-ws.org/Vol-2552/Paper14.pdf

[Kolesnikova, 2016] Kolesnikova O. (2016) Survey of Word Co-occurrence Measures for Collocation Detection. Comp. y Sist. [online]. 2016, 20(3), 327-344. ISSN 1405-5546. https://arxiv.org/abs/1809.04052

[Maheshan et al, 2018] Maheshan M., Harish B. S., Revanasiddappa M. B. (2018) Indexing-Based Classification: An Approach Toward Classifying Text Documents Information Systems. Design and Intelligent Applications, 1, 894-902. DOI: 10.1007/978-981-10-7512-488

[Marcus, 1967] Marcus S. (1967) Algebraic Linguistics; Analytical Models. Academic Press, New York, 1967, XIV +, 254.

[Martin Jurafsky, 2017] Martin D. Jurafsky D. (2019) Speech and Language Processing. An introduction to natural language processing, computational linguistics, and speech recognition. Third Edition draft. 621. URL: https://web.stanford.edu/ jurafsky/slp3/ed3book.pdf

[McCann et al, 2017] McCann B., Bradbury J., Xiong C., Socher R. (2017) Learned in translation: Contextualized word vectors. Advances in Neural Information Processing Systems, 6294–6305.

[Mohit et al, 2018] Mohit I., Wieting J., Gimpel K., Zettlemoyer L. (2018) Adversarial example generation with syntactically controlled paraphrase networks. Proceedings of NAACL-HLT, 1875–1885.

[Moschitt, 2004] Moschitt A. (2004) Complex Linguistic Features for Text Classification: a comprehensive study. Lecture Notes in Computer Science. 26 European Conference on IR Research, Sunderland, UK, 181-196.

[Moskvina et al, 2016] Moskvina A. D., Orlova D., Panicheva P. V., Mitrofanova O. A. (2016) Development of a parser kernel for the Russian language based on NLTK libraries//Computer linguistics and computational ontologies. Works of the XIX International Joint Scientific Conference "Internet and Modern Society" (IMS-2016). - St. Petersburg. ITMO University. Page.44–45. http://openbooks.ifmo.ru/ru/file/4103/4103.pdf

[Moulton Jiang, 2018] Moulton R. Jiang Y. (2018) Maximally Consistent Sampling and the Jaccard Index of Probability Distributions. International Conference on Data Mining, Workshop on High Dimensional Data Mining 2018, 347–356. URL: https://arxiv.org/abs/1809.04052

[Osochkin et al, 2020] Osochkin A., Piotrowska X., Fomin V. (2020) Comparative Research of Index Frequency - Morphological Methods of Automatic Text Summarisation. NESinMIS-2020. Proceedings of the XV International Conference "New Educational Strategies in Modern Information Space", Saint-Petersburg, Russia, March 25, 2020. CEUR-WS, Vol-2401, 73-86. URL: http://ceur-ws.org/Vol-2630/paper8.pdfURL: http://ceur-ws.org/Vol-2630/paper8.pdf

[Osochkin et al, 2021] Osochkin A., Piotrowska X., Fomin V. Automatic Identification of Authors' Stylistics and Gender on the Basis of the Corpus of Russian Fiction Using Extended Set-theoretic Model with Collocation Extraction // Glottometrics 50, RAM-Verlag, 2021. Pp. 76-89

[Piotrowska X., 2005] Piotrowska X., (2005) Computer-assisted language learning. The quantitative-linguistic basis of CALL methods. // Quantitative Linguistik / Quantitative Linguistics, 2005: 897-908

[Piotrowska X., 2014] Piotrowska X., (2014) A Survey of Text mining. Izvestia: Herzen University Journal of Humanities Sciences. 168, 128-134. (In Rus.)

[Ribeiro et al., 2020] Ribeiro M., Tongshuang W., Guestrin C., Singh S.(2020) Beyond Accuracy: Behavioral Testing of NLP Models with CheckList», Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4902–4912 URL: https://www.aclweb.org/anthology/2020.acl-main.442

[Roul et al., 2017] Roul R., Sahoo J., Kushagr A. (2017) Modified TF-IDF Term Weighting Strategies for Text Categorization. Proceedings of 14th IEEE India Council International Conference (INDICON), 1-6.

[Salton et al, 1994] Salton G., Allan J., Buckley C. (1994) Automatic structuring and retrieval of large text files. Communications of the ACM, 37(2), 97-108

[Wang Zhu, 2019] Wang Y. Zhu L. (2019) Research on improved text classification method based on combined weighted model. National Natural Science Foundation of China, 7(11), 783-796.

[Yang et al, 2020] Yang Zh., Dai Zi., Yang Y., Carbonell J., Salakhutdinov R., Quoc Q. (2020) XLNet: Generalized Autoregressive Pretraining for Language Understanding. Proceedings of Advances in Neural Information Processing Systems 32 (NIPS 2019) URL: https://arxiv.org/abs/1906.08237