Building a graph of a sequence of text units to create a sentence generation system

Maksim Kaminskiy
Samara National Research University
Samara, Russia
beefiestracer@gmail.com

Igor Rytsarev
Samara National Research University;
Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS
Samara, Russia
rycarev@gmail.com

Alexander Kupriyanov
Samara National Research University;
Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS
Samara, Russia
alexkupr@gmail.com

Maximilian Khotilin
Samara National Research University
Samara, Russia
turbomax.1994@gmail.com

***Abstract*—The article is devoted to the development of a text data analysis system. The approaches to the presentation of text from the posts of a single page in the form of a dictionary of phrases for sentence generation and applying a developed system for correcting the results of neural network generation are considered. Within the framework of the work, data collection, filtering and processing using Big Data technologies were implemented.**

Keywords—annotation, social networks, big data, graph, machine learning

I. Introduction

The 'social network' notion was used by sociologists back in the 1920s for investigating the interrelations between participants of different communities. The psychologist Iacob Moreno offered sociograms representing graphs on which separate individuals were represented by points, and interrelations between them – by lines. The idea of using the apparatus of the theory of graphs for studying interrelations between people was taken by specialists in such areas as sociology, psychology, anthropology, politology, economics – thus, the Social Network Analysis flow was established, dealing with studying structural properties of social interrelations modeled in the form of graphs and networks. Building the model based on various data from printed media, additional inquiries and questioning was an important but rather time-consuming stage of such investigation [1].

Contemporary social networks substantially have made the life of researchers easier, having presented to them the developing and easily-accessible source of big data. Every day the users of social networks generate large volumes of data of different type. The analysis results of this information may become a perfect material for investigations of various fields [2]. For example, Social Media Marketing (SMM), is an important tool for promoting in the Internet for many companies. Social networks are an environment in which all users unconsciously work as focus groups, and do not hesitate to share their opinions, argue, prove their case, express their needs and wishes. Companies are constantly looking for client insights that people share on social networks [3]. One of the tools for these studies is content analysis - a text analysis method that is carried out by counting the occurrence of components in the analyzed information, used in sociology, as well as in computer technology. The purpose of this method is to identify or measure various facts and trends reflected in the investigated documents. Using content analysis, it is possible to establish both the characteristics of information sources and the characteristics of the communication process. Content analysis can be used to study most of the documentary sources, but it works best with a relatively large amount of single-order data [4]. Hence, it is so vital to be able to represent these data in the form convenient for the efficient analysis.

From a commercial point of view, the most successful Natural-language generation (NLG) applications have been data-to-text conversion systems that generate text summaries of databases and datasets. These systems usually perform data analysis as well as text generation. Research has shown that text-based resumes can be more effective than graphics and other visual elements for decision support, and that computer-generated texts can outperform (from the reader's point of view) human-written texts. There is currently considerable commercial interest in using NLG to aggregate financial and business data. Gartner has said that NLG will become the standard tool for 90% of modern BI (Business intelligence) and analytics platforms. NLG is also used for commercial purposes for automated journalism, chatbots, creating product descriptions for e-Commerce sites, and compiling brief medical records [5].

The text annotating methods can be broken down into two groups: extracting and generating. Among the extracting methods of automatic annotating the method on the basis of the theory of graphs can be distinguished, where the text is presented as a graph, which nodes are text fragments, and edges are relations among them [6].

II. Task settings

The modern world is dynamic, computerized, the employee is required to complete a task fast and qualitatively to the greatest possible extent. The software that uses the developed algorithm in its work can be used by employees with the occupations, where it is necessary to print the text for drawing up similar-in-content documents decreasing the time spent for such task, or in organizations servicing the citizens with disabilities (static and dynamic disorders of upper limbs, visual impairments) on the quoted places, duties of which are directly related to the work with computers. Also, the software tool can be used in the field of education, providing students with the opportunity to save time on reporting on completed work. With its help, it is possible to

facilitate blogging on social networks for professional, entertaining or educational purposes, as the algorithm will learn the style of the written texts and begin to suggest the most suitable words for input.

III. Collection and working with data

The algorithm developed in the framework of the research, first of all, collects data, then filtering it in order to obtain the crucial text information, then building a graph of key words, when passing on the chains of words are built. Further on, if required, the system can be additionally improved adding new texts belonging to other authors, for style combining [7].

One of the most known weblog platforms LiveJournal was chosen as a source of data, which represents the possibility of publishing own records and commenting on others. This large resource abounds with weblogs on various topics, being an excellent source of large volumes of text information. All obtained information is stored in the text file to work with after that.

The data must be prepared for further work with the text. Hyperlinks, emojis, punctuation marks, special characters are filtered out, all other letters are converted to lowercase. The words with the length less than four characters are filtered out as well in order to exclude the majority of auxiliary parts of speech. After that the text is structured into separate key words. Lemmatization of tokens, i.e. reducing the words into their initial form, is performed after that. Under lemmatization the parts of speech are transformed according to the following type: nouns – singular, nominative case; adjectives – singular, masculine, nominative case; verb – indefinite form (infinitive). Example of lemmatization can be seen on Figure 1.

посмотрела	посмотреть
кошки	кошка
субтитрами	субтитр
знаю	знать
смотрится	смотреться
русском	русский
языке	язык
глазами →	глаз
детей	ребёнок
может	мочь
надо	надо
ради	ради
коллекции	коллекция
впечатлений	впечатление
жалею	жалеть
впечатления	впечатление

Fig. 1. Example of transformation of words into lemma.

The vocabulary is created after these transformations, arranged by the frequency of application of key words, based on which the phrase matrix is built, the terms of which there will be a number of repetitions among the words in the text. Further on, having the phrase matrix and the vocabulary $W = w_1 w_2 w_3 \dots w_n$, a graph can be built. The nodes of the graph will be key words w_i from the vocabulary W, the edges connect them into phrases from the text. The number of repetitions among the words is given as the edge weight.

When improving a new portion of processed information is introduced to the graph. However, since the weight of a new bond primarily will be less than that of the bonds already existed in the graph, for compensation purposes a new structure at every node is introduced, which is represented as a stack of words ($K = k_1 k_2 k_3 \dots k_m$, where k_j is a word taken separately from the stack). It has the latest bonds created after the node. The priority for output will be given to new data, and the compensation of the low weight of the bond will be performed by means of introducing a coefficient s, which depends on the position of the word in the stack, selecting which logic chains for two sets of data at once can be built, but the second one will have some priority, because it was used for improving the system. The summarized scheme of work of the described algorithm is given on Figure 2.

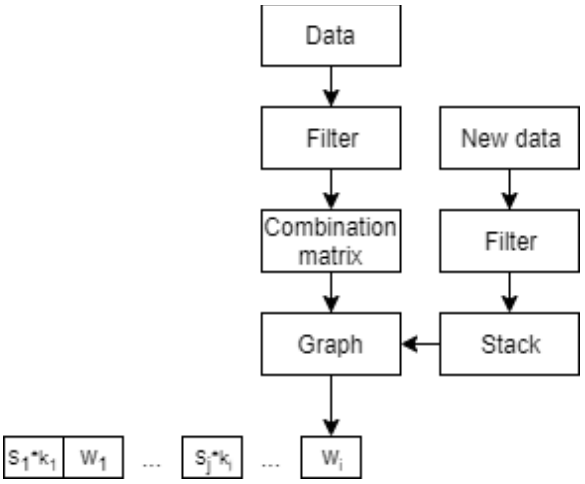

Fig. 2. Schematic representation of work of algorithm.

IV. Comparing of styles of different authors

Two posts dedicated to the "Cats" screen musical by two different authors were taken for the research. Having obtained the text, the data were filtered and processed for making the vocabulary of key words and matrix of phrases. Two weighted graphs were built after that, and they are presented on Figure 3 and 4.

The first graph has 297 nodes and 385 edges, the second one has 296 nodes and 384 edges. Having compared them totally 49 nodes having the same name were found. For these 49 coincidences the big difference between neighboring nodes can be observed, making a conclusion that the frequency of coincided words with the authors differs.

Further on, we consider the total capacity of every node from the graph. As seen from Figure 5 and 6, that the first author frequently uses certain words (for example, the variation of the word «быть» ("be")), while the usage of words by the other author is more even.

In the result of these comparisons it can be concluded that the lexicon of the authors substantially differs regardless the writing of articles on the topic alike.

The developed algorithm was also applied to texts generated by GRU and LSTM neural networks to eliminate word errors and increase contextual connectivity. As a dataset for training neural networks of text generation, a text consisting of speeches of the characters of Shakespeare's plays was taken. To check the generated texts, the GLTR

(Giant Language model Test Room) was selected, which is a tool for detecting text that was automatically generated. This instrument can use any text data and analyze what language model GPT-2 would predict in each position. Each text is analyzed according to how likely each word will be a predicted word, taking into account the context on the left. If the actual word used would be in the top 10 predicted words, the background is colored in green, for the top 100 in yellow, the top 1000 in red, otherwise in purple. On Figures 7 and 8, you can see the results of the analysis of texts, and Figures 9 and 10 show histograms where the number of predictions for each of the texts is calculated [7].

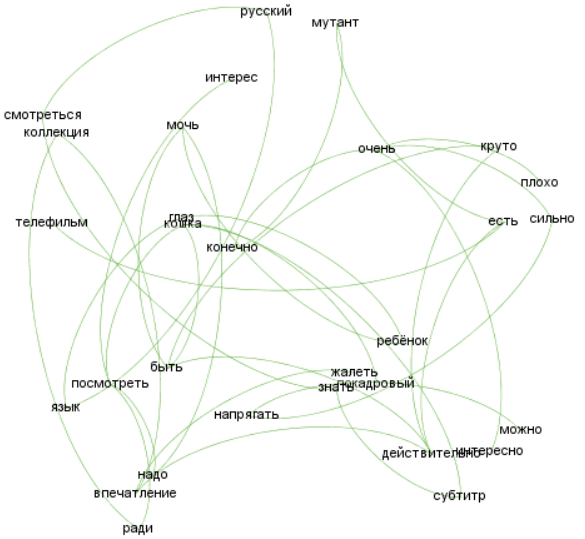

Fig. 3. Simplified representation of the graph drawn by the "Musical "Cats" post in cinema: mutants not able to sing", created by the user named shakko_kitsune.

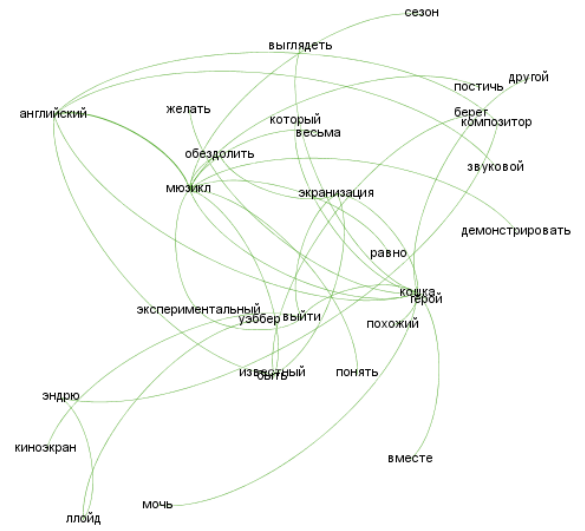

Fig. 4. Simplified representation of the graph drawn by the "Cats: purring musical", created by the user named carabas.

V. Correction of Generated Texts

As can be seen from the result of the generation, the text generated by GRU turned out to be not too contextually connected and 7 words were displayed with errors, in the LSTM generated text there are slightly more errors in words – 9, but according to the data from the histograms, it surpassed the previous neural network.

To correct the received texts, an algorithm was developed that, in conjunction with the constructed graph, allows to correct errors in words and increase contextual connectivity.

быть	27
который	14
фильм	12
танцевать	11
вообще	10
реально	8
только	8
человек	8
большой	6
даже	6
кошка	6
очень	6
роль	6
сейчас	6
экранизация	6
посмотреть	5

Fig. 5. Capacity of nodes of first graph.

который	18
мюзикл	18
кошка	14
актёр	10
выглядеть	10
поэтому	10
фильм	10
английский	7
выступление	6
есть	6
однако	6
понять	6
потому	6
похожий	5
автор	4
быть	4

Fig. 6. Capacity of nodes of second graph.

ROMEO:
Nor needs again, awake, sir; yet bring Edward's power
In every party of his uncles did.
DUKE OF AUMERLE:
Where is devote, and I am satisfied;
With our foes ere it gized.
GLOUCESTER:
KING EDWARD IV:
Now, princely Buckingham, awake! Gentle God, I pray thee, upon;
WASWARD:
Cloffic! have prepared conduction there,
Or not my father: shall you go with us,
'Twas fourteen purpose to assail in every foe.
RICHARD:
Then leans that name,
No bare adother, if he did forswear it;
Inferreth argues and mortal feet,
Nor incated happily he gave me
For gifth; and bring it talk, oracle,
And leave him not before him.
CAMILLO:
Say you have, by right out of these traitor than
the way it to be saying, aumpose lies
make his demand.
MISTRESS OVERDONE:
What's this?
FRIAR LOUCESTER:
Sweet saw on me, Gregurater.
Now, no more: my mother, young York,
And this so far fortune, visited
A king of heaven: the neare have learn'd these arts:
Now shall you give unto them told and show me how
That she shall poink

Fig. 7. GRU text generation result.

ROMEO: the happy the news;
And she's done no day: hast thou be seen in Henry's glory;
And the manifested with this born.
Clown:
Ay, are too sore to know that ladies,--whither do thou but this most islands,
Here with thy sorrow's torture; but that you ave
Many lord of this action.
Stand gates:
Why, ure still Cousin of Pisa harm.
FRIAR LAURENCE:
Hold, farewell: this Tarquion as the sun,
Some way their love, with such direction to his ears Show in the bod of some death
And threaten us from the jest of it.
DUCHESS OF YORK:
Ay, to speak business.
HORTENSIO:
Tailor:
You bid me seek revengeant as
you cannot:
You lie, a none will not King Richard and myself
For most impedition be set down
expear your pear-fair inform and nature
show our tender heart proportion nor gow:
Sirrah young grain I weigh, no dancing,
But stops thy son: who wilt lose his eyes?
JULIET:
Must be that sorry thine abundance;
For never was not so resolved which wash'd
The importand of the mind being naked.
ELBOW:
Come, what m

Fig. 8. LSTM text generation result.

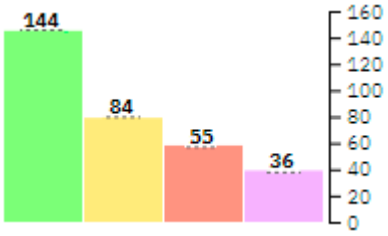

Fig. 9. Prediction histogram for GRU-generated text.

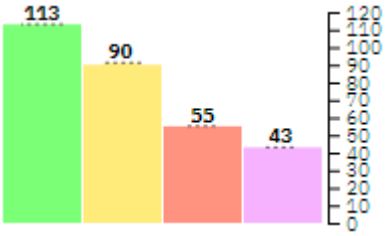

Fig. 10. Prediction histogram for LSTM-generated text.

To build a graph, the text on which the neural networks were trained was used (Figure 11).

Then the triples of words h_i, h_{i+1} и h_{i+2}. are examined. Since we are working with a large text data set, the "windows" of the three words considered during the operation of the algorithm are enough to correct the text. Each word h_i in the sentence is checked for its presence in the column, and then the words associated with it are considered, if there is no word in the column, then we shift our "window" by 1 step. Then check for the presence of related words h_{i+1}. If h_{i+1} is in the list, then we shift the "window" by 1 step and continue checking. If not, then look at the word h_{i+2}. We check the presence of this word in the graph; if it is absent, we shift the "window" by 3 steps. If the word is found, then we check through which words to which the links from h_i depart, it is possible to establish a connection with h_{i+2}. If there are connecting words, put h_{i+1}in place of one of them, if not, then shift the "window" by 2 steps. The choice of words that can be put in place of h_{i+1} is carried out by calculating the probability by the formula:

$$p_i = \frac{g_i}{\sum_{j=1}^{N} g_j}, \quad (1)$$

where g_i is the weight of the edge between h_{i+1} and the word connecting it to h_ii, p_i is the probability of choice and N is the number of connected words. A generalized scheme of the described algorithm is presented on Figure 12.

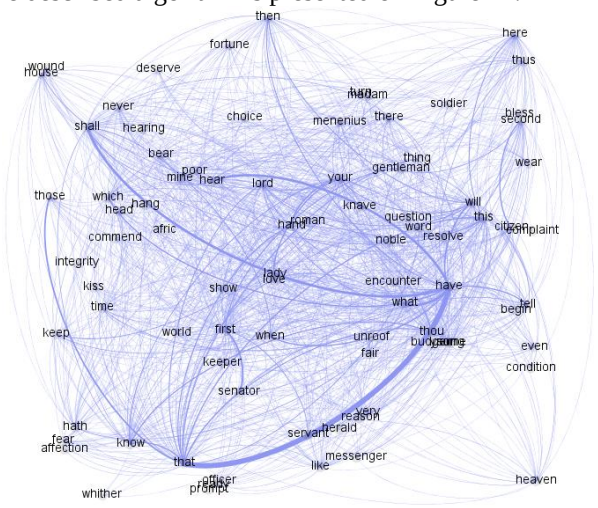

Fig. 11. Simplified image of a graph composed from speeches of characters from Shakespeare's plays.

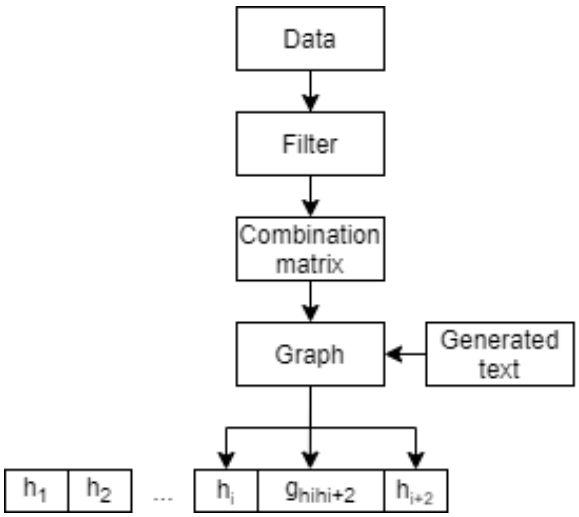

Fig. 12. Schematic representation of work of algorithm.

After processing the text generated by GRU and LSTM, almost all incorrectly composed words were replaced and the contextual connection between words in sentences improved slightly. Figures 13, 14, 15 and 16 show the results after adjustment.

As a result of applying the algorithm, 6 out of 7 incorrectly composed words were eliminated in the text generated by GRU, and 7 out of 9 in the text from LSTM, GLTR analysis results were also improved.

VI. Conclusion

We presented the learning system of annotating on the basis of the theory of graphs allowing to build the chains of words similar in style to the texts of authors, which if required can be additionally improved downloading the texts of another authorship or another topic. At further development of the algorithm of work of the elaborated system we consider its usage in typing sizable texts to be expedient, which allows to increase the rate of their writing.

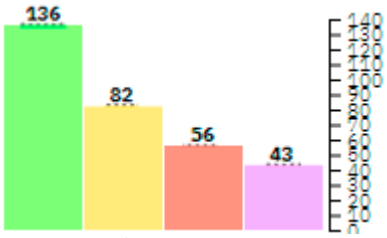

Fig. 13. Prediction histogram for corrected text generated by GRU.

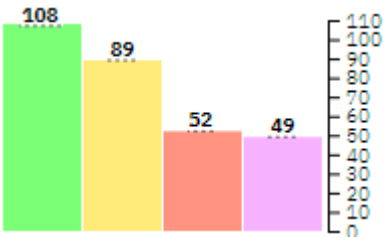

Fig. 14. Prediction histogram for corrected text generated by LSTM.

ROMEO:
Nor needs must, awake, sir: yet bring Edward's power
In every party of his uncles here
DUKE OF AUMERLE:
Where is devote, and I am satisfied;
With that foes ere it gized.
GLOUCESTER:
KING EDWARD IV:
Now, princely Buckingham, good! Gentle God, I pray thee, upon;
WASWARD:
Cloffic! have prepared that there,
Or not my father: shall you go with us,
was that purpose to assail in every foe.
RICHARD:
Then wills that name,
No bare adother, if he did forswear it;
Inferreth argues and mortal feet,
Nor incated happily he gave me
For gifth; and bring it talk, oracle,
that leave this not before him.
CAMILLO:
Say you have, by right out of these words than
the way it to be saying, marcius lie
make good demand.
MISTRESS OVERDONE:
What's this?
FRIAR LOUCESTER:
Sweet saw on me, Greguraters.
Now, no more: my mother, have York,
what this so far fortune, visited
A king of heaven: the neare have learn'd these arts:
Now shall have give leave them told king show me how
That have shall poink

Fig. 15. The result of applying the corrective algorithm on the text generated by GRU.

An algorithm was also developed that, when used in conjunction with a graph of keywords, can be applied to texts generated by neural networks to correct incorrectly generated words and increase contextual connectivity. In addition to this, this algorithm improved the results of the analysis of the GLTR text.

This application has an extensive scope: for example, in the current situation, given the unfavorable epidemic situation caused by coronavirus infection, the majority of the population have a need to master the work remotely. At the same time, the specificity of each sphere implies a certain terminology, a set of the most used speech turnovers when creating a product description, correspondence with customers and partners. This program will simplify and accelerate the work on a set of textual material, which in turn will increase the productivity of the labor process, help to more effectively deal with deadlines (which, by the way, have already entered the norm of modern life).

ROMEO: soul happy that news;
And she's done no day: hast thou be seen in Henry's glory;
And the manifested effect this heart.
Clown:
Ay, are too sore to know with ladies, whither do thou angelo more islands,
Here with give sorrow's torture; but that you ave
Many lord of this valiant.
Stand gates:
Why, ure still Cousin of Pisa here.
FRIAR LAURENCE:
madam farewell: this Tarquion as the sun,
Some slain their lords, with will direction to his ears Show in the bod of some death
that threaten us from time jest of it.
DUCHESS OF YORK:
Ay, to speak night.
HORTENSIO:
Tailor:
You bid me seek revengeant as
you would
You lie, a none will not King Richard cannot myself
false most impedition be set down
send your shall-fair inform that nature
show theirtender heart proportion nor gow:
Sirrah young grain I weigh, no dancing,
But stops thy son: who wilt lose mine eyes?
JULIET:
Must be that theey thine abundance;
For never was not so resolved that there
The importand of the mind leave naked.
ELBOW:
Come, what m

Fig. 16. The result of applying the corrective algorithm on the text generated by LSTM.

Also, this program can provide assistance in the preparation of advertising articles, political campaign materials. Allowing you to analyze large textual volumes (for example, articles on the Internet or in print media) to determine the intentions, psychological state of target groups, identify attitudes, interests and values, belief systems by highlighting the most commonly used expressions and turns. Subsequently, relying on these stable constructions, using them in composing his own texts, the author acts between the lines on the readers unconscious mind, letting him know that they speak the same language, their problems and ideals are the same, thereby increasing the level of openness for the information presented and trust in it.

But there is also a category of people who find it difficult to type texts on a computer keyboard due to limited health capabilities. For example, a person with spastic disorders in the upper extremities who works on a PC. Each movement is much more difficult for him, with greater efforts than a conditionally healthy one, and besides, his exhaustion and fatigue comes much faster. And here, the use of the application will act as a significant assistant, allowing you to minimize arbitrary movements, therefore, the energy expended.

Thus, the first algorithm presented:

- Simplifies typing, as it learns the style of the author's letters and suggests the most appropriate words for subsequent input;
- Saves time, because instead of manual typing, you can use the options of the displayed words provided by the algorithm, which partially automates the process of working with text;
- Increases productivity, reducing time costs, making it possible to do more work in the same amount of time.

As a result, the totality of these advantages allows you to increase productivity.

And the second one algorithm:

- Fixes errors in incorrectly generated words by replacing them;
- Increases the contextual coherence of the text by replacing words with those that have associations within the graph, and which are met in the original text, written by a human.

This transformation brings the text closer to the style of the author, allowing the text to look less similar to the one that was compiled by a machine.

ACKNOWLEDGMENT

The work is done with the financial support from the Russian Foundation for Basic Research (No. 18-37-00418, No. 19-29-01135, No. 19-31-90160) and the Ministry of Science and Higher Education of the Russian Federation (grant # 0777-2020-0017) in the framework of fulfilling the governmental task of Samara University and FSRC "Crystallography and Photonics" of RAS.

REFERENCES

[1] W. Tan, B. Blake and L. Saleh, “Social-Network-Sourced Big Data Analytics,” Open systems. DBMS, no. 8, pp. 37-41, 2013.

[2] I.A Rytsarev, D.V. Kirish and A.V. Kupriyanov, “Clustering of media content from social networks using bigdata technology,” Computer Optics, vol. 42, no. 5, pp. 921-927, 2018. DOI: 10.18287/2412-6179-2018-42-5- 921-927.

[3] “Social network analytics: 10 ways to use monitoring systems,” YouScan - Social Media Monitoring System, 2019 [Online]. URL: https://youscan.io/ru/blog/10-instrumentov-analiza-socsetei/.

[4] I.V. Dmitriev, “Content analysis: essence, tasks, procedures,” PSI-FACTOR. – Center for Practical Psycology, 2005 [Online]. URL: https://psyfactor.org/lib/k-a.htm.

[5] “Natural-language generation,” Wikipedia [Online]. URL: https://en.wikipedia.org/wiki/Natural-language_generation.

[6] P.G. Osminin, “Modern approaches to automatic summarization,” Bulletin of South Ural State University. Series: Linguistics, no. 25, pp. 134-135, 2012.

[7] I.A. Rytsarev, A.V. Blagov and M.I. Khotilin, “Development and implementation of services to collect social networking data in order to improve the human environment,” Collected papers of ITNT. Information technologies and nanotechnologies, pp. 2452-2457, 2018.

[8] H. Strobelt and S. Gehrmann, “Catching a Unicorn with GLTR: A tool to detect automatically generated text,” Catching Unicorns with GLTR, 2019.