Topic Organization of E-hypertext Media: Corpus Driven Research *

Valerii Shulginov [1]
prostovalera@yandex.ru

Vadim Shulginov [2]
vadim.shulginov@yandex.ru

Olga Mitrofanova [3]
o.mitrofanova@spbu.ru

[1]Higher School of Economics, Moscow
[2]Rostelecom, Vladivostok,
[3]Saint Petersburg State University, Saint Petersburg,
Russian Federation

Abstract

This article focuses on the principles of topic analysis of electronic hypertext. E-hypertext is defined as a communicative-cognitive phenomenon, which has all the signs of textuality, and is also characterized by a complex structure and non-linear connection between text fragments. We are developing an algorithm that reveals the thematic connections in the three-part elements of e-hypertext. As a result, thematic dominants in the structure of media discourse are identified, as well as two strategies of topic organization of e-hypertext: monothematic and polythematic transitions.

Keywords: *e-hypertext, topic modeling, text coherence, semantic proximity, hypertextuality*

1 Introduction

Over the past decades, hypertext has become a complex phenomenon that has been interpreted in various sciences (cybernetics, sociology, linguistics, psychology, etc.), and this suggests a metaphorization of the concept. It made it possible to define hypertext as a method of inquiry, "a lens through which other topics are investigated" [Atzenbeck and Nürnberg, 2019, 29]. Thus, hypertext has become not just an object of study, but a special method of studying various theoretical and practical issues. Topic modeling of the structure of electronic hypertext (e-hypertext) is an important task in Natural Language Processing and Computational Linguistics, as it allows to provide prediction of semantic relations between linear texts, which, in the end, is closely connected with the task of automatic text generation, development of dialogue programs and creation of artificial intelligence.

In this paper we regard the phenomenon of hypertext as a complex structure that includes two levels of organization. On the first level, it is type of system, characterized by provision of links or other structure to users. This type of structure has formed the basis of the Internet, and it is characterized by nonlinearity, multimedia, openness, eternity and provides storage and exchange of information between users. All data on the Internet forms a universe of electronic documents [Ryazantseva, 2010] consisting of nodes and multidirectional links between them.

On the second level, electronic hypertext is a special type of text that is transformed under the influence of hypertext media into a new type of communicative and cognitive element, which has all the features of textuality (cohesion, coherence, intention, acceptability, informativeness, situationality, intertextuality), and is characterized by a complex structure and non-linear links between fragments [Shulginov 2016]. This approach is based on the hypertext conception suggested by the French literary theorist G. Genette [Genette, 1982]. In Genette's theory the term "hypertextuality" is used to refer to the type of relationships between fiction texts only, where one or more texts derive from the initial text by means of direct transformation or imitation. Thus, hypertext as a special information system creates a non-linear infinite space, which determines the formation of a new type of discourse based on dialogical connections between text chunks.

Author of the term hypertext T. Nelson argues that electronic hypertext changes the positions of the writer and reader. His notion of hypertext is understood by "non-sequential writing — text that branches and allows choices to the reader, best read at an interactive screen... this is a series of text chunks connected by links, which offer a reader different pathways" [Nelson, 1993, 2]. It manifests the poststructuralist concept of "death of the author" [Bart, 1994], according to which the source of the text is not in writing, but in reading. The reader ceases to be a passive recipient of information, he or she constructs the message in cooperation with the author.

However, such freedom of perception and interpretation of e-hypertext by the reader turns out to be quite conditional, since it is the author who defines its composition. Actually, the e-text readers freedom is limited by the choice of one of two strategies of hypertext activation: "selecting the text semantically related to the previously read section (coherence strategy) or choosing the most interesting text, delaying reading of less interesting sections (interest strategy)" [Salmeron, Kintsch, Canas, 2006, 1157]. But the author creates a three-component structure that includes the initial text, the target text and the hyperlink. Moreover, if we study electronic hypertext from the author's point of view, it is the target text that becomes the primary one, as it is the stimulus for further construction of hypertext. The main means of ensuring the cohesion and coherence of electronic hypertext is the semantics of the hypertext transition, which is usually marked by the nomination of the links. Depending on the authors strategy, the semantic proximity between the nomination of the hyperlink and the target text could be variable: the target text performs the function of interpretation (in this case, there are relations of title/text) or hides the semantics of the hypertext transition (the hyperlink indicates the possibility of hypertext transition, but doesn't give a semantic characteristic of the target text). In addition, it is important to take into account which texts can form hypertext structures with each other.

2 Methods

To analyze the topic structure of electronic hypertext, we have created a corpus of three-component hypertext elements, the search engine allows to find links according to the parameters: part-of-speech characteristics; the number of words in the link nomination; syntactic position of the link in the sentence. In addition, the corpus has the ability to search by keywords in the source and target texts and to identify which links connect these text fragments. The corpus includes texts that relate to the Internet news discourse. The following media were used as sources of information:"Kommersant", "Izvestia", "RBC", "Novaya Gazeta", "TASS", "Dozhd, "Vedomosti", "Interfax", "HabraHabr", "Kremlin". The data set includes 53000 articles which include 12 million tokens in total. These articles are connected in 70000 three-component hypertext elements. These data are processed in three stages: data collection, preprocessing and modeling, and topic modeling. Data collection is carried out with the use of parser developed on the basis of Python libraries: requests (access to web pages), BeautifulSoup (reading of HTML-content), re and NLTK (selection of necessary elements). Parsing is performed in two stages.

Firstly, the parser analyzes Internet news resources, indexes all links on the page, collects pairs: source/target text, links, and domains. If a link is found on the target text, it is assigned the status of the source text and the algorithm is repeated. Secondly, the parser analyzes the pairs of texts to find the full-text fragments. When the necessary tags are found, all elements of source and target text are loaded into the database. The set of analyzed components includes media title, article content, title, subtitle, name of author, tags, date of publication. So the database allows us to identify correlations between linguistic (text subject, frequency characteristics) and extra-linguistic features.

Each text goes through a preprocessing stage, during which tokenization, normalization, lowercase transformation and removal of punctuation marks are made. We used standard dictionaries to remove stop words, but the word "год" (year) and acronyms (млн. (million), млрд (billion)) were also removed due to the peculiarities of web discourse. After data preprocessing, we generalized keywords to collocations taking into account frequency of their co-occurrence. Collocation analysis was performed by means of phrases module in gensim library for Python (`https://radimrehurek.com/gensim/`). It has allowed combining semantic neighbors in one token that has reduced the matrix dimension. We used a TF-IDF scheme to detect keywords in each document. The weight of each keyword was calculated using the wellknown TF-IDF (Term Frequency — Inverse Document Frequency) formula.

As a result, we received a matrix of 2000 tokens per 53,000 documents (base-matrix), which takes into account the weight of each token for the text fragment in which it is used.

At the stage of topic modeling, we used multiple learning method t-Distributed Stochastic Neighbor Embedding (`https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE`), which is particularly well suited for the visualization of high-dimensional datasets. This method transforms high-dimensional objects by a two-dimensional points in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability. The algorithms starts by calculating the probability of similarity of points in high-dimensional space and calculating the probability of similarity of points in the corresponding low-dimensional space. Then it tries to minimize the Kullback-Leibler divergence between the two distributions using a gradient descent method with respect to the locations of the points in the low-dimensional space. Using this method, we were able to predict the topic clustering of the hypertext corpus. After the algorithm

had worked, we got a set of points characterizing the distribution of our data set. Then we clustered it with the DBSCAN (`https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html`) algorithm, one of the advantages of which is that it does not require specifying the number of clusters in advance. The DBSCAN algorithm can be abstracted into the following steps:

- find the points in the (eps) neighborhood of every point, and identify the core points with more than minPts neighbors;
- find the connected components of core points on the neighbor graph, ignoring all non-core points;
- assign each non-core point to a nearby cluster if the cluster is an (eps) neighbor, otherwise assign it to noise.

As a result of DBSCAN's work, the clusters depicted in the Figure1 are highlighted.

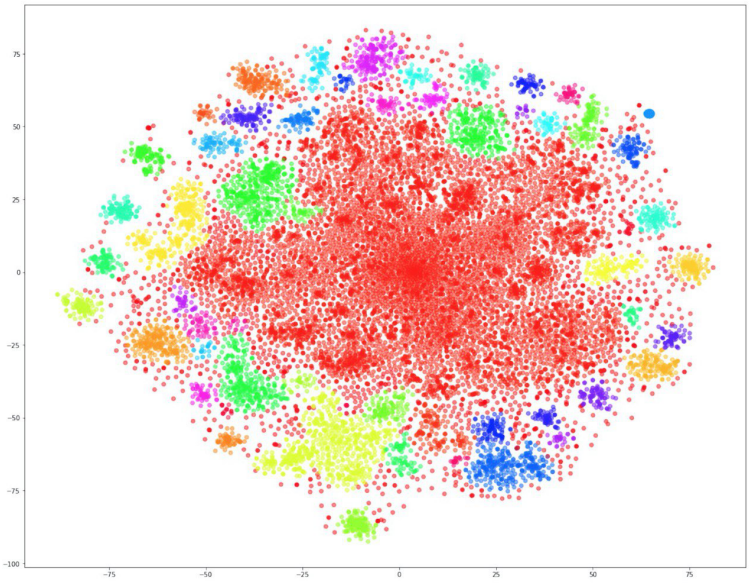

Figure 1: Preliminary topic clustering of the data set

On the basis of the obtained data, we have identified 40 basic topics, which have become features for further clustering of the data set. Based on our understanding of the overall thematic structure of the enclosure, we applied the non-negative matrix factorization method to identify the exact topic clusters. Base-matrix is represented by the two smaller matrices M1 and M2 (with the size of Number of documents * R, R* Number of words), where R = 40 (number of basic topics in the data set).

We got the weight of the words for each of the identified features. The words with the highest weight determine the content of each topic.

Topic features correspond to the matrix column numbers. We analyze the words with the maximum coefficients for each column and then assign titles to the topics according to the set of the most significant words. As a result, we obtained the following markup for the features (See Appendix, Table 4-6).

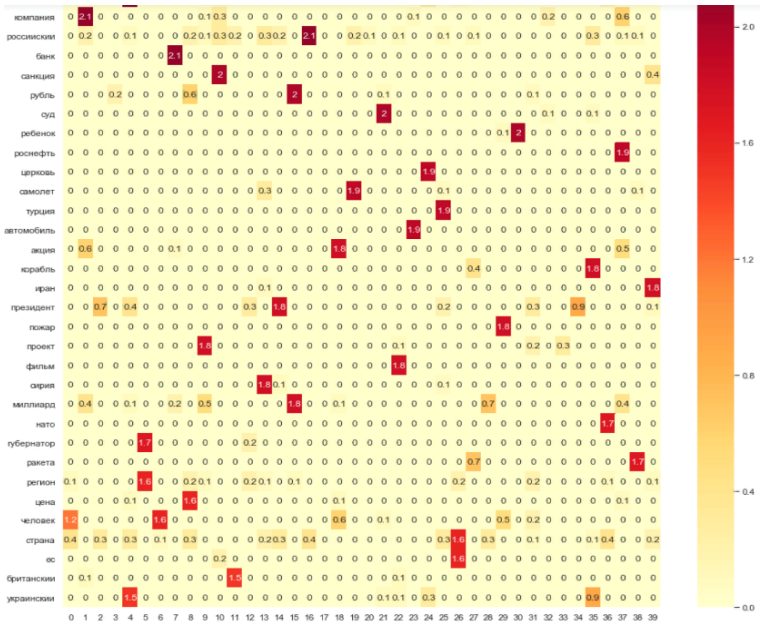

Figure 2: Weight of the words for each of the identified features

Similarly, the next step is to categorize the text into 40 features, taking into account the weight of the keywords. This allows us to identify the topic of each text fragment in our database, as well as to find out the texts which topics are actively involved in the formation of three-part hypertext structures. Thus, we can formalize the dialogical connections in electronic hypertext.

3 Results

Automatic analysis of the topical structure of the three-component structures identified 40 topics, which were then clustered into 21 main themes and 16 sub-themes. As a result, at the macro level the thematic structure includes *Accidents, People, Foreign policy, Economy, Justice, Construction, Pension reform, Government, Elections, Bank, Football, Cars, Aviation, Cosmos, Profits, Demonstrations, Internet, Family, Cinema, Military, Orthodoxy.*

Some of the topics are represented by a set of sub-themes, which is a marker of their discursive function in the electronic media. An example of such a topic is Foreign politics, the texts of which describe Russia's relations with various states: *Ukraine, Turkey, Iran, the European Union, the United States, Syria, China and England.*

However, these topics may be represented by even smaller thematic entities. In particular, the subtopic *USA* is split into the subtopic *Government* (k-words: *США, американский, трамп, дональд трамп, вашингтон, президент, белыи дом, вмешательство, конгресс, соединить, штат, заявить, администрация, штат, обвинение, американец, выбор, расследование, twitter, заявление, кндр*) and *Sanctions* (k-words: *санкция, против, сша, ввести, минфин, отношение, список, ограничение, введение, попасть, запрет, россиискии, бизнесмен, мера, американский, вводить, новое*).Topics *Economy* (Business, Oil, Budget), *Justice* (Court, Investigation), *Military* (Army, Navy), *Government* (President, Head of the administration) and *Internet* (Social networks, Blockades, Internet resources) are

also characterized by a complex structure.

Keyword frequency analysis within each topic showed that the greatest value is attached to the topic *Foreign policy*, the subtopics of which are ranked in the next sequence: *USA*– 6%, *Ukraine* – 4,2%; *England*– 4,1%, *China* – 3,8%; *Turkey я* – 3,7%; *Iran* – 3,5%; *European Union* – 3,4%; *Syria* – 3,2% . This topics are also of high value for the news discourse: *People* (about 5% of the total thematic landscape), *Justice* (Investigation- 4.2%; court - 3.3%); *Economy* (*Oil* - 3.8%, *Business*- 3.5%); *Pension reform* (3.4%); *Construction* (3%).

Topic analysis of the three-component structures of e-hypertext has shown that the author can use two different strategies to create it: fragments of texts can be combined by a common topic (monothematic hypertext) or fragments belong to different topic groups (polythematic hypertext).

The monothematic three-part structures include texts on the following topics: F*ire, Syria, Blockages, Cosmos, Orthodoxy.* So, hypertext structures of this type are characterized by "thematic deafness".

Table 1: Links-reactions to topic groups of texts

Keywords	**Names of links**	**Numbers**
Бюджет	заявить	99
	писать	92
	сообщать	70
	говорить	49
	предложить	45
Фонд	сообщать	103
	заявить	90
	писать	79
	опубликовать	63
	объявить	59
	говорить	50
	рассказать	38
Сумма	сообщить/сообщать	205
	писать	102
	заявить	100
	говорить	47
	объявить	45
	рассказать	41
	арестовать	40

The data in table 1 demonstrate, that these topics have the lowest inter-thematic transition coefficient (Itc $< 0{,}5$), which is calculated by the formula: inter-thematic transitions / (intra-thematic transitions + inter-thematic transitions).

The minimal inter-thematic potential is characterized by the topic *Orthodoxy*, which is represented by the following k-words: *церковь, решение, октябрь, принять, передача, заявить, глава, порошенко, структура, русский, действие, отменить, восстановление, создание, москва, признать, действовать.* The analysis of nominations of links in monothematic three-part hypertext structures shows that links reflect the topic of these text fragments (*предоставление автокефалия* (20), *разорвать* (15), *разорвать связь*

Table 2: Correlation between topic and type of hypertext transition (monothematic hypertext)

Text Topic	Number of intra-thematic transitions	Number of inter-thematic transitions	Inter-thematic connectivity coefficient
Orthodoxy	680	75	0,099338
Cosmos	1126	340	0,231924
Blockages	447	136	0,233276
Syria	756	262	0,257367
Fire	455	158	0,257749

Table 3: Correlation between topic and type of hypertext transition (polythematic hypertext)

Text Topic	Number of intra-thematic transitions	Number of inter-thematic transitions	Inter-thematic connectivity coefficient
Internet	9	91	0,91
Budget	222	552	0,713178
Profits	256	619	0,707429
Military	286	593	0,67463
Foreign policy	356	651	0,646475
People	749	1175	0,610707
Construction	596	788	0,569364
Business	698	902	0,56375
European Union	478	543	0,531832

константинопольский патриархат (13), *прекращение участие структура* (12)) or express the tonality of the authors reception (*пригрозить* (4), *скандал* (3)). In general, the coherence of the Orthodoxy texts is reflected in the nomination of the links, which often refers to the content/title of the target text. Thus, for example, the link to the nomination "разорвать" ("break") is the authors reaction to fragments with the titles: *РПЦ разорвала отношения с Константинопольским патриархатом, Константинополь отлучили от РПЦ: что означает церковный раскол.*

The strategy of inter-thematic hypertext transitions is most often used when the initial text relates to the topic Internet. The target texts are fragments characterized by the subthemes Oil, Business, USA, "Ukraine. In our opinion, this is explained by the functional specifics of this type of connection: the vocabulary of "Internet" thematic group marks the hypertext transition to the fragments, the function of which is to confirm the evidentiality of the initial text. Thus, the inter-thematic potential of hypertext structures is determined by the authors attitude to confirm the veracity of the published information.

As Table 1 shows, the texts related to *Politics, Economics* and *Construction* fields show the greatest potential for generating inter-thematic transitions (Itc $>$ 0,5). Text fragments of one of the most active topics Budget are represented by the following key words *миллиард, рубль, бюджет, фонд, сумма, составить, долг, минфин, средство, расход, выручка, триллион рубль, объесть, доход, акция, выплата, вырасти, актив, триллион, кредит.*

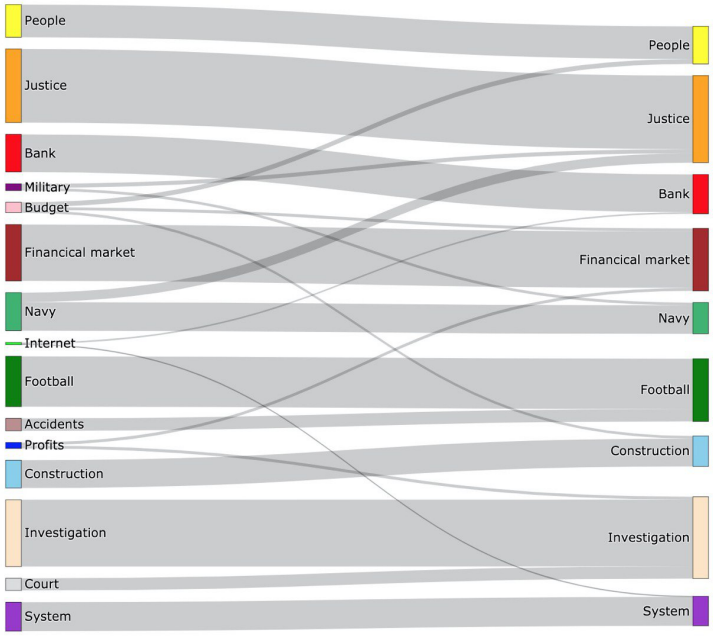

Figure 3: The strongest topic connections in three-part hypertext structures

We have identified the nominations of links, which are the author's reaction to text fragments with keywords *бюджет, фонд, сумма.*

Most of the frequency links do not inform about the topic of the target text, and do not express the tonality of the authors reception. They indicate the format/genre of the target text, which can be determined by the inter-thematic transition, but also by the context of the broader expression of the semantics of the hypertext transition, as well as the popularity of the texts of the topic (and thus the use of the most frequent verb links). We have identified the strongest thematic connections in three-part hypertext structures, limiting the minimum number of hypertext connections to 20 examples (see Fig.3).

As the data show, the structures of three-part hypertext elements often form the texts of related topics: Budget – People, Construction, Military – Navy, Profits – Financials market, Court – Investigation. However, hypertext connections also reveal nontrivial topical connections, which are explained by the specifics of the publicist discourse of the given period. For example, the connection between the themes of the Index and the Budget is determined by the criminal action of Russian footballers Kokorin and Mamaev, which is widely reflected in the media. The correlation between the Navy and Justice topics appeared

4 Discussion

In this article we carried out topic modeling of hypertext structures and revealed two types of the author's strategy: the creation of monothematic and polythematic hypertexts. The analysis of three-part hypertext elements has shown that the potential for inter-thematic transitions is determined not only by the specifics of the author's reception, but also by the specificity of the intersection of texts in the media discourse. To a large extent, the topic structure of e-hypertext is determined by the thematic dominants of the publicist discourse

in a certain period of time. The approach to creating metrics that define the potential for intertext transitions is the discussion.

The final sample did not include the topic Meetings, which is low-frequency in the corpus, but creates thematic links with 33 topics: Court, Investigation, Business, Accidents, Head, Elections, Cinema, Orthodoxy, Construction, Ukraine, EU, Oil, People, Foreign Affairs, Revenue, Navy, Cars, Bank, President, Social Networks, Blocks, Budget, Aviation, Internet, Cosmos, Football, Iran, USA/elections, Army, England, USA/ Sanctions, Turkey, Syria. The question of correlation between the author's strategy and the type of semantic connection between the nomination of the link and the target text also remains controversial.

Acknowledgements

The reported study was funded by RFBR according to the research project № 18-312-00010.

References

[Atzenbeck, Nürnberg, 2019] Atzenbeck C., Nürnberg P.J. Hypertext as method HT 2019 - Proceedings of the 30th ACM Conference on Hypertext and Social Media. pp. 29-38

[Ryazanseva, 2010] Ryazantseva T. I. (2010) Hypertext and electronic communication, Moscow: LKI. 256 p.

[Shulginov, 2016] Shulginov V. A. (2016) Cognitive model of hypertext Bulletin of Kemerovo State University, Vol. 4, pp. 233–238.

[Genette 1982] Genette G. (1982) Palimpsestes. La litterature au second degree / G. Genette. - Paris: Editions du Seuil, 1982, 467 p.

[Nelson, 1993] Nelson T (1993). Literary machines. - Sausalito, CA: Mindful Press, 1993. -220 p.

[Patterson] Patterson N (2000) Hypertext and the Changing Roles of Readers. G. English Journal,Vol. 90. № 2, pp. 74–80.

[Bart] Bart R. (1994) Selected Works by Bart R. Semiotics. Poetics.Moscow: Progress: Univers, 616 p.

[Salmeron, Kintsch, Canas, 2006] Salmerón, L., Kintsch, W.Cañas, J. (2006a) Reading strategies and prior knowledge in learning from hypertext. Memory and Cognition, 34, pp. 1157–1171

[Dedova, 2008] Dedova O. V (2008).The theory of hypertext and hypertext practice in RuNet. Moscow: MAKS Press, 2008. 284 p.

[Commer, 1979] Comer D. (1979) The ubiquitous b-tree. Computing Surveys, 11(2), June 1979, pp. 121–137

[Kutuzov, Kuzmenko] Kutuzov A., Kuzmenko E. (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham, pp.155-161

[Mangen, 2008] Mangen A. (2008) Hypertext fiction reading: haptics and immersion. Journal of Research in Reading, Vol. 3, pp. 404 – 419

Appendix

Table 4: Features and k-words. Part 1

Topic ID	NMF - components
34	['система', 'дать', 'использовать', 'код', 'файл', 'функция', 'работа', 'устройство', 'задача', 'использование', 'сеть', 'сервер', 'приложение', 'например', 'модель', 'пример', 'значение', 'каждый', 'помощь', 'нужно']
0	['человек', 'очень', 'самый', 'большой', 'еще', 'говорить', 'время', 'хороший', 'хотеть', 'жизнь', 'просто', 'стать', 'знать', 'делать', 'сказать', 'мир', 'работать', 'сделать', 'вопрос', 'думать']
6	['человек', 'октябрь', 'произойти', 'полиция', 'погибнуть', 'район', 'город', 'взрыв', 'москва', 'результат', 'сообщать', 'сообщить', 'пострадать', 'пострадавший', 'инцидент', 'находиться', 'сообщаться', 'мужчина', 'ранее', 'здание']
4	['украина', 'украинский', 'киев', 'крым', 'порошенко', 'донбасс', 'территория', 'заявить', 'днр', 'президент', 'власть', 'республика', 'газпром', 'риа новость', 'конфликт', 'область', 'депутат', 'страна', 'слово', 'журналист']
31	['дело', 'сотрудник', 'скр', 'следствие', 'уголовный дело', 'расследование', 'задержать', 'следователь', 'фсб', 'следственный', 'управление', 'следственный комитет', 'фигурант', 'обвинять', 'обыск', 'адвокат', 'орган', 'убийство', 'бывший', 'подозревать']
16	['цена', 'рост', 'рынок', 'уровень', 'топливо', 'нефть', 'снижение', 'ставка', 'вырасти', 'повышение', 'инфляция', 'рубль', 'баррель', 'цена нефть', 'курс', 'экономика', 'прогноз', 'стоимость', 'валюта', 'правительство']
2	['сша', 'американский', 'трамп', 'дональд трамп', 'вашингтон', 'президент', 'белый дом', 'вмешательство', 'конгресс', 'соединить штат', 'заявить', 'администрация', 'штат', 'обвинение', 'американец', 'выбор', 'расследование', 'twitter', 'заявление', 'кндр']
1	['компания', 'сделка', 'акция', 'доля', 'группа', 'рынок', 'бизнес', 'акционер', 'крупный', 'продажа', 'совет директор', 'рбк', 'миллион', 'актив', 'гендиректор', 'group', 'инвестор', 'ооо', 'владелец', 'бизнесмен']
17	['законопроект', 'возраст', 'госдума', 'закон', 'пенсия', 'повышение пенсионный', 'пенсионный', 'правительство', 'поправка', 'депутат', 'пенсионный реформа', 'женщина', 'гражданин', 'документ', 'внести', 'реформа', 'принять', 'предложить', 'изменение', 'право']
3	['суд', 'адвокат', 'решение', 'дело', 'судья', 'арест', 'приговор', 'иск', 'обвинять', 'ходатайство', 'заседание', 'защита', 'срок', 'признать', 'судебный', 'москва', 'обвинение', 'мера пресечение', 'право', 'арестовать']
27	['проект', 'строительство', 'развитие', 'газпром', 'мост', 'реализация', 'инфраструктура', 'построить', 'объект', 'работа', 'газа', 'завод', 'участок', 'инвестиция', 'создание', 'финансирование', 'мощность', 'программа', 'северный', 'новый']
5	['президент', 'встреча', 'путин', 'владимир путин', 'лидер', 'переговоры', 'саммит', 'трамп', 'государство', 'вопрос', 'кремль', 'визит', 'состояться', 'дмитрий песков', 'сказать', 'глава', 'заявить', 'дональд трамп', 'обсудить', 'песок']

Table 5: Features and k-words. Part 2

Topic ID	NMF - components
25	['глава', 'губернатор', 'регион', 'отставка', 'пост', 'республика', 'рбк', 'правительство', 'должность', 'администрация', 'господин', 'источник', 'президент', 'назначить', 'министр', 'заместитель', 'александр', 'назначение', 'руководитель', 'область']
12	['выбор', 'кандидат', 'партия', 'выборы', 'тур', 'голосование', 'голос', 'цик', 'единый', 'избиратель', 'кпрф', 'депутат', 'мэр', 'результат', 'сентябрь', 'регион', 'пройти', 'глава', 'парламент', 'политический']
10	['банка', 'банк', 'цб', 'кредит', 'кредитный организация', 'банковский', 'актив', 'финансовый', 'ставка', 'регулятор', 'клиент', 'сбербанк', 'втб', 'операция', 'капитал', 'средство', 'вклад', 'кредитный', 'открытие', 'счет']
20	['матч', 'команда', 'футболист', 'клуб', 'сборный', 'игра', 'футбол', 'чемпионат мир', 'игрок', 'спортсмен', 'счет', 'сезон', 'победа', 'минута', 'играть', 'тур', 'состав', 'стать', 'московский', 'октябрь']
21	['автомобиль', 'машина', 'водитель', 'модель', 'продажа', 'гибдд', 'дорога', 'двигатель', 'тысяча', 'новый', 'движение', 'авария', 'рынок', 'место', 'штраф', 'транспорт', 'продать', 'скорость', 'производство', 'пассажир']
11	['британский', 'великобритания', 'солсбери', 'лондон', 'вещество', 'новичок', 'март', 'дипломат', 'расследование', 'инцидент', 'обвинение', 'заявить', 'полиция', 'мид', 'сотрудник', 'бывший', 'агент', 'спецслужба', 'москва', 'разведка']
18	['ракета', 'авария', 'роскосмос', 'мкс', 'союз', 'запуск', 'экипаж', 'полет', 'космический', 'спутник', 'октябрь', 'старт', 'произойти', 'комиссия', 'причина', 'космос', 'станция', 'источник', 'система', 'земля']
9	['санкция', 'против', 'сша', 'ввести', 'минфин', 'отношение', 'список', 'ограничение', 'введение', 'попасть', 'запрет', 'российский', 'лицо', 'август', 'ес', 'бизнесмен', 'мера', 'американский', 'вводить', 'новое']
22	['самолет', 'борт', 'авиакомпания', 'пассажир', 'аэропорт', 'рейс', 'полет', 'находиться', 'экипаж', 'вертолет', 'погибнуть', 'упасть', 'минобороны', 'минута', 'двигатель', 'причина', 'транспорт', 'сообщить', 'мчс', 'место']
36	['страна', 'ес', 'нато', 'европейский', 'евросоюз', 'европа', 'альянс', 'соглашение', 'германия', 'саммит', 'совет', 'министр', 'великобритания', 'заявить', 'грузия', 'решение', 'государство', 'польша', 'отношение', 'парламент']
13	['сирия', 'сирийский', 'боевик', 'террорист', 'удар', 'израиль', 'оон', 'группировка', 'операция', 'атака', 'район', 'территория', 'минобороны', 'сила', 'город', 'организация', 'войско', 'урегулирование', 'военный', 'нанести']
7	['миллион', 'тысяча', 'рубль', 'около', 'сумма', 'размер', 'штраф', 'составить', 'доход', 'составлять', 'зарплата', 'данные', 'месяц', 'стоимость', 'деньга', 'россиянин', 'средний', 'общий', 'число', 'квартира']
15	['акция', 'митинг', 'навальный', 'задержать', 'полиция', 'проведение', 'участник', 'организатор', 'москва', 'человек', 'мероприятие', 'активист', 'против', 'задержание', 'сторонник', 'город', 'согласовать', 'власть', 'пройти', 'полицейский']
8	['российский', 'рф', 'москва', 'мид', 'федерация', 'международный', 'дипломат', 'иностранный', 'организация', 'посольство', 'гражданин', 'представитель', 'россиянин', 'спортсмен', 'сообщить', 'министр', 'сказать', 'крым', 'официальный', 'отметить']

Table 6: Features and k-words. Part 3

14	['пользователь', 'facebook', 'google', 'сервис', 'яндекс', 'соцсеть', 'приложение', 'дать', 'реклама', 'социальный сеть', 'информация', 'контент', 'компания', 'персональный', 'доступ', 'интернет', 'apple', 'сайт', 'twitter', 'личный']
29	['ребенок', 'школа', 'родитель', 'семья', 'подросток', 'детский', 'женщина', 'образование', 'больница', 'класс', 'медицинский', 'врач', 'помощь', 'минздрав', 'март', 'рождение', 'полигон', 'жизнь', 'известие', 'московский']
37	['военный', 'минобороны', 'армия', 'вооружение', 'ракета', 'военнослужащий', 'комплекс', 'войско', 'нато', 'вооружённый сила', 'сила', 'оборона', 'оружие', 'граница', 'операция', 'техника', 'система', 'боев', 'российский', 'положение']
28	['миллиард', 'рубль', 'бюджет', 'фонд', 'сумма', 'составить', 'долг', 'минфин', 'средство', 'расход', 'выручка', 'триллион рубль', 'объесть', 'доход', 'акция', 'выплата', 'вырасти', 'актив', 'триллион', 'кредит']
24	['фильм', 'режиссер', 'картина', 'театр', 'премьера', 'роль', 'сцена', 'премия', 'герои', 'алексей', 'история', 'культура', 'хороший', 'выйти', 'январь', 'экран', 'письмо', 'главный', 'получить', 'женщина']
23	['китай', 'пошлина', 'товар', 'китайский', 'торговый', 'ввести', 'импорт', 'торговля', 'введение', 'продукция', 'миллиард', 'тариф', 'мера', 'поставка', 'воина', 'страна', 'вашингтон', 'американский', 'ограничение', 'объем']
26	['турция', 'турецкий', 'саудовский аравия', 'поставка', 'журналист', 'отношение', 'убийство', 'газпром', 'страна', 'октябрь', 'власть', 'российский', 'попытка', 'заявить', 'президент', 'турист', 'сторона', 'газа', 'граница', 'связь']
39	['иран', 'ядерный', 'соглашение', 'иранский', 'сделка', 'нефть', 'программа', 'подписать', 'выход', 'санкция', 'договор', 'баррель', 'израиль', 'вашингтон', 'переговоры', 'действие', 'страна', 'международный', 'май', 'поставка']
32	['корабль', 'украинский', 'судно', 'ноябрь', 'фсб', 'морской', 'порт', 'военный', 'провокация', 'задержать', 'положение', 'россииский', 'инцидент', 'экипаж', 'ремонт', 'возбудить уголовный', 'войти', 'сила', 'применить', 'доставить']
33	['пожар', 'мчс', 'человек', 'здание', 'площадь', 'погибший', 'погибнуть', 'март', 'пострадавший', 'произойти', 'центр', 'место', 'жертва', 'сообщить', 'ск', 'безопасность', 'возникнуть', 'четыре', 'результат', 'пострадать']
30	['роснефть', 'нефть', 'акция', 'сделка', 'компания', 'месторождение', 'нефтяной', 'рбк', 'соглашение', 'газпром', 'пакет', 'добыча', 'глава', 'иск', 'министр', 'миллион', 'продажа', 'правительство', 'покупка', 'ноябрь']
38	['telegram', 'роскомнадзор', 'мессенджер', 'блокировка', 'фсб', 'заблокировать', 'реестр', 'информация', 'пользователь', 'доступ', 'закон', 'ведомство', 'требование', 'апрель', 'сообщение', 'сайт', 'оператор', 'ресурс', 'решение', 'сервис']
19	['церковь', 'решение', 'октябрь', 'порошенко', 'украина', 'украинский', 'общение', 'принять', 'передача', 'заявить', 'глава', 'структура', 'русский', 'действие', 'отменить', 'восстановление', 'создание', 'москва', 'признать', 'действовать']