European Proceedings of
Computers and Technology
EpCT

www.europeanproceedings.com

e-ISSN: 2672-8834

DOI: 10.15405/epct.23021.3

HMMOCS 2022
International Workshop "Hybrid methods of modeling and optimization in complex systems"

SYSTEM OF AUTOMATED TEXT MESSAGES CLUSTERING BY SEMANTIC PROXIMITY BASED ON NLP AND MACHINE LEARNING METHODS

Iuliia Khudonogova (a), Leonid Lipinskiy (b), Anastasiya Polyakova (c)*
*Corresponding author

(a) Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky rabochy prospect, Krasnoyarsk, 660037, Russian Federation
(b) Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky rabochy prospect, Krasnoyarsk, 660037, Russian Federation
(c) Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky rabochy prospect, Krasnoyarsk, 660037, Russian Federation, polyakova_nasty@mail.ru

Abstract

At the present moment the relevance of natural data processing problem solving is rising. A massive data amount of text data has been accumulated in recent years. Classical analytical methods, such as machine learning methods, are not capable of dealing with raw text data, which complicates the analysis significantly. Therefore, a modern set of methods of text data vectorization has been developed, which gained massive popularity in the recent years for analyzing text data, specifically for solving text clustering problem, as one of the most relevant text data related analytical problems. In this paper, a few of these methods were researched; a new dictionary optimization approach has been proposed and tested on the real text datasets; a number of conclusions on the effectiveness and of the methods for the given tasks has been made. For the future work a more thorough research on the dictionary optimization scheme (genetic algorithm parameters) and vectorization method are planned.

Keywords: Text message clustering, semantic proximity, machine learning

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

1. Introduction

To date, the science of artificial intelligence has made great strides in many areas of its application. However, one of the unexplored areas remains human-machine communication. This topic has spawned a new field at the intersection of linguistics, computer science and artificial intelligence - natural language processing.

One of the main and most important tasks in this area is the task of text messages clustering, on which this work is focused.

2. The Problem of Natural Language Processing and Approaches to Text Vectorization

2.1. Relevance of the problem of natural language processing

Natural language processing (NLP) is an interdisciplinary area of knowledge, technologies and methods that focuses on working with natural, i.e., human languages, extracting knowledge from data representing messages in natural languages using various methods.

NLP can be considered one of the directions for the development of artificial intelligence, since it allows applying machine learning algorithms to text and speech. NLP techniques can be applied to many different tasks: speech recognition, text summarization, chat-bot creation, auto complete, and many others. On the basis of NLP methods, quite a lot of apps and technologies have been developed today, which are widespread among ordinary users, for example, apps such as Cortana, Yandex Alice, Siri are very popular. In addition to that, we all use, albeit unknowingly, NLP technologies when working with email, since today it is very popular to use NLP technologies when sorting emails to categories, sorting out spam messages etc. One of the most promising, at the same time least studied problems in NLP is the problem of analyzing text messages.

The text, in the broadest sense of the word, is one of the most accurate, complete and at the same time universal and understandable ways of presenting human thought. It is accurate in the presentation of thoughts, since in most developed languages there is a sufficiently large set of tools for expressing thoughts, capable of conveying subtle shades of meanings. The completeness is provided by the practical unlimitedness of this toolkit, as well as the possibility of languages for evolution. In addition, the textual presentation of information can be called more understandable compared to other representations, such as pictures, layouts, formulas, tables, since the apparatus of text comprehension is embedded in people from an early age and does not require any additional knowledge and skills for understanding. Based on the facts presented above, the usefulness and importance of text messages and their analysis is proven.

The main value of text data is their number. Over the years of its existence, mankind has accumulated a huge number of texts in various human languages, a very large number of them can now be found in electronic format (e-books, archival articles, etc.).

With the advent of new technologies for communication and information transfer, the rate of generation of new text data has arisen at times. Examples include microblog posts, book and film reviews, reviews of events and places, publications in online post.

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

It is especially useful to analyze multiple texts, rather than individual statements. Solving the problem of analyzing sets of statements is useful for many real-life problems. By analyzing several related statements, one can draw conclusions about the opinions, interests of various groups of people, which is a very useful skill in the fields of commerce, politics, education and many other fields of activity. Collecting and analyzing feedback from product users, researching the characteristics and opinions of the target audience, determining the criteria for buyers product choosing – these are just some of the possible tasks in the solution of which the analysis of text messages is of great benefit.

The complexity of working with text data is as great as the value of the knowledge extracted from this data. Indeed, extracting semantics from textual data is a non-trivial task for an intelligent system, and direct application of classical methods of working with numerical data, for example, machine learning methods, to this task is impossible. When working with texts, it is required to carry out a thorough multilevel preprocessing, and only then proceed to the analysis.

2.2. Texts vectorization and preprocessing

Before proceeding with the application of analysis methods, the texts must be vectorized.

To solve the problems of text analysis, data must be presented in numerical format. There are several ways to translate text statements into numerical expressions. They can be classified in various ways, for example, according to the basis of work; within the framework of this classification, frequency based and prediction based approaches can be distinguished (Sarwan, 2017). According to the data format, methods can be divided into working with single words and working with whole statements. There are two main technologies for vectorizing text: word2vec and doc2vec. They differ significantly from each other in many aspects. However, in order to fully understand the description of these technologies, let us first describe the basic necessary NLP terms.

Before analysis, the text should be split into text units - tokens. The token, depending on the technology used, can be symbols (letters that make up a word), words themselves, sentences, etc. The tokens look different in word2vec and doc2vec, but in general they can be described as a parsed text unit.

The complete set of tokens is called a dictionary. The dictionary can be sorted, changed in order to optimize the solution of the problem (if the token is not useful for solving a specific problem, it can be excluded from the dictionary).

A document (statement, message) is a set of tokens characterized by certain semantics, i.e., it carries some specific meaning. As a document, you can consider a phrase, sentence or a set of coherent sentences (text). For example, a tweet, a paragraph from a book, or an entire article or review can be considered a document.

The corpus is the general set of all the documents under consideration. Documents inside one corpus can be of different lengths (for example, considering the corpus of movie reviews, one document can consist of just a couple of words - "Good movie", while the other can be expanded into several paragraphs)

As stated above, vectorization techniques can be divided into two broad categories. Let's consider them in more detail.

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

2.3. Word2vec Technology

Word2vec is one of the first vectorization technologies developed. Word2vec is a general name for a set of different vectorization models that use different learning algorithms but obey the same workflow. It belongs to the type of predictive vectorization. There is also a software of the same name from the Google team, released in 2013 and became the first large-scale and popular implementation of word2vec technology.

The essence of the technology is as follows: word2vec models work with a corpus of texts, collecting a dictionary from all tokens, while tokens in this case are individual words. Thus, it is not the vectorization of texts, sentences or phrases that is performed, but the vectorization of words, which follows from the name of the method.

The principle of work of word2vec models is as follows: it is necessary to find connections between word contexts under the assumption that words in similar contexts tend to mean similar things, i.e., be semantically close. More formally, the problem is as follows: maximizing the cosine proximity between vectors of words (scalar product of vectors) that appear next to each other and minimizing the cosine proximity between vectors of words that do not appear next to each other. «Next to each other» in this case means in close contexts (Nikitinsky, 2015).

The vectorization process itself can be performed in two ways. The first method is called Continuous Bad of Words (CBoW). It is used for the tasks of predicting a word relative to its surrounding by context. The second method is called Skip-gram.

The main problem with this technology is that the corpuses have to be very large in order to provide quality results. Research shows that in order to obtain adequate analysis results, a corpus must consist of at least 10,000 documents.

This technology is used for a wide, but very specific range of tasks, such as clustering concepts and terms by semantic proximity, word prediction tasks based on semantic proximity (for example, for the previously mentioned auto-complete mechanism), as well as for problems of restoring blanks in a text.

As can be seen from the listed tasks, word2vec is focused on working with words, and in most tasks, it is not connected texts that are analyzed, but individual words. To use word2vec for analyzing text corpuses, it becomes necessary to work with multidimensional matrices of vectors, which overly complicates the analysis. Therefore, for tasks in which it is necessary to consider a message as an individual without diving into the details of the characteristics of each word in this message, the doc2vec technology was developed.

2.4. Doc2vec technology: Bag of words and TF-IDF

By analogy with the previous technology, doc2vec (doc-to-vec, document-to-vector) is a vectorization technology, its main difference from word2vec is that it matches a numeric vector not to a word in a statement, but to a whole statement in a corpus. The corpus consists of documents (statements), then using vectorization methods from the doc2vec technology, these documents are vectorized and further analysis is carried out on the basis of documents, and not individual words in documents.

Let's describe the main methods of vectorization used in the doc2vec technology.

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

The first method is a bag of words, it can rightfully be considered the easiest vectorization method to understand and execute. However, despite the seeming primitiveness, this method in a number of problems is more effective than others. The bag of words method is widely used for various tasks of text analysis.

The essence of the method is as follows: from the text documents under consideration, the entire existing structure is cleared, for example, paragraphs, punctuation marks, and all the unique words in them - tokens - are extracted from the continuous texts. This is the process of tokenization (Yordanov, 2019).

Then all unique tokens from all considered documents are collected into a single array and numbered (for example, in alphabetical order). Such an array is called a corpus dictionary.

When the corpus dictionary is compiled, for each document it is calculated how often each of the words entered in the dictionary occurs in the document. Based on the results of this process, we obtain numerical vectors characterizing each statement (document) in the corpus. The length of the vector corresponds to the length of the dictionary (see Figure 1).

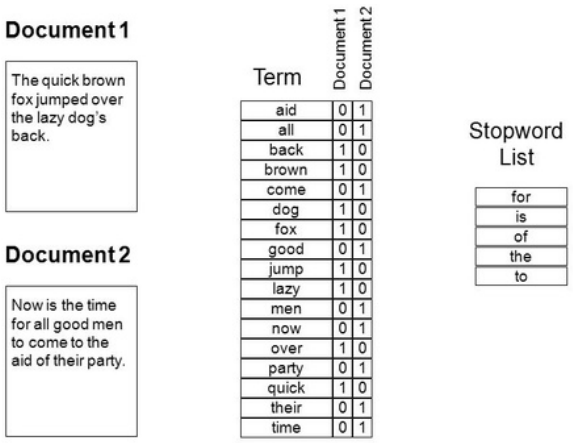

Figure 1. Bag of words vertorization

In certain cases, you can use binary vectorization using a bag of words. Then the vector will consist only of 0 and 1, 0 will mean that the word from the dictionary is not present in the document, 1 - that it is present, regardless of how many times this word is repeated in the document.

However, if this operation is performed without preliminary processing, the dictionary will be littered with elements that are unimportant for analysis - punctuation marks, prepositions, conjunctions, pronouns (in most cases they are not an important part of the document, but there are exceptions), special characters.

In order to prevent the influence of all of the above and other problems on the vectorization process, it is necessary to carry out preprocessing of the data before starting tokenization.

As part of preprocessing, you need to do several steps with tokens:

- Convert all characters to the same case (this should be done first);
- Remove punctuation and other special characters from text;

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

- Clear the text from the so-called "stop words". For NLP problems, there are public stop word dictionaries for various languages, including Russian and English. However, it is recommended to supplement this vocabulary in accordance with the data of a specific task to improve the efficiency of vectorization;
- Carry out the procedure for normalizing words (lemmatization). Every word has a normal form or lemma. For nouns, this is the nominative singular; for adjectives - nominative, singular, masculine; for verbs (as well as participles and gerunds) - an infinitive in an indefinite form of an imperfect form. To reduce the form of a word to its lemma, there are special libraries that include huge dictionaries with various forms of words and their lexical characteristics (grammemes). Sometimes the lemmatization procedure is replaced by the stemming procedure. The difference is that lemmatization brings a word to the form of the lemma - with its normal, standard form. Stemming, on the other hand, is the process of finding the “base” of a word (its unchangeable part that expresses the lexical meaning). The stem of a word does not always correspond to its root (Yordanov, 2019).

The Figure 2 below shows the data preprocessing diagram.

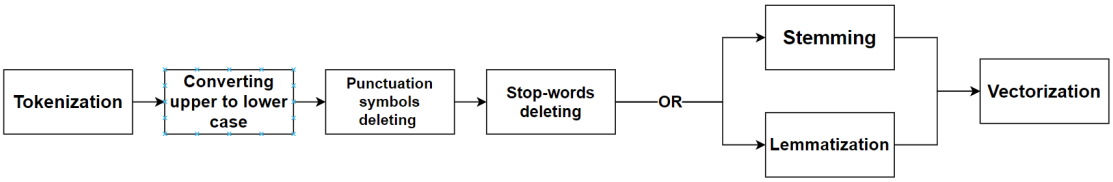

Figure 2. Preprocessing scheme

After completing all the above procedures, it is possible to generate a complete noise-free dictionary.

The word vectors formed according to such rules are ready-made material for the operation of machine learning algorithms when solving various problems (Müller & Guido, 2017).

The main disadvantages of the “bag of words” method are that, firstly, the final matrix of vectors turns out to be very sparse, since it contains a large number of zeros (the more documents in the corpus, the more zeros in the vectors, if any dictionary optimization was carried out). Secondly, the resulting vectors do not contain any information about the order of words in sentences, and information about the grammatical form of a word is also lost, although sometimes these two indicators can hide a significant part of the semantics of the phrase.

Now let's look at the second vectorization method related to the doc2vec technology - TF-IDF. The abbreviation TF-IDF stands for term frequency - inverse document frequency, formally defined as a statistical measure used to assess the importance of a word in the context of a document that is part of a document collection or corpus (Selivanov, 2020). As the name suggests, the method consists of two stages or steps, we will consider each in more detail.

Step TF (term frequency) - at this step, a measure of the frequency of occurrence of a certain term (word) in a certain document is calculated.

$$tf(t,d) = \frac{n_t}{\sum_k n_k}$$

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

where the numerator is the number of occurrences of the word t in document d, the denominator is the total number of words in this document.

Upon completion of this step, each document and each word is assigned its own TF score.

The inverse document frequency (IDF) step is a measure of the importance of a term. It is calculated as follows:

$$idf(t, D) = \log\left(\frac{|D|}{|\{d_i \in D | t \in d_i\}|}\right)$$

where the numerator is the number of documents, the denominator is the number of documents in the D corpus in which the word t occurs (it is not equal to zero).

Multiplying these two estimates, we obtain the TF-IDF measure for each word in each document:

$$tf\ idf(t, d, D) = tf(t, d) * idf(t, D)$$

High weight in TF-IDF will be given to words with a high frequency within a specific document and with a low frequency of use in other documents.

The doc2vec technology has several drawbacks, for example, when forming a document vector, the word order in a sentence or phrase is not taken into account in any way, although sometimes this factor plays a very important role in the semantics of a message, since, depending on the word order, a message can take on polar opposite meanings (Heidenreich, 2018).

2.5. Application of the genetic algorithm to the dictionary optimization

The process of creating a dictionary for text vectorization was described above. However, it is not always possible to obtain the best clustering results with a complete dictionary.

In this paper a variant of dictionary optimization using a genetic algorithm is proposed. The general scheme of the genetic algorithm is shown in the Figure 3 below.

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

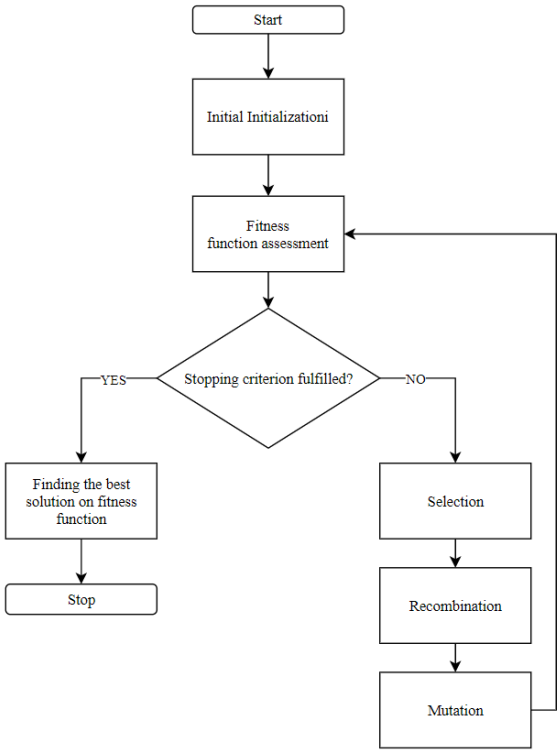

Figure 3. Genetic algorithm scheme

At the first step, an initial population of individuals is created in a random way, then the operators of selection, crossing and mutation are applied. During the selection stage, based on the value of the fitness function, the most adapted individuals are selected. The most popular selection operator is tournament selection (Andreev, 2011).

Then the crossing operator forms the next generation of individuals by recombining the chromosomes of the two parent individuals. After this, the stage of mutation begins - a random change in one (or several) positions in the individual's chromosome.

The above sequence of actions is carried out until the stop criterion is met. According to the result of GA work, the best individual from the last population is found.

In the considered problem, GA is proposed to be used for dictionary optimization. The dictionary is encoded with a binary string (the length of the string corresponds to the number of words in the complete dictionary of the corpus), in which 1 means the inclusion of this feature (word) in the dictionary, 0 means its exclusion from the dictionary. At the first iteration, the line is completely filled with ones. A string containing zeros is a truncated dictionary representation.

Conducting clustering based on the obtained vocabulary and calculating the silhouette criterion, we obtain the suitability of this vocabulary.

3. Solving Test Problems of Text Messages Clustering

Three tasks were chosen to test the proposed approach. The data for the first two tasks were taken from the UCI Machine Learning repository; the data for the third task was collected using special tools for

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

administering groups of the social network «Vkontakte». The clustering method used in all three tasks is k-means.

3.1. Task 1: BBC health tweets dataset

The Figure 4 below shows the raw data - a set of short headlines for news articles from the BBC (British Broadcasting Corporation) Twitter, tagged with "health" topic. Data set contains 3898 headlines (documents) of various lengths.

```
Breast cancer risk test devised
GP workload harming care - BMA poll
Short peoples heart risk greater
New approach against HIV promising
Coalition undermined NHS - doctors
Review of case against NHS manager
VIDEO: All day is empty, what am I going to do?
VIDEO: Overhaul needed for end-of-life care
Care for dying needs overhaul
VIDEO: NHS: Labour and Tory key policies
Have GP services got worse?
"A&amp;amp;E" waiting hits new worst level
Parties row over GP opening hours
Why strenuous runs may not be so bad after all
VIDEO: Health surcharge for non-EU patients
VIDEO: Skin cancer spike from 60s holidays
80 might die in future outbreak
Skin cancer linked to holiday boom
Public back tax rises to fund NHS
VIDEO: Welcome to the designer asylum
VIDEO: Why are we having less sex?
Five ideas to transform the NHS
```

Figure 4. BBC tweets dataset

The table 1 below shows examples of tokenized documents from the corpus:

Table 1. Examples of data tokenization in BBC tweets dataset

Original documents from the corpus	Tokenized documents
Ambulance progress not fast enough	ambulance progress not fast enough
Guinea declares Ebola emergency	guinea declare ebola emergency
VIDEO: Sitting down poses health risk	video sit down pose health risk

A multilevel clustering approach was proposed: clustering is performed once, then the most numerous cluster acts as a new set and is subjected to clustering again.

For this dataset, it was decided to perform two-level clustering. The Figure 5 below shows the resulting clusters, keywords that can be used to describe the documents of each cluster and also the size of each cluster.

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

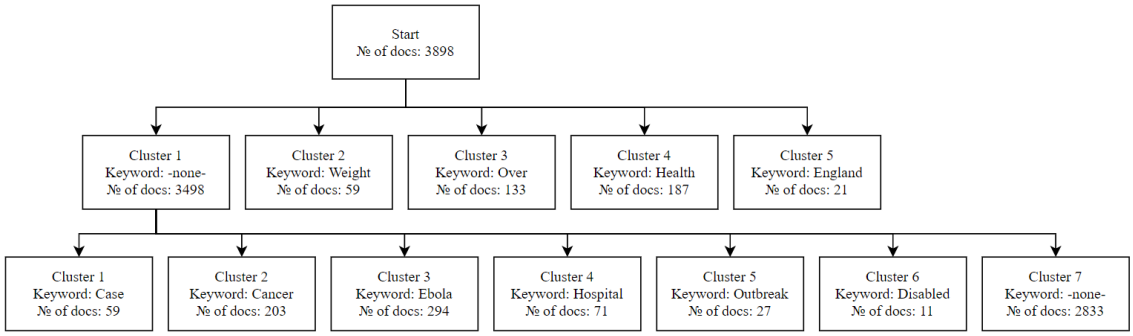

Figure 5. BBC tweets clustering tree

As can be seen from the figure 5, all found clusters turned out to be informative, since the semantics of cluster messages is clear by the keyword. In addition, there are no repetitions of topics, they are clearly separated.

The resulting clusters are also fairly universal semantic categories, since there are no clusters with a relatively small population.

3.2. Task 2: News headlines from Kaiser Health News

The Figure 6 below shows the raw data - a set of news articles headlines from the Kaiser Health News. Dataset contains 3433headlines (documents) of various lengths.

```
Tougher Vaccine Law In Calif. Clears First Hurdle
A new sort of extracurricular activity seeing patients.
Houston firefighters have another tool at their disposal when answering calls digital doctors
Will Montana expand its Medicaid program Some moderates there have revived the idea
Rand Pauls campaign first day abortion and the budget
Enroll America, A Group Dedicated To Promoting Health Law Sign Ups, To Cut 100 Jobs
RT LVGillespie Ever look up health symptoms on a computer Retired nlm_news director led the charge
PatchUpJob Obamacare
Still getting your taxes done Your 2014 tax bill could be affected by your insurance.
In medicine and use a computer You probably have this man to thank ... LVGillespie reports
RT cnnhealth Two doctors fight for their own choice of how to die. Via KHNews
RT drshow Evidence just doesnt support annual physicals, JennyAGold reports But that info isnt getting out
Still getting your taxes done Your 2014 tax bill could be affected by your insurance. Watch
Rule Proposed On Providing Mental Health 'Parity' In Medicaid Program
Battle For Mental Health Parity Produces Mixed Results
HighDeductible Health Plans Can Ruin Finances
RT JennyAGold Maybe You Should Skip That Annual Physical nprnews KHNews
Billionaires Harness Money, Technology In Pursuit Of Fountain Of Youth
Ritual, Not Science, Keeps The Annual Physical Alive
Medicare Is Stingy In First Year Of Doctor Bonuses
Cancer Survivor Worries About Supreme Court Ruling On Obamacare Subsidies
Got a minute Watch how the your 2014 tax bill could be affected by your health insurance
Consumers Contributing Less To Health Savings Accounts, Study Finds
RT Julie_appleby Its tax time. Heres a clip on what you need to know if you received a health insurance sub
Got a min Your 2014 tax bill could be affected by your health insurance.
```

Figure 6. KHN dataset

The table 2 below shows examples of tokenized documents from the corpus:

Table 2. Examples of data tokenization in KHN dataset

Original documents from the corpus	Tokenized documents
Todays headlines Obama Administration Simplifies Application For Health Insurance	todays headlineobama administration simplify application health insurance

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

Medicare falls behind with project to allow patients to get hospice and treatments to cure them at the same time	medicare fall behind with project allow patient get hospice and treatment cure same time

The clustering results are shown in the Figure 7 below:

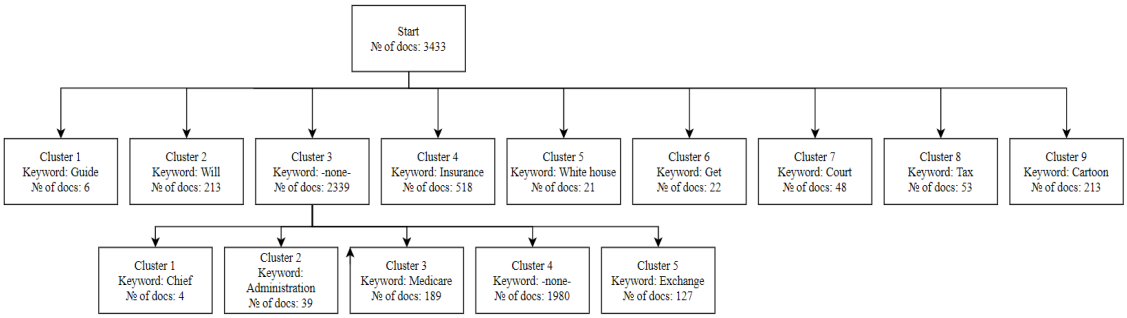

Figure 7. KHN clustering tree

A wider semantic spectrum occurred during this dataset clustering. Not all clusters are concentrated on the topic of “health”. It can be seen that some clusters have formed around "insurance" and "taxes" that are not exclusively medical-related topics. This can be explained, among other things, by the broader focus of the publication.

However, clusters with obvious health topics (cluster 3 in the first level and cluster 2 in the second level) combine more documents.

General conclusion on the two English language sets: clusters are well formed relative to nouns, they contain the main semantic load of the message. Verbs play a supporting role.

3.3. Task 3: Reshetnev University applicants’ messages

Unlike the two previous tasks, this dataset is formed in Russian language and does not consist of news headlines, but of short messages from applicants, which contain any question regarding the selection committee, admission, etc.

Since the data is real and not standardized (unlike news headlines, for which there are generally accepted rules), it was decided to increase the number of clustering levels to 6. The final clustering tree looks as follows:

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

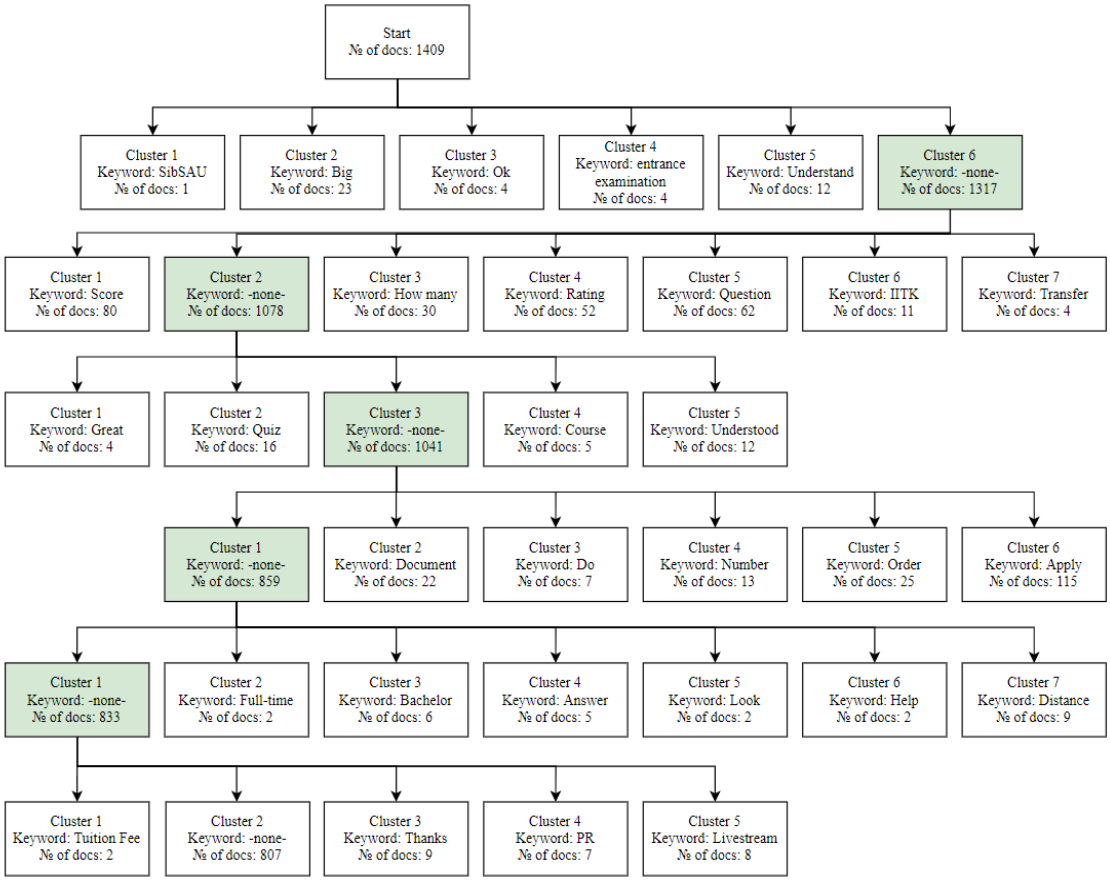

Figure 8. Reshetnev University clustering tree

As you can see from the Figure 8, some clusters turned out to be less informative compared to the English-language datasets. One of the reasons for this may be that in Russian, verbs often carry a semantic load, compare with nouns, but outside the context they can be interpreted in different ways, therefore, highlighting a verb as a cluster key word does not give a complete picture of the meaning of messages in this cluster.

The general theme can be traced closely to the theme of university and admission. In addition, an increase in the levels of clustering made it possible to get to unobvious clusters.

4. Conclusion

In this paper, the problem of clustering text messages was solved. Two methods of text vectorization were tested: bag of words and tf-idf. The task of clustering text data was solved with the help of k-means method. The dictionary was optimized with the help of genetic algorithm. In the future work further research of the purposed optimization method is planned.

References

Andreev, M. (2011). *Genetic Algorithm*: *In simple way about the complex*. https://habr.com/ru/post/128704

https://doi.org/10.15405/epct.23021.3
Corresponding Author: Anastasiya Polyakova
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

Heidenreich, H. (2018). *Natural Language Processing: Count Vectorization with scikit-learn* https://towardsdatascience.com/natural-language-processing-count-vectorization-with-scikit-learn-e7804269bb5e

Müller, A. C., & Guido, S. (2017). *Introduction to machine learning with Python: a guide for data scientists.* O'Reilly Media, Inc.

Nikitinsky, N. (2015). *A little about word2vec: a useful theory*. http://nlpx.net/archives/179

Sarwan, N. S. (2017). *An Intuitive Understanding of Word Embeddings: From Count Vectors to Word2Vec* https://medium.com/analytics-vidhya/an-intuitive-understanding-of-word-embeddings-from-count-vectors-to-word2vec-8231e18dbe92

Selivanov, D. (2020). *Analyzing Texts with the text2vec package*. https://cran.rproject.org/web/packages/text2vec/vignettes/text-vectorization

Yordanov, V. (2019). *Fundamentals of Natural Language Processing for Text*. https://habr.com/ru/company/Voximplant/blog/446738/