ПЕРВЫЕ ШАГИ В НАУКЕ

Представляем работы молодых ученых

Aleksey Gorgadze

National Research University
Higher School of Economics
St. Petersburg, Russia
e-mail: alexei.gorgadze@gmail.com

Alina Kolycheva

National Research University
Higher School of Economics
Moscow, Russia
e-mail: avkolycheva@hse.ru

УДК 003 (05)

Mapping Ideas: Semantic Analysis of "PostNauka" materials

Abstract: "PostNauka" is a project (quite similar to the global set of conferences "TED") about modern fundamental science and the scientists who create it. The website was established in 2012 as a platform, participating in the popularization of scientific knowledge. More than 3000 materials were published on it. Surprisingly, no one has tried to carry out the analysis of these materials until now. We decided to fill this gap by conducting our own research on PostNauka. An inspiring example to be followed was the performance by Sean Gourley and Eric Berlow, who showed the map of themes similarity in the TED-speeches and the way new topics were born on the periphery of global themes. We conducted the same type of analysis for PostNauka to show the existence or the lack of links between topics, to define the extent of interdisciplinarity of the project and to check the rightness of sections proposed by the website. We took all video-materials transcripts published on www.postnauka.ru. Two types of methods were combined for analysis: semantic analysis methods and network analysis. As a result, we got the mind-maps of PostNauka materials. They demonstrate how the materials are tied, which disciplines are exceled and how they are sorting with the official classification of the website.

Keywords: Text Mining, Semantic networks, PostNauka, interdisciplinarity, co-word map.

Introduction

This research was inspired by the excellent S. Gourley's and E. Berlow's TED talk "Mapping ideas worth spreading"[1], devoted to the look of "global conversation", represented by the platform they were the speakers of. The main goal was to demonstrate the connection between talks through extracting and comparing the key concepts from each of the speeches transformed into text. The authors conducted geographical topic analysis, tested the role of gender in choosing the TED talk to be viewed and commented; pointed out central, peripheral and unique (isolated) topics. Following in their steps, we decided to draw the mathematical structure of ideas in Russian academic world, taking the transcripts of PostNauka speeches as a basis. PostNauka[2] is a popular scientific journal, which has been functioning as a learning platform alike TED, Arzamas or Coursera for 3 years. The project was started in 2012: it got more than 100 000 subscribers in the social networks, more than 600 scientists from different research fields took part in the project, and more than 3000 materials were published. The platform involves videos, lectures and courses of different disciplines. The speakers of different backgrounds perform there, regularly combining the theories or examples from a variety of fields, and hence the platform is considered to be interdisciplinary.

According to M. Burawoy, "each discipline, after all, offers but a partial perspective on that world, so interdisciplinarity offers a more complete picture. Interdisciplinarity can only enrich our understanding of the world" [*Burawoy*, 2013, p. 7]. There is no doubt, that interdisciplinarity is becoming more and more fashionable in the academic sphere. Our aim is to show how these disciplines are interconnected in the talks, published on PostNauka, so we could get a better understanding of the relationship between different scientific fields through the language they use. In respect that our project is on its early stage, we work only with the key words, chosen by PostNauka organizers as the markers to present a topic of talk. Further we are planning to include all words from the talks' transcripts and use LDA in combination with LSA solving the stated problem. In this article we will watch this relationship in dynamic, so it would be possible to notice the changes in penetration of topics and disciplines. In addition we will look at some statistics on the platform users and the activities changing the structure we investigate.

Related works

E-learning platforms became extremely popular with the researchers, paying attention to their influence on higher education [*Ruiz, Mintzer, Leipzig*, 2006; *Sife, Lwoga, Sanga*, 2007], economic costs [*Matei, Vrabie*, 2013], students' behavior [*Graf, List*, 2005; *Paechter, Maier, Macher*, 2010] and etc. We did not manage to find any publications on exploring the educational platforms as text corpora to extract the topics (certain distribution over the words [*Blei*, 2012]) and track their points of coincidence, that's why our research seems essential and original. The talks on PostNauka will be treated as "bags of words" to find "hidden thematic structure in large archives of documents" [*Blei*, 2012, p. 77] and follow the change of topics over time [*Wang and McCallum*, 2006]. Quite a similar project was

[1] http://www.ted.com/talks/eric_berlow_and_sean_gourley_mapping_ideas_worth_spreading
[2] https://postnauka.ru/about

conducted by D. Hall and his colleagues, who investigated the textual structure of three conferences (COLING, ACL, EMNLP) to designate their level of ideas' diversity [*Hall, Jurafsky, Manning*, 2008]. They figured out that all those conferences were becoming broader eventually, thou each of them shown different rate of "topic entropy". Taking this method in account, we will try to find out which talks or scientific fields have a narrow focus on a small amount of topics and which of them expand the borders using the language.

The Russian pioneers in this area of studies V. Nalimov and Z. Mul'chenko were among the first researchers, investigating the language or the "slang" of different spheres of knowledge to find the common words they have and the common ideas they use [*Nalimov, Mul'chenko*, 1969]. Those times the computer methods of text analysis were not as well developed as now, so in our research we can rely on more friendly ways of working with huge text corpora. "Topic modeling" is applied by modern researchers, who work with the citations, try to construct a hierarchical model, build semantic networks or draw co-word maps [*Lehmann*, 1992; *Danowski*, 1993; *Blei et al.*, 2003; *Li, McCallum*, 2006; *Dietz et al.*, 2007]. A great variety of options can be chosen to visualize the results researchers come up with, including Pajec, ORA, Gephi and etc. [*Vlieger, Leydesdorff,* 2011; *Danowski*, 2009]. The combination of Text Mining and SNA has not become very popular yet, but these two methods have already been used in a couple of works [eg. *Mei, Cai, Zhang, Zhai*, 2008; *Leydesdorff, Nerghes*, 2016] and we hope, that our research will somehow supplement the list of the latest investigations within this area.

Data and methods

As it was mentioned in the previous parts, our research is based on video-talk materials published on the platform PostNauka. We collected detail information about all video-talks, which were published from 2012 to 2014 on the site (1228 talks). The data contains transcripts (or parts of transcripts), number of views, comments and reposts on other social network sites, date of publishing, duration of talk, key-words, speaker's name, speaker's academic degree and profession. We also decided to take the language of each talk (some speakers made their oral presentations in English) and the gender of a speaker into account.

Of course, as any researchers, we faced with a certain number of challenges. The data was uploaded in January 2016, when PostNauka had not yet changed its interface and had not structured the information about each video-talk. The other issue was related to the fact, that one word could have several meanings: "Words may have different meanings in different contexts, although one expects the meaning of specific words to be stable within a single text" [*Leydesdorff, Nerghes*, 2016: 12, cit. ex: *Leydesdorff, Hellsten*, 2006]. Inasmuch as we decided to leave the step, where we could work with all amount of words from the transcripts, we did not really have to deal with homonimia. This problem will be faced with on the next research stage. We have already conducted it, but the results we gor are so complicated, that we came up with the decision to dwell on key-words only in this article.

The preprocessing of texts included lemmatization of all texts using Yandex free soft for morphological analysis of the Russian language texts [*Segalovich*, 2003]. For data analysis and visualization we combined the methods of Text Mining and SNA, and used R-packages ("tm", "mallet", "dplyr", "rJava" ect.) and Gephi.

We created the network of video-talks based on shared key words. This graph shows how talks are thematically close to each other. We also used SNA parameters for analyzing network differences of topics and speeches, such as Degree Centrality (DC), Betweenness Centrality (BC), Closeness Centrality (CC).

Results and conclusions

The final graph (Figure 1) has a low density (0.151) and a high modularity (0.577). We got five major clusters (modularity classes) and four isolates. We decided to call them "humanitarian", "linguistic", "psychologic", "physical" and "biological" clusters of the network (Figure 2). Humanitarian cluster contains the speakers who identify themselves as historians, philosophers, sociologists, economists, philologists, culture experts, folklorists, geographers, anthropologists, lawyers etc. However, there is a separate psychological cluster, which includes eleven psychologists and one sociologist. Talks from this cluster have links with humanitarian cluster as well as biological cluster, acting as a bridge between them. Furthermore, there is also segregate linguistic cluster, which consists of linguists, philologists, historians and etc. It links humanitarian cluster with physical one. Physical cluster combines physicist, mathematicians, astrophysics, chemists, astronomers, IT-specialists, engineers and etc. Biological cluster — biologists, microbiologists, medics, anthropologists, biophysics, chemists, neuroscientists, bioinformatics etc.

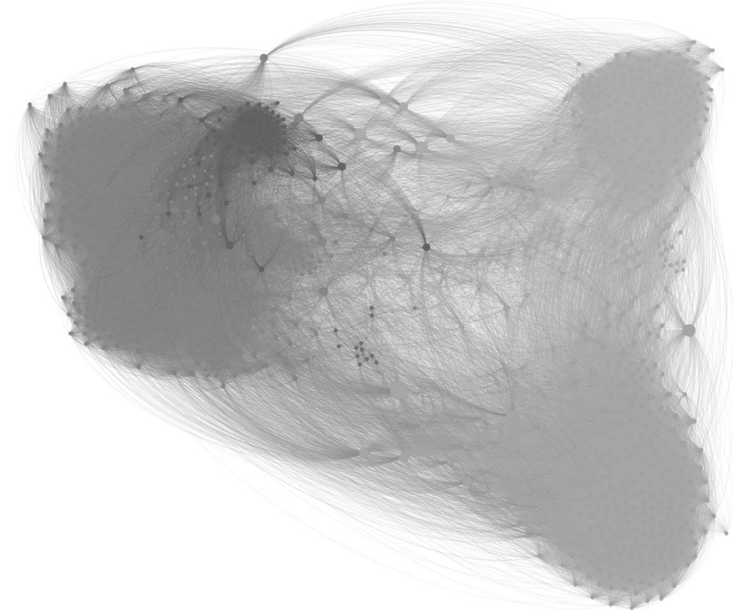

Figure 1. Network of all video-talks, based on shared key words. Size of nodes — Betweenness Centrality. The weight of edges represents the number of intersections between key words

Our attention must be payed to several other facts. First, we can see that anthropologists are divided into two large clusters: humanitarian and biological clusters. In the first case, there are cultural anthropologists, who explore mode of life, culture, rites and etc. In second — anthropologists, studding biology factors, human evolution, genetics, race and etc.

In addition to this, there are chemists, who are presented in two different clusters: physics and biology. «Physical» chemists focus on inorganic aspects, and are trying sometimes to link them with organic issues. For example: «Physical properties of the polymers and polymeric materials using a protein structure prediction problem»[3]. «Biological» chemists are close to medicine, immunity, biochemistry and etc.

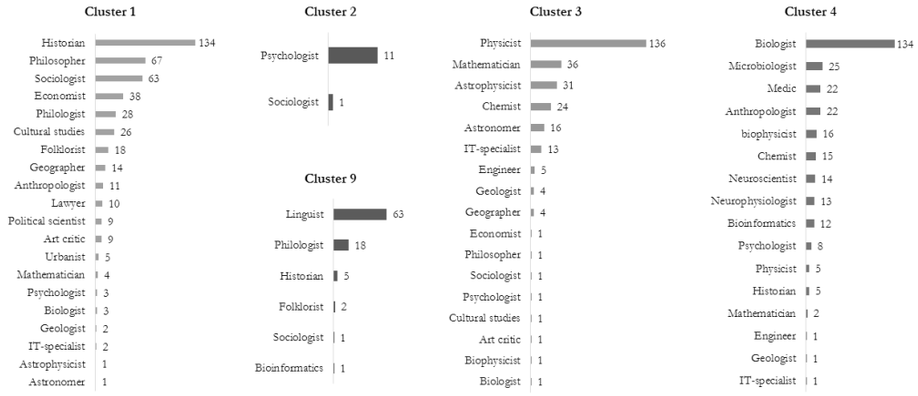

Figure 2. Distribution of spikers' professions by clusters (modularity classes)

We used ANOVA test to check significant differences of video-talk characteristics. There is no significant distinction by all indicators between Male and Female. Average page views, comments, repost, recording duration, and network metrics of centrality (degree, betweenness closeness) do not differ by gender.

We also checked whether there were differences between the talks by the years (2012, 2013 and 2014). We assumed that the videos posted earlier had more opportunities for users to watch them and write comments. Talks in 2012 had significantly more views, indeed. However, we have seen a sharp decline in 2013 and rise in 2014. In addition to this, the users' activity expressed through the number of comments and reposts was higher in 2013. In other words, despite the fact that in 2013, there were fewer views, the video comment activity was much more than during other years. We also found a significant increase in the duration of talks from 2012 to 2014.

Dedicated Clusters differ on these indicators. The cluster "2" ("psychology") has greatly more views and reposts. We also found that cluster "4" ("biology") has a larger average number of ties (DC) and average value of Closeness Centrality.

The language of talk has an effect on some indicators too. English language influences negatively the number of comments. Nevertheless, English-speaking talks have significantly higher level of Betweenness Centrality.

[3] https://postnauka.ru/video/38347

Table 1

Distribution of some talks' characteristics by spiker's sex, year of publishing, modularity class and language of talk

		Number	Views	Comments	Reposts	Time (sec.)	DC[4]	BC[5]	CC[6]
Sex	Female	195	6846.4	5.3	193.6	723.1	179.8	0.00096	0.5
	Male	933	7177.4	4.5	194.4	725.3	168.3	0.00085	0.5
	p-value		0.674	0.217	0.965	0.837	0.073	0.501	0.058
Year	2012	209	8251.4	4.7	167.0	621.5	176.5	0.00096	0.5
	2013	407	5773.6	6.0	206.5	730.6	166.8	0.00095	0.5
	2014	512	7728.9	3.4	195.6	762.6	170.5	0.00075	0.5
	p-value		0.002	0.000	0.083	0.000	0.371	0.275	0.127
Modularity class (cluster)	1	448	7059.3	5.4	207.1	733.5	172.0	0.00078	0.51
	2	12	12753.8	5.1	265.4	733.4	41.9	0.00006	0.46
	3	277	5948.5	3.3	146.3	726.0	142.6	0.00084	0.49
	4	296	7882.7	4.6	178.7	715.0	217.2	0.00106	0.53
	9	90	8035.4	4.75	325.3333	712.0	119.5	0.00085	0.47
	p-value		0.038	0.041	0.000	0.345	0.000	0.291	0.000
Language	English	63	4869.43	2.33	180.73	723.83	183.49	0.0016	0.52
	Russian	1065	7253.36	4.73	195.04	724.98	169.52	0.0008	0.51
	p-value		0.065	0.027	0.598	0.946	0.185	0.004	0.211

The graph of key words (Figure 3) illustrates the basic topics of all talks. The words with the highest BC are: history, biology, physics, Russia, mathematics, culture, philosophy, technologies, society, sociology. They shape and define the main thematic blocks of the talks.

Based on categorization of nodes proposed by Nerghes A. (Nerghes, 2016) we classify words by Degree and Betweenness Centralities. There are four classes:

- *Globally Central (GC) — High DC, High BC*
- *Locally Central (LC) — High DC, Low BC*
- *Gatekeeper (G) — Low DC, High BC*
- *Marginal (M) — Low DC, Low BC*

[4] Degree centrality (DC).
[5] Betweenness Centrality (BC).
[6] Closeness Centrality (CC).

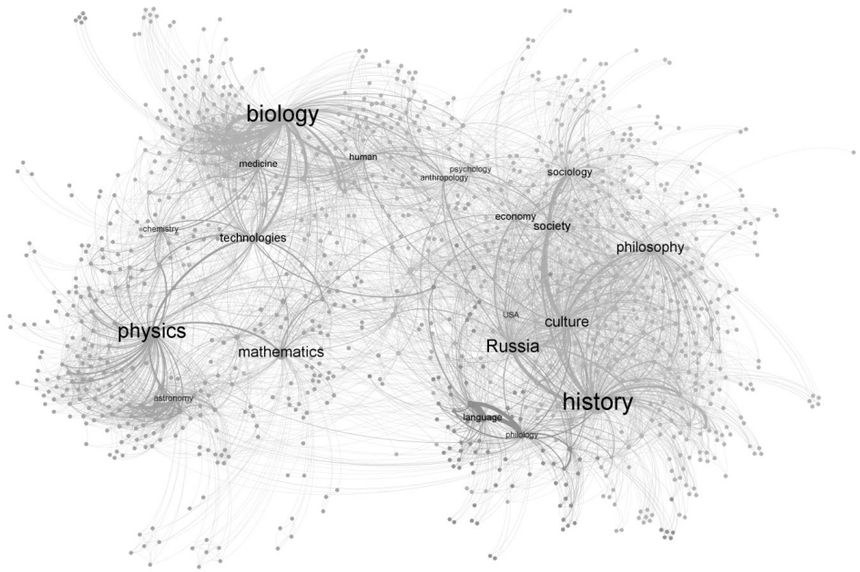

Figure 3. Network of key-words, with dedicated Top-20 words. Size of nodes — Betweenness Centrality. The weight of edges represents the number of videos, where key words were used together

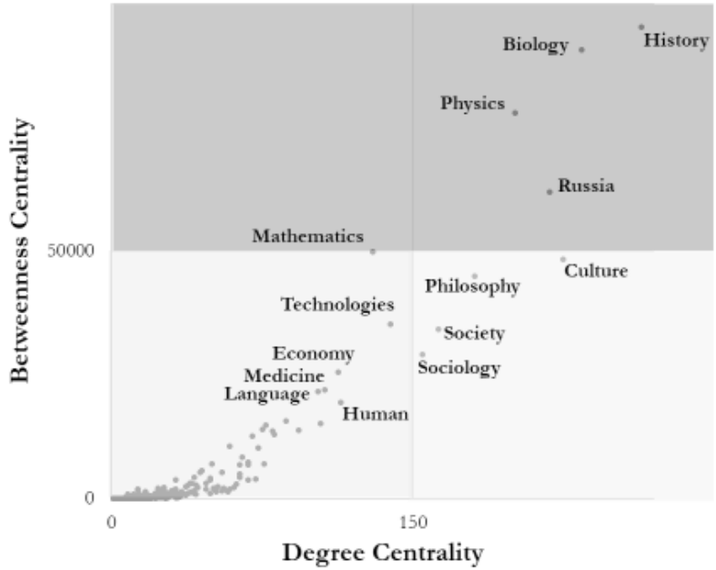

Figure 4. Key words centralization structure

As we can see on the Figure 4, there are four key words, which have high Degree and Betweenness Centralities: History, Biology, Physics, Russia. It means that these words are often found with other keywords (High DC) as well as they are connected with the words from different clusters. These terms though associate different clusters, but it is likely to be the cause of a large number of links. Moreover, we cannot see Gatekeepers — the words which have small number of links, but come from different thematic clusters. We can turn our attention to the fact that «mathematics» is the closest word to the Gatekeeper's status. If we look at the previous word network, we could also see that the mathematical cluster is located between different clusters and can serve as a bridge between them.

There are also four words in Locally Centrality class: Culture, Philosophy, Society, Sociology. It means that these terms are common with other words but coming from the same cluster. They have few links with other thematic clusters.

The vast number of words has low levels of Betweenness Centrality and Degree Centrality.

In conclusion it must be noted that the results of the research presented above are only offering the backlog for the further survey. Using the key-words for analysis is a rough method and it cannot be estimated as reliable as working with all words from the transcripts. There are some talks on the platform which have few key-words or none of them. In that case the key-words cannot be taken as preferable data for analysis, but they can help to compose the first impression on the interconnection of disciplines existing on the platform. For the next step we are planning to investigate the big corpora of texts to find out whether the speakers representing the disciplines use the language or the "slang", which consists of the words, uniting certain scientific areas. By combining Topic modeling method and Social Network Analysis we are expecting to evaluate our results.

References

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *The Journal of machine Learning research*, no. 3, pp. 993–1022.

Blei, D.M. (2012), "Probabilistic topic models", *Communications of the ACM*, vol. 55, no. 4, pp. 77–84.

Burawoy, M. (2013), "Sociology and Interdisciplinarity: The Promise and the Perils", *Philippine Sociological Review*, vol. 61, no. 1, pp. 7–20.

Danowski, J.A. (1993), "Network analysis of message content", *Progress in communication sciences*, no. 12, pp. 198–221.

Danowski, J.A. (2009), *Inferences from word networks in messages. The content analysis reader*, pp. 421–429.

Dietz, L., Bickel, S. and Scheffer, T. (2007, June), "Unsupervised prediction of citation influences" in *Proceedings of the 24th international conference on Machine learning*, ACM, pp. 233–240.

Graf, S. and List, B. (2005), *An evaluation of open source e-learning platforms stressing adaptation issues*, IEEE, pp. 163–165.

Hall, D., Jurafsky, D. and Manning, C.D. (2008, October), "Studying the history of ideas using topic models" in *Proceedings of the conference on empirical methods in natural language processing*, Association for Computational Linguistics, pp. 363–371.

Lehmann, F. (1992), "Semantic networks", *Computers & Mathematics with Applications*, vol. 23, no. 2–5, pp. 1–50.

Leydesdorff, L. and Nerghes, A. (2016), *Co-word Maps and Topic Modeling: A Comparison Using Small and Medium-Sized Corpora (n<1000)*, *available at:* https://arxiv.org/ftp/arxiv/papers/1511/1511.03020.pdf

Li, W. and McCallum, A. (2006, June), "Pachinko allocation: DAG-structured mixture models of topic correlations" in *Proceedings of the 23rd international conference on Machine learning*, ACM, pp. 577–584.

Matei, A. and Vrabie, C. (2013), "E-learning platforms supporting the educational effectiveness of distance learning programmes: a comparative study in administrative sciences", *Procedia-Social and Behavioral Sciences*, vol. 93, pp. 526–530.

Mei Q., Cai D., Zhang D. and Zhai C. (2008, April), "Topic modeling with network regularization" in *Proceedings of the 17th international conference on World Wide Web*, ACM, pp. 101–110.

Nalimov, V. and Mul'chenko, Z. (1969), *Scientometrics: Studies on the development of science as an informational process [Naukometriya: Izuchenie razvitiya nauki kak informacionnogo processa]*, Publishing house "Nauka", Moscow, Russia.

Nerghes, A. (2016), *Words in Crisis: A relational perspective of emergent meanings and roles in text*, available at: *http://dare.ubvu.vu.nl/bitstream/handle/1871/54227/complete_dissertation.pdf?sequence=1*

Paechter, M., Maier, B. and Macher, D. (2010), "Students' expectations of and experiences in e-learning: Their relation to learning achievements and course satisfaction", *Computers & education*, vol. 54, no. 1, pp. 222–229.

Ruiz, J., Mintzer, M. and Leipzig, R. (2006), "The impact of e-learning in medical education", *Academic medicine*, vol. 81, no. 3, pp. 207–212.

Segalovich, I. (2003), "A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine" in *MLMTA,* pp. 273–280, available at: *http://download.yandex.ru/company/iseg-las-vegas.pdf*

Sife A., Lwoga, E. and Sanga, C. (2007), "New technologies for teaching and learning: Challenges for higher learning institutions in developing countries", *International Journal of Education and Development using ICT*, no. 3(2).

Vlieger, E. and Leydesdorff, L., (2011), "Content analysis and the measurement of meaning: The visualization of frames in collections of messages", *Public Journal of Semiotics*, vol. 3, no. 1, pp. 28–50.

Информация для авторов и требования к рукописям статей, поступающим в журнал «Социология науки и технологий»

Социология науки и технологий
Sociology of Science and Technology

Журнал ***Социология науки и технологий*** (СНиТ) представляет собой специализированное научное издание.

Журнал создан по инициативе Санкт-Петербургского филиала Института истории естествознания и техники имени С. И. Вавилова Российской академии наук (СПбФ ИИЕТ РАН) в 2009 г. и издается под научным руководством Института.

Учредитель и издатель: Издательство «Нестор-История».

Периодичность выхода — 4 раза в год.

Свидетельство о регистрации журнала ПИ № ФС77–36186 выдано Федеральной службой по надзору в сфере массовых коммуникаций, связи и охраны культурного наследия 7 мая 2009 г.

Журнал имеет международный номер ISSN 2079–0910 (Print), ISSN 2414–9225 (Online). Входит в перечень рецензируемых научных изданий, рекомендованных ВАК (по специальностям 07.00.00 — исторические науки и археология; 22.00.00 — социологические науки; 09.00.00 — философские науки). Включен в российский индекс научного цитирования (РИНЦ), в европейский индекс журналов по общественным и гуманитарным наукам ERIH-PLUS.

Журнал публикует оригинальные статьи на русском и английском языках по следующим направлениям: наука и общество; научно-техническая и инновационная политика; социальные проблемы науки и технологий; социология академического мира; коммуникации в науке; история социологии науки; исследования науки и техники (STS) и др.

Публикации в журнале являются бесплатными для авторов. Гонорары за статьи не выплачиваются.

Требования к статьям

Направляемые в журнал рукописи статей следует оформлять в соответствии со следующими правилами (требования к оформлению размещены в разделе «Для авторов» на сайте журнала http://sst.nw.ru/):

1. Рукопись может быть представлена на русском и английском языках.
2. Рекомендуемый объем рукописи — до 40 000 знаков (включая — на русском и английском языках — название, аннотацию, ключевые слова, авторскую справку и список литературы). Текст предоставляется в форматах: .doc, .docx, .odt. Шрифт — Times New Roman, 12 кегль, интервал 1,5. Поля: слева — 3 см, сверху и снизу — 2 см, справа 1,5 см. Текст размещается без переносов. Абзацный отступ — 1 см.
3. Материалы для разделов «Рецензии», «Хроника научной жизни» и др. не должны превышать 10 000 знаков.
4. Автору необходимо представить:

а. Название статьи, аннотацию (на русском языке — в пределах 150 слов, на английском — от 250 до 300 слов). Машинный перевод категорически запрещен. Требования к аннотации — в разделе «Для авторов» на сайте журнала.