ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

Полянская Анна Кирилловна

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ МЕЖДИСЦИПЛИНАРНОСТИ В ВЫПУСКНЫХ КВАЛИФИКАЦИОННЫХ РАБОТАХ СТУДЕНТОВ ФГН НИУ ВШЭ

Topic Modeling Approach to Evaluating Interdisciplinarity in Theses of Faculty of Humanities, HSE University

Выпускная квалификационная работа студента 4 курса бакалавриата группы 171

Академический руководитель образовательной программы канд. филологических наук, доц. Ю.А. Ландер

_____

« » _____ 2021 г.

Научный руководитель

кандидат филол. наук, доцент Д.А. Скоринкин

_____

Москва 2021

Table of contents

Abstract

Recent studies in the field of scientific research analysis have proven computational methods to be extremely useful for gaining insight into the history, structure, and current state of scientific knowledge. This paper aims to contribute both to the existing research of academic texts and to the educational and research community of HSE University (and anyone else interested in science and education) by applying Topic modeling methods, namely LDA (Latent Dirichlet Allocation), to theses of students of Humanities Faculty and gaining useful insight into its structure and scientific tendencies.

Keywords: topic modeling, LDA, academic texts, interdisciplinarity.

Topic Modeling Approach to Evaluating Interdisciplinarity in Theses of Faculty of Humanities, HSE University

Scientific research analysis (SRA) is a very interesting research field, distinct from others mainly in several ways. First of all, the field itself can be called relatively new, although such type of analysis is very common and well known – it is conducted for every published paper and usually presented as a literature review (just as one in this paper). The mentioned 'novelty' lies in the methods it uses to achieve the same goals, meaning to give an insight to understanding and describing history, structure, and current state of scientific knowledge. Secondly, analysis of scientific research is to a certain degree a meta-filed, as its main research object is the research itself. Last, but not least, SRA is a very multidisciplinary field, as it combines a wide variety of methods from all the different more common fields, including Sociology, Mathematics, History, Literature studies and Computational Linguistics. I would consider SRA to be a part of Digital Humanities, as they share a lot of characteristics, including ones described above.

The purpose of this research is twofold. My first motivation is to propose an analysis of theses, written by the students of Faculty of Humanities at HSE University, to model interdisciplinarity and access to which extent my methods are suitable for such a task. The second motivation is my interest in working with Russian language because it is not as popular as English, and I aim to create a corpus of academic texts, suited for automatic analysis, which has never been done before (at least to my knowledge). It is also one of the reasons for choosing such type of texts as an object of my research. A thesis is a product of both science and education, and I believe that my work would contribute to better understanding of Humanities and scientific tendencies of students' work.

I also put forward two hypotheses and will try to prove whether they are true or false using the Topic Modeling Approach:

1. Master theses are distinct from bachelor in a way that they come either from a narrower field of research or, on the contrast, combine more than one broad research fields.
2. Masters programs such as *Cultural and Intellectual History: Between East and West* and *History of Artistic Culture and the Art Market* indeed show a more interdisciplinary approach, as it is advertised in the program description, while *Philosophy and Religious Studies* and *Russian as a Foreign Language in Cross-Linguistic and Cross-Cultural Perspective* represent highly specific topics, not presented anywhere else.

Literature Review

Digital methods for Scientific Research Analysis

The idea of using digital methods for scientific research analysis has been present for some time now. A variety of possible techniques is available for such kind of task, and some notable trends can be observed over the last three to five decades.

With the development of the web, it became possible to create very large databases of scientific papers such as the well-known Google Scholar. Due to the nature of research texts, nearly all of them are linked together through references and citations, making them extremely suitable for methods of network analysis. Thus, Citation analysis is one of the earliest and most popular approaches (Small, 1973) (Braam, Moed, & van Raan, 1991). As it was stated later in (Wagner, et al., 2011), "Assessment of research outputs should be broadened beyond those based in bibliometrics". The focus shifts to working with textual data that can be retrieved from the papers, for example, using keywords for classification tasks (Dutta, 2008). However, working with larger volumes of text such as abstracts and full texts requires applying different techniques, one of them being Topic Modeling (TM).

TM approaches has been successfully applied to collections of research papers of different popular fields of science, example being journals of Informatics (Zhu, Zhang, & Wang, 2016), *PubMed* (Älgå, Eriksson, & Nordberg, 2020), Library and Information Science journals (Han, 2020). I would like to mention (Hall, Jurafsky, & Manning, 2008) as a first case of topic modeling of Computational Linguistics, (Paul & Girju, 2009) as a great inspiration for my research, as they used TM on journals from three field: Linguistics, Computational Linguistics and Education and provided a thorough analysis of the resulting topis and trends inside each field, both in synchrony and diachrony, and (Bakarov, Kutuzov, & Nikishina, 2018), which is a more diachronic study of Russian NLP (although, unfortunately, on English texts). Studying interdisciplinarity in scientific research (IDR) and measuring it is also quite popular. Among the mentioned methods, it can also exploit less obvious data like texts of award proposals (Nichols, 2014).

Topic Modeling

Topic modeling is a machine learning approach, or rather a group of algorithms, developed for analyzing collections of documents. The main assumption here is that such collection has a latent structure, which can be described in terms of 'themes' or 'topics'.

More technical explanation is given in (Hofmann, 2001) as following:

"Each document in a given corpus is thus represented by a histogram containing the occurrence of words. The histogram is modeled by a distribution over a certain number of topics, each of which is a distribution over words in the vocabulary. By learning the distributions, a corresponding low-rank representation of the high dimensional histogram can be obtained for each document".

This approach is quite straightforward and intuitive. A person tasked with describing some corpora would do it in a same way, for example by dividing a pile of magazines into categories like "sport, cooking, cars, fashion, science", and each category then described by the prototypical words like "play, football, game, ball, speed" vs "food, recipe, bread, oven, etc." for the first ones.

The most well-known algorithms are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM). I will briefly describe all of them to provide a better understanding of the internal mechanism of Topic modeling.

LSA

Latent Semantic Analysis was proposed in (Deerwester, 1988) and (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) as a technique for information retrieval task, making use of Singular Value Decomposition dimension reduction of a transformed Term-Document matrix. As described in (Dumais, 2005), "LSA is a fully automatic statistical approach to extracting relations among words by means of their contexts of use in documents, passages, or sentences. It makes no use of natural language processing techniques for analyzing morphological, syntactic, or semantic relations. Nor does it use humanly constructed resources like dictionaries, thesauri, lexical reference systems (e.g., *WordNet*), semantic networks, or other knowledge representations. Its only input is large amounts of texts. LSA is an unsupervised learning technique. It starts with a large collection of texts, builds a term-document matrix, and tries to uncover some similarity structures that are useful for information retrieval and related text-analysis problems." This general features of LSA are also true for all the other algorithms mentioned above.

PLSA

Probabilistic Latent Semantic Analysis was developed by (Hofmann, 1999). It follows the following steps: "Documents are represented as a multinomial probability distribution over topics (which are assumed but not directly observed). The generative model for a term-document pair is the following: select a document with probability *P(d)*, select a latent class or topic with probability

P(z|d), and generate a term with probability *P(t|z)*. Expectation maximization, a standard machine-learning technique for maximum likelihood estimation in latent variable models, is used to estimate the model parameters" (Dumais, 2005).

LDA

Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) is a Bayesian version of the previous algorithm, it uses Dirichlet priors both for topic- and word-distributions. In somewhat simple words, PLSA uses "document → topic → word" sampling sequence. For LDA, generative steps can be thought of like this: "dirichlet distribution 1 → topic distribution → topic *Z* → dirichlet distribution 2 → word distribution of topic *Z* → word". The advantage of LDA is that it can work with new documents. In PLSA document probability is fixed, meaning that there no data for a document the algorithm has not seen. In LDA it can just be sampled from a distribution.

CTM

Correlated Topic Model is an extension to LDA model by (Blei & Lafferty, 2007). The difference is summarized in (Mahmood, 2013): "LDA cannot model the correlations among topics. For example the topic "genetics" is more likely to be similar to "disease" than to "astronaut". <…> CTM can model the correlations among topics".

Furthermore, there are several algorithms that take into consideration the time period of the document creation, like A Non-Markov Continuous-Time Method or Dynamic Topic Models (Alghamdi & Alfalqi, 2015). These models are called Topic Evolution Models and can be applied to many different tasks involving diachrony, for example, analyzing topic evolution in the scientific literature over time.

Data

My research is different from most studies described above in two ways. I use Topic Modeling primarily as a method for analysis and only secondarily as a tool for generating features for further classification. The reason is obvious: there is no need for me to label works because I already have a two-level classification as the texts were written by students of different Educational Programs, and those programs are linked to certain Schools of HSE Humanities Faculty. This allows me to compare the results of my algorithm to 'gold standard' classification, which is presented in Figure 1. The motivation behind this particular classification (combining bachelor's and master's programs into so-called schools) is to balance the distribution of texts (see Figure 2) and to link theses to a certain broad scientific field.

```
School of Cultural Studies:
        Visual Culture
        Applied Cultural Studies
        Cultural Studies
School of History:
        History of Knowledge and Social History
        History
        History of Artistic Culture and the Art Market
        Historical Knowledge
        History of Arts
        Medieval Studies
School of Linguistics:
        Fundamental and Computational Linguistics
        Computational Linguistics
        Language Theory and Computational Linguistics
        Russian as a Foreign Language in Cross-Linguistic and Cross-Cultural Perspective
        Linguistic Theory and Language Description
School of Philological Studies:
        Philology
        Comparative Studies: Russian Literature in Cross-cultural Perspective
        Cultural and Intellectual History: Between East and West
        Language Policy in the Context of Ethnocultural Diversity
        Russian and Comparative Literature
School of Philosophy:
        Philosophical Anthropology
        Philosophy
        Philosophy and Religious Studies
        Philosophy and History of Religion
```

Figure 1. The structure of Faculty of Humanities, schools and programs.[1]

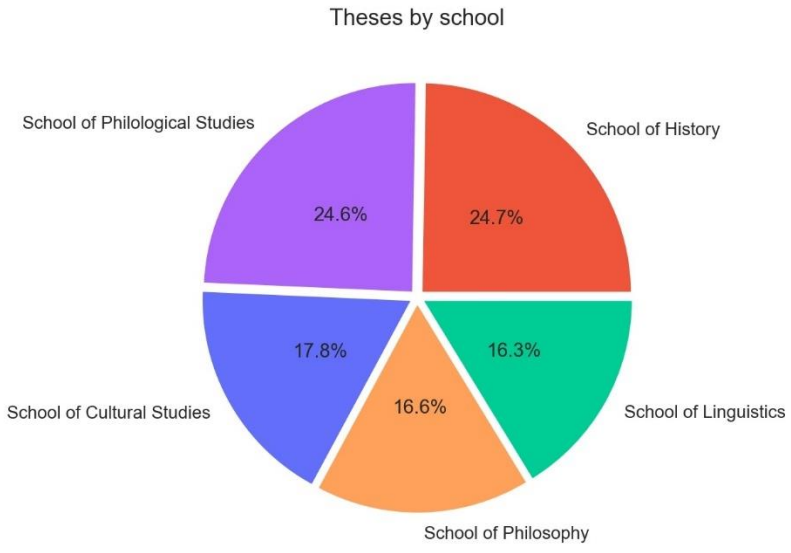


Figure 2. The distribution of theses over schools.

[1] It should be noted here, that from summer of 2020 School of Philosophy and School of Cultural Studies were combined together into one department, but I kept them separate to be able to distinguish between these scientific fields.

My approach also differs in a type of texts I am using. Most studies are conducted on corpora, made only of papers' abstracts and keywords, because this information is always publicly accessible in contrast to papers' full texts. Fortunately, HSE University has an open database[2] of all theses from 2015 (about 51,000 entries), and with over 18,000 available texts.

In this study, I am only focusing on one faculty, so raw dataset consists of 889 available texts. However, due to some technical difficulties during the very first step of data processing, this number dropped to 578 (for the reasons see chapter "Methods"). Total token counts for raw and processed texts are 9,688,788 and 4,607,842 respectively.

It should also be noticed here that 25 of the original 889 texts come from master's program *Creative Writing* and are not academic papers, but rather chapters from student's literature pieces, so they are dismissed from the data set immediately. I also excluded theses written in English as they would interfere with text processing and, more importantly, with LDA, affecting word distribution and presumably contributing to some "most common English words" topics, meaningless for the analysis.

Methods

All the file and text processing, data analysis, and visualization are done in Python programming language[3].

Getting Theses Files and Metadata

Starting from 2015, it is obligatory for Higher School of Economics' student to upload their term papers and thesis to LMS (Learning Management System), so that their work can be checked for plagiarism and added to an internal database. Metadata for the thesis can be accessed through the HSE website[4]. While looking into the structure of the site, I assumed that the search was implemented by sending a query to an external source, getting all the entries and only then rendering them in HTML and soon found out that I was right, and the internal database (in fact it might be any type of data storage system, but for the simplicity, I use term 'database') has an API, that I can use to send my query and get all the metadata for all found entries, including links to download theses files through LMS. At the time of research, this required sending data via the 'POST' method with the right headers and a certain payload to an API URL[5], which does not seem to be working anymore.

[2] https://www.hse.ru/edu/vkr/
[3] Code is available at https://github.com/polyankaglade/Theses_LDA (Jupyter Notebooks).
[4] https://www.hse.ru/edu/vkr/
[5] https://www.hse.ru/edu/vkr/api/list

This might seem to be possible via the 'GET' method with the same parameters as in the search URL another API URL[6], but response contains only the author's name, title, supervisor's information, program title, and faculty (also rating if available), no links or ids for the file download.

I used *requests* package for working with HSE Thesis API and downloading files from LMS[7] into the right file formats (*.pdf*, *.doc*, and *.docx*).

Processing Files

The next step of the data preparation required extracting textual data from the downloaded files. I used *docx*, *textract* and *pdfminer* packages, but faced some difficulties:

1. *.doc* files could not be processed on my machine at all, as some of *textract* package's components (namely, *antiword*) do not work on my OS.
2. *pdfminer* package was very inconsistent in extracting Cyrillic texts and, if worked properly, extracted footnotes. It was also very unintuitive to use.
3. Footnotes of most PDF files contained the whole bibliography in various formats (or no clear format at all), making it almost impossible to remove them. I could not leave such references with authors names and full titles, containing all sorts of Russian and foreign words, since they would affect words distribution, same as English texts mentioned in Data section.

These circumstances lead to being able to use only *.docx* files. I extracted texts via *docx* package, which has proven itself to be a reliable tool, and saved them to *.txt* files.

Preprocessing Texts

At this stage, *.txt* files contained all content from the original file (excluding footnotes, figures, and images), so it was necessary to extract only the main body of the text by leaving out everything before 'Introduction' and after 'References'. My pipeline here was as follows:

1. Compile one regular expression to match most of the 'Introduction' variations and about seven regular expressions for 'References / Bibliography', since they were significantly less uniform.
2. Make a function to get all the matched instances for each case, select only the last one, and return its start or end position.
3. Run algorithm and evaluate results visually.

[6] https://www.hse.ru/n/vkr/api/?faculty=139191145&year=2020&text_available=yes
[7] http://lms.hse.ru/ap_service.php?getwork=1&guid={id}

4. Make some manual changes to files to ensure that every file is processed correctly. For example, replace `Вступление` for `Введение`, replace `ЛИТЕРАТУРА` for `Библиография`, add or delete newlines
5. Repeat steps 3-4 until every case is captured.
6. Save only the text between the end of 'Introduction' and the start of 'References'

I also plotted the length of the text before introduction and after references in relation to the whole length of a document (Fig. 3), and manually reviewed cases, that showed unintuitive results, for example, theses that had less than 60% of the text before the bibliography. However, they all turned out to be processed correctly, just had an impressively long list of cited works.

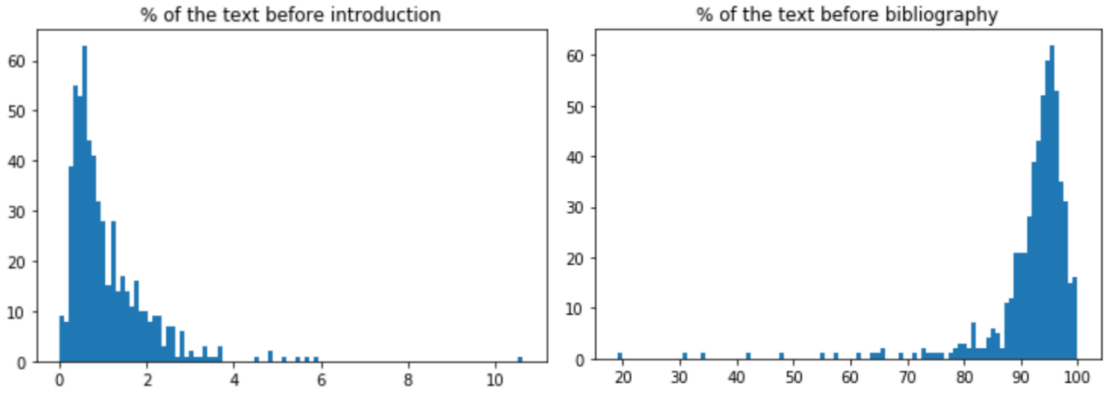


Figure 3. Percent of the text's length before introduction and after references, respectively.

Processing Texts

The purpose of this data manipulation part is twofold: to increase the semantic fullness of texts and reduce the volume of the corpus. It can be achieved by removing as much noise (punctuation, numbers), "general" words with no distinct meaning (stop words) and reducing the variance of similar words. The steps in my processing pipeline are:

0. Delete inline references, matched by a special regex pattern. This approach does not produce a perfect result because of the diversity of citation formats (or the nonexistence of such).
1. Tokenize texts via *razdel* package.
2. Delete any non-alphabetic characters.
4. Lemmatize and POS-tag tokens via *pymorhy2* package. POS-tags are then used for working with n-grams.
5. Delete stop words from Russian, English, and German obtained from *NLTK* package, and some corpora specific words such as *'author', 'article', 'work'*. Looking back on the final

list, it would have been useful to include French stop words as well, although it probably would not make much difference in terms of general word distribution.

6. Join frequent bi- and trigrams into one token, for example, *'russian_language', 'machine_learning', 'soviet_union'*. I used tools from *NLTK.collocations* to obtain n-grams, set a dynamic minimal frequency threshold (0.8 quantile of lemmas frequency for a given document), and also filtered them by POS (where 'None' means a non-Russian word in *pymorhy2*'s output):
    a. bigrams: Noun/Adjective/None + Noun/None
    b. trigrams: Noun/Adjective/None + Noun/Adjective/None + Noun/None

Latent Dirichlet Allocation

In this work, I settled on *Gensim* implementation of the LDA algorithm (Rehurek, R. and Sojka, P., 2010), and chose LdaMulticore for faster training.

Dictionary and Corpus

LDA model requires two objects to be trained. The first one is a Dictionary, which holds [word – id] pairs ("token – id" to be more precise), counts term and document – term frequencies and can be filtered by these two measures. After being applied to the full corpus, it contained 187,520 wordforms (unique tokens), with 66,177 having term frequency equal to 1. To reduce the number of tokens which would be used for training and exclude too rare and too common words I applied *filter_extremes()* function, filtering out tokens, that appear in less than 5 documents or in more than 50% of documents. These parameters proved to be optimal to get meaningful and mostly interpretable topics later.

Second main object is Corpus, which holds [word – measure] representations for each document. Experiments showed that using BOW representations lead to better LDA outcome, compared to Tf-Idf representations. While using the latter one, almost all topics had a surname as the word with the highest probability and were not interpretable in general.

Model

LDA model has many parameters, but it can be hard to find ones most suitable for the given corpora, because there are few reliable measures for evaluating model quality, so most assertion is still done manually (or rather, visually). For my model, I experimented with three parameters that

affect the quality of the outcome: number of topics (num_topics) and two affecting the "concentration parameters for the Dirichlet distribution"[8] (alpha, eta).

After training about 100 models with different combinations of mentioned parameters and comparing their coherence (u_mass and c_v), I decided to set alpha equal to 'asymmetric' and eta to 'auto'. There were two values of num_topics, which resulted in higher coherence, ~25 and ~250. On the one hand, 25 topics lead to somewhat good visualization, but very low topic interpretability. On the other, 250 topics lead to quite convoluted visualization but very distinct and clear topics (as far as I can judge ones that are outside my field of expertise). Further trials showed that setting the number of topics to 50 was an optimal decision and resulted in both appropriate topics and reasonable data generalization (both visually and statistically).

Other training parameters were as following:

- passes = 4,
- workers = 3 (on a machine with 4 cores),
- chunksize = 1000 (whole corpus is less than 80 Mb in *.txt* format),
- eval_every = False (to speed up training),
- random_state = 42 (for somewhat replicable results)

I strongly believe that, while increasing the number of topics to 100-150 would presumably contribute to higher coherence, model's quality would increase insignificantly and such large number of topics would be much more difficult to analyze and interpret manually.

Dimensions and Visualization[9]

To sum up and refresh that have been said in the previous sections, LDA is "generative probabilistic model" (Blei, Ng, & Jordan, 2003) which is trained over a corpus of texts represented in BOW format, and for each document it returns the probability of each topic. As discussed above, a topic is a list of words with numbers, representing words' weights in a particular topic. Thus, I use resulting topic distributions as a kind of vector representations for the texts, allowing me to locate and compare them in a multidimensional space.

Resulting vectors are 50-dimensional, which makes them suitable for many types of mathematical analysis, but not very convenient for visualization and thorough manual analysis. A solution is to reduce the dimensionality to 2 or 3 to be able to plot them on a plane or in 3D space.

[8] https://stats.stackexchange.com/questions/37405/natural-interpretation-for-lda-hyperparameters
[9] Interactive plots are available at https://polyankaglade.github.io/Theses_LDA/.

For the first attempt in dimensionality reduction, I chose to apply t-SNE with 2 and 3 components and used them as X and Y (and Z) axes (Fig. 4). Although there seems to be some visible clusters, subsequent manual analysis reveals that these groups have little to none in common, except for two large clusters of Linguistics and Philology. Independently of the perplexity value, this visualization did not provide a comprehensive representation of the data. Theses' locations were mostly random and impossible to interpret.

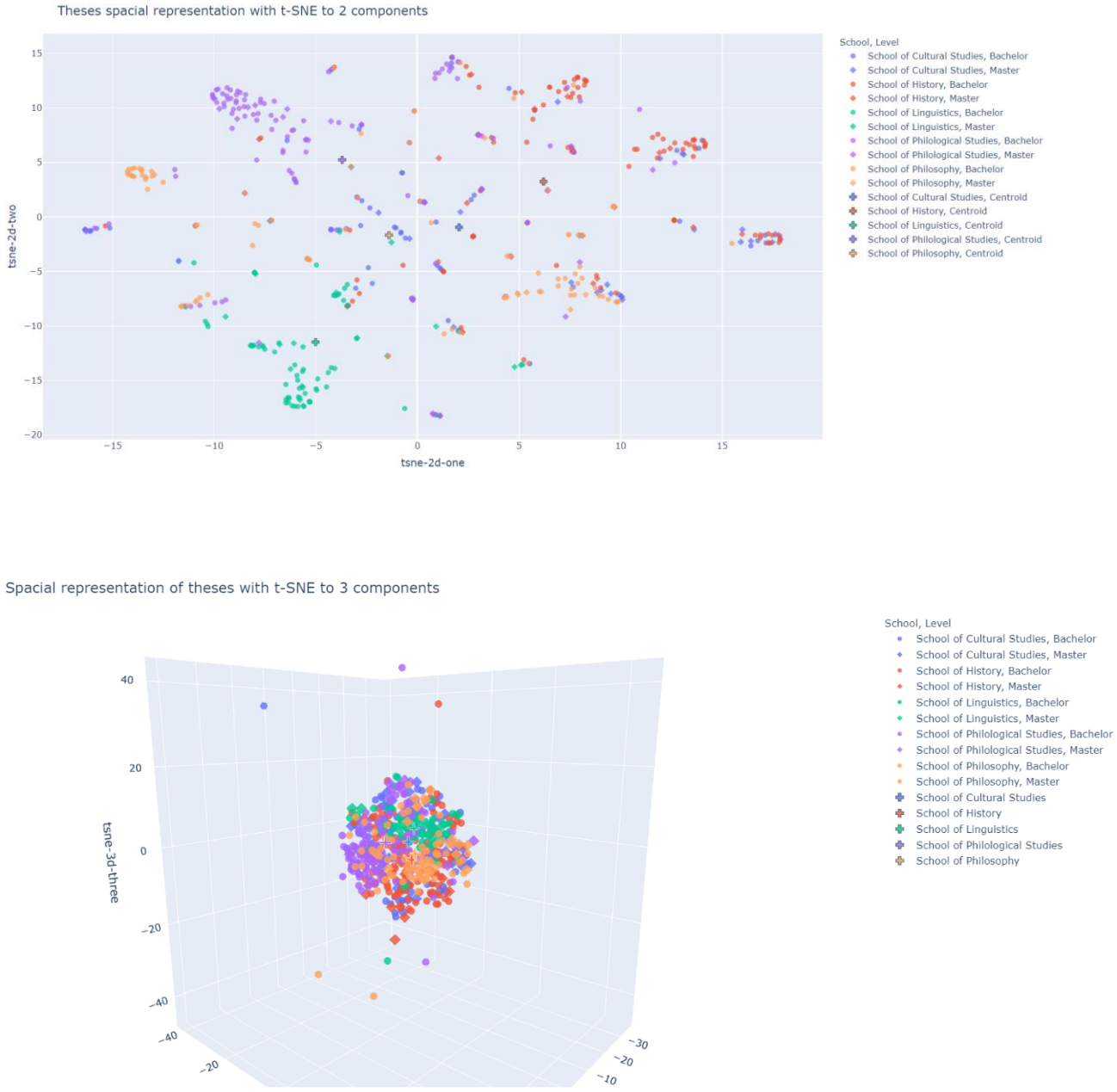


Figure 4. 2D and 3D visualization with only t-SNE.

For the second attempt I tried applying PCA and t-SNE algorithms successively, hence it is one of the most common approaches. I ran PCA with 10 components and then t-SNE with 2 components (with perplexity equal to 90) and used them as X and Y axes (Fig. 5). The result was also poor, but slightly better. For the task of comparing Bachelor to Masters works the representation was also unsuitable.

Figure 5. Visualization with PCA + t-SNE, theses grouped by School and Level

The final and most successful visualization was done by applying only PCA with 5 components and using the first three on them for 3D visualization (Fig. 6) and all possible combinations of them for a 5D → 2D type of visualization (Fig. 7).

Figure 6. Theses in 3D space

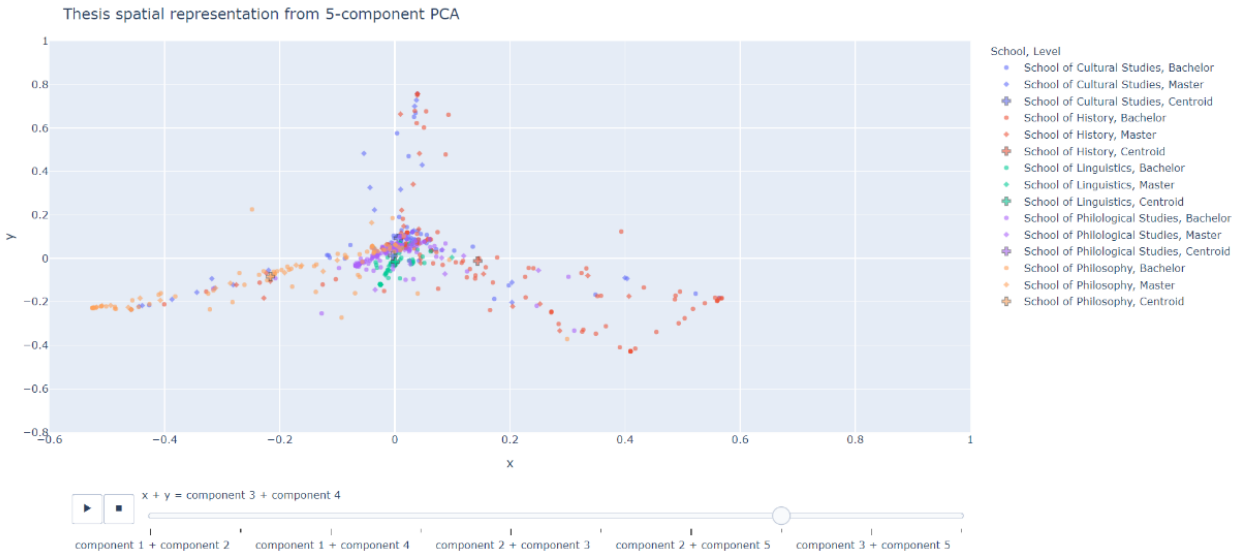


Figure 7. 2D representation of 3 and 4 components.

Analysis

Here I provide the methods and measures I used to study the data both manually/visually and quantitatively. For later I used mathematical functions mainly from *numpy* and *scipy* packages. All data points were plotted via *plotly* package, distribution plots – via *seaborn*.

Qualitative

To better understand the data, I conducted in-depth manual analysis of the resulting 3D space. Several its features are quite prominent even at the first glance. First of all, we can distinctively see green, purple and orange groups, almost creating lines (Fig. 8). Red one is less visible, but still

recognizable. If we take a closer look at the titles of the works, constituting each so-called line, it is possible to define their general theme. Thus, green line will be Linguistics Line, purple – Philology (or Literature) Line, orange – Philosophy Line, red – History Line. Cultural Studies Theses seem to be scattered a lot, mostly around red line and the center of the shape.

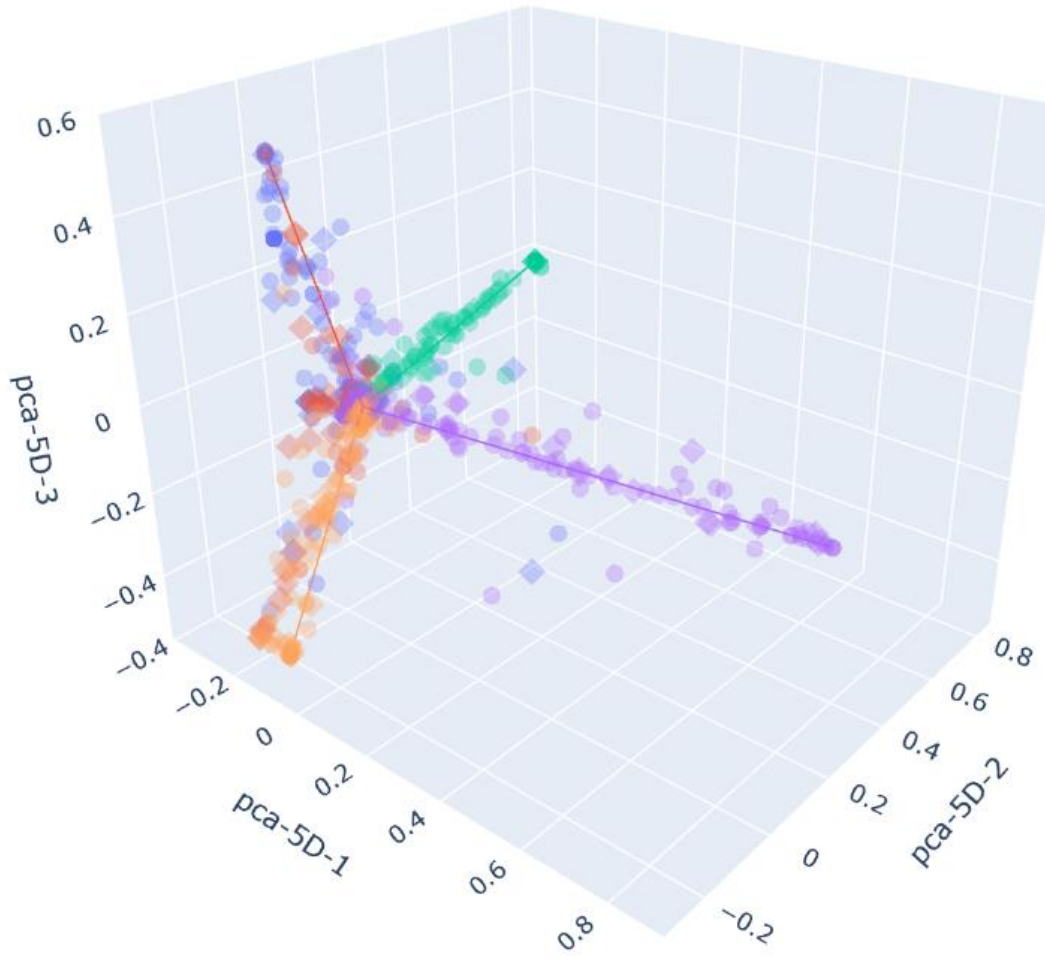


Figure 8. 3D with approximated central lines.

It is also possible to imagine dividing this space by two planes, almost perpendicular to each other. The first one contains Linguistics and Philology Lines, second – Philosophy and History Lines. They can be interpreted as Language and Non-Language Planes.

Some points are located significantly further from the central lines of their color and it can be interpreted through looking at this works' titles. I also provide abstracts, with phrases I marked as contributing to one or other scientific area. It should be noticed than I have little expertise in the mentioned fields and approach such a task as a naive reader. Points, marked in Figure 9 are:

1. “Students, Scientists and Tricksters in the Works of Geoffrey Chaucer: Literary Traditions and Historical Context” (Medieval Studies, School of History)

> The work is devoted to the traits of the trickster archetype that can be found in some of the characters of Geoffrey Chaucer's "Canterbury Tales". The auther attempts to identify the reason of said traits' existance in those characters through literary traditions and historical context surrounding the work.

2. “Generational Synthesis Principle and the Problem of Reflections on Serfdom among Russian Peasantry after 1861” (History, School of History)

> This research is devoted to the problem of what it was like to be a serf before the period of Great Reforms, how this biographical and traumatic experience has formed the newly freed society, it’s identity, views of life and self-understanding. Based on biographical sources such as memoires and autobiographies, that were created by ex-serfs, this research should provide it’s auditory with the very special “peasant narrative” in Russian history in the nineteenth century, which is, unfortunately, so unknown, unpopular and yet undiscovered among the Russian mass-audience and scientific community. The understanding of this narrative with all it’s complexity and details will provide a clearer view on the generations, that were born as a free folk after 1861.

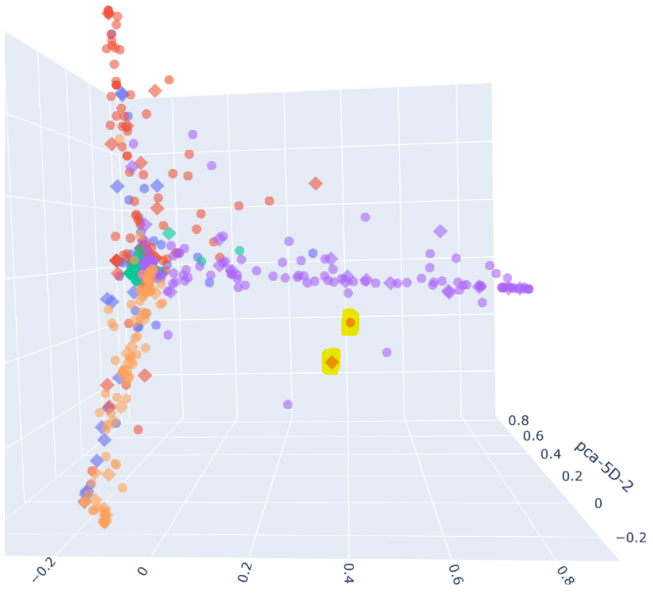


Figure 9. Two History works, located between Philosophy and Philology Lines

More examples of such works, showing the interdisciplinarity of their School would be:

1. “L.N. Tolstoy and Russian Intelligentsia: a Philosophical Anallysis of Historical Conflicts” from School of Philosophy, but located very close to History Line (Fig. 10a, Fig. 10b)

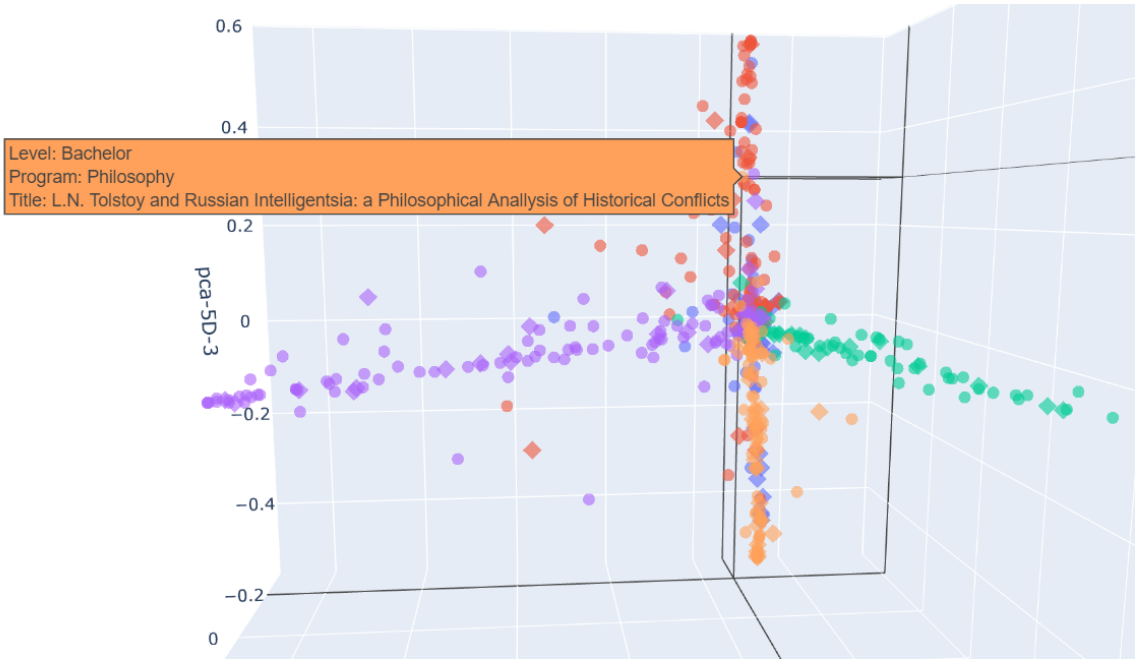


Figure 10a. View 1 of the first thesis

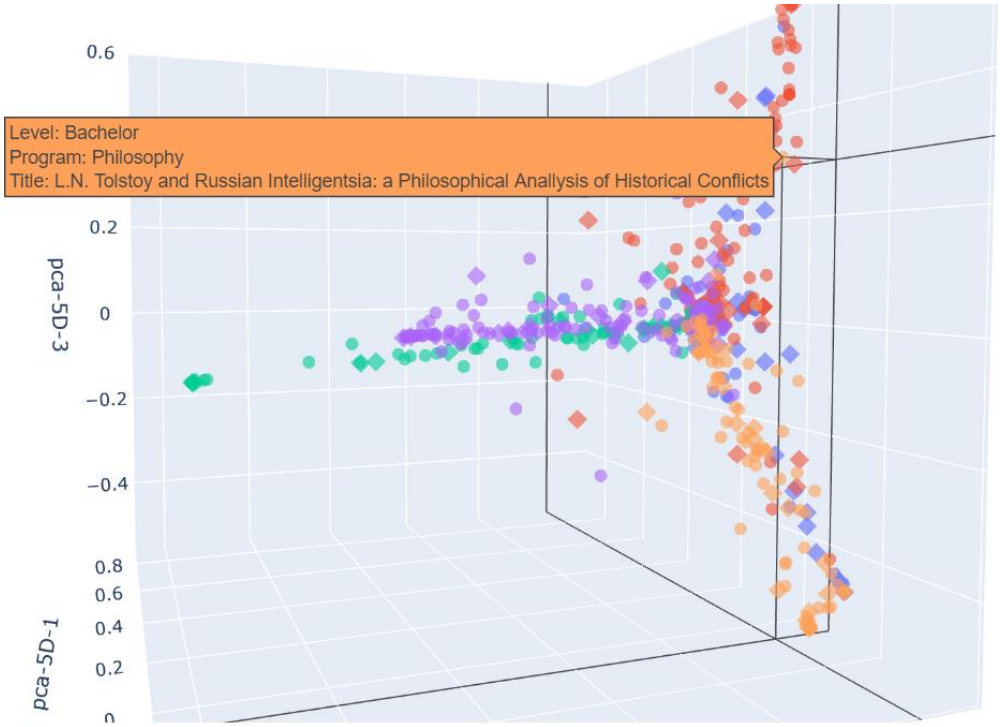


Figure 10b. View 2 of the first thesis

2. “Linguistic Representation of Speakers' Values in Modern Russian Discouse” from School of linguistics, but located between Linguistics, History and Philology Lines (Fig. 11a, Fig. 11b)

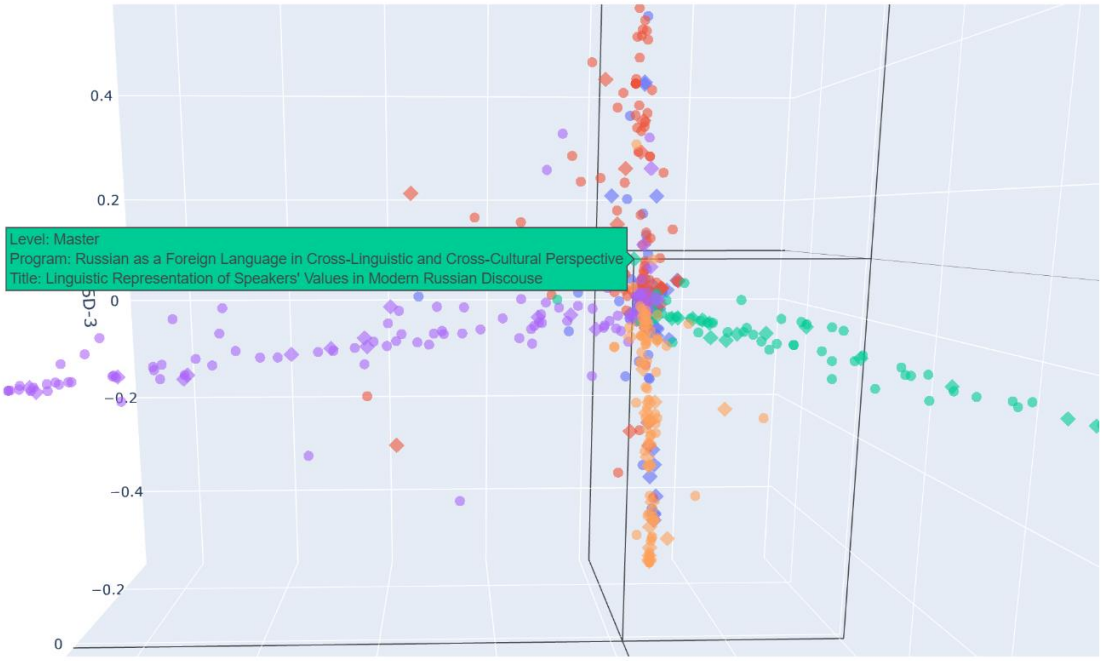


Figure 11a. View 1 of the second thesis

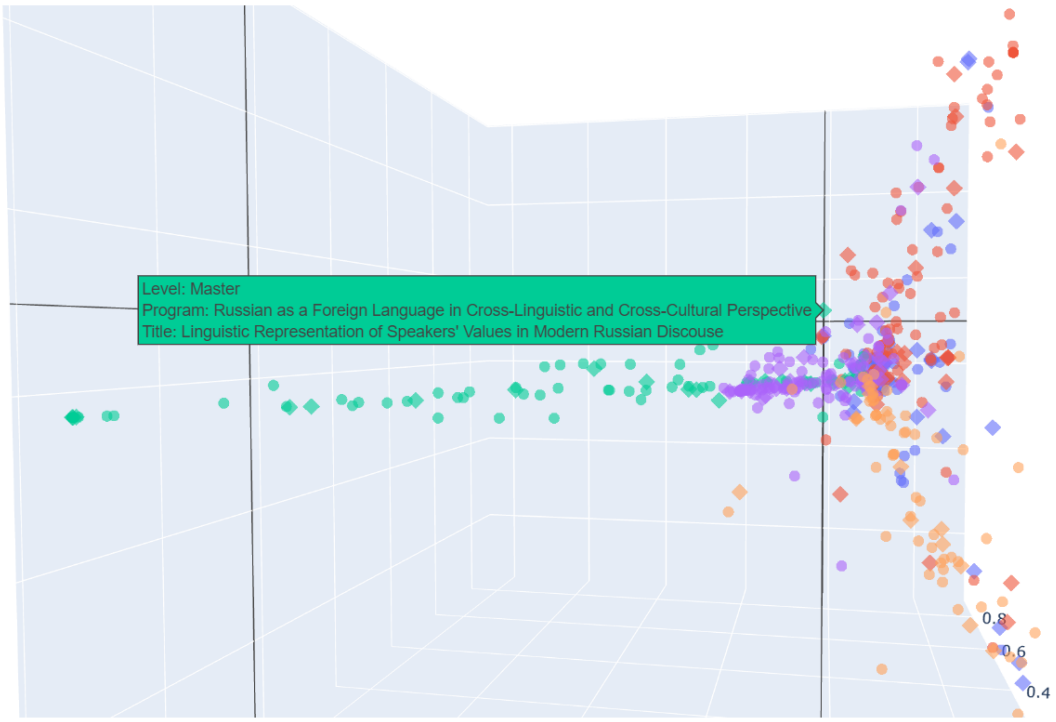


Figure 11b. View 2 of the second thesis

Another important conclusion can be made from the way the data is distributed in the 3D space. Such prominent lines indicate that there is a strong correlation between a class and some features in the vector, meaning a strong correlation between a School and certain topic. To capture this correlation, I fitted a Logistic Regression model on vectors and Schools and was able to juxtapose weights of the topics with their top 5 words (Fig. 12):

```
School of Cultural Studies
    2.5939254083769234  king, newspaper, tolstoy, historian, government
    2.055191156701277   soviet, ussr, military, sport, document
    1.8250179225713132  rubens, artist, perform, portrait, mausoleum
    1.7443224485903976  artist, depict, master, composition, italian
    1.5628192147774067  temple, polish, architect, brand, prince
    1.2909069980251406  museum, artist, exhibition, visitor, spectator
    1.2005788395136283  god, prince, saint, constantine, Socrates
    0.842710507922798   god, rilke, mall, angel, urn
    0.6727720629758619  artist, building, music, photography, fragment
    0.6209975021368711  bulgakov, record, novel, dostoevsky, daughter
----------
School of History
    2.0790912426402004  film, spectator, director, performance, theater
    1.4791515271619156  flashmob, community, respondent, interview, mosk
    1.459256490754906   museum, artist, exhibition, visitor, spectator
    1.288933579733873   subject, being, serf, individual, philosophy
    1.1422777790051437  film, scene, romance, kim, character
    1.0483772604727568  artist, viewer, james, turgenev, athlete
    0.9751584569185027  artist, building, music, photography, fragment
    0.9168403291434661  vowel, baroque, mikhailov, music, agent
    0.8994251837450844  baratynsky, aksakov, theater, performance, spectator
    0.7657869073286515  shaman, village, migrant, tribe, resident
----------
School of Linguistics
    4.5797990292789     verb, construction, russian_language, student, experiment
    2.698619364173861   corpus, algorithm, parameter, markup, training
    1.7069751950989493  verb, lexeme, frame, used, idiom
    1.2564008093738928  gesture, russian_language, student, informant, language
    1.1476902796131687  adverb, construction, felix, speaker, noun
    1.0909678278953139  particle, block, kuzmina, cycle, verb
    0.8719651279420416  composite, corpus, noun, collocation, component
    0.659525346800817   festival, cluster, burning, film, tragedy
    0.6435330935440234  judgment, witgenstein, interview, university, network
    0.44717760760168923     ontology, pisemsky, peirce, graph, invention


School of Philological Studies
    4.214722555517609   poet, poem, novel, writer, character
    1.918018414437308   romance, theater, character, heroine, del
    1.2252242678719127  novel, petersburg, white, storyteller, elle
    1.052802943520902   gypsy, moscow, poem, dostoevsky, writer
    1.0433600895433097  bulgakov, record, novel, dostoevsky, daughter
    0.8235255216539572  god, prince, saint, constantine, socrates
    0.7771675506818128  novel, adaptation, clitic, character, performance
    0.6312565260405699  gesture, russian_language, student, informant, language
    0.6012701180659712  poem, poetry, poetess, niva, collection
    0.5605436886525359  film, zone, tarkovsky, script, story
----------
School of Philosophy
    3.82530670570969    philosophy, philosopher, phenomenology, god, philosophical
    2.0522356134870794  subject, being, serf, individual, philosophy
    1.6903110792650833  kant, sublime, reason, metaphysician, knowledge
    1.3749039986547562  metaphor, speaking, romance, religion, noise
    1.1869145865118285  nietzsche, heidegger, truth, bacon, sculptor
    0.8931199797589219  russo, schmitt, thinker, philosopher, ibn
    0.8201097986064299  los, franc, being, god, philosophy
    0.6701065316866388  kropotkin, yell, capitalism, aristotle, revolution
    0.4586465558022146  artist, yoga, yogi, liberation, even
    0.4205409750517246  bradbury, strauss, writer, self, hobbes
```

Figure 12. Top 10 topics per School with top 5 words per topic.

Quantitative

For the quantitative analysis I treat texts' topical representations as vectors and PCA components' values as texts coordinates in 3D space, visualized above. For each School I compute a

Centroid vector in 50D space (as the mean vector of the group) and a Centroid point in 3D space (as the mean coordinates, (Fig. 13)). For each thesis I then compute:

1) Euclidean distance to Centroids,
2) Manhattan (or Cityblock) distance to Centroids,
3) Cosine distance to Centroids.

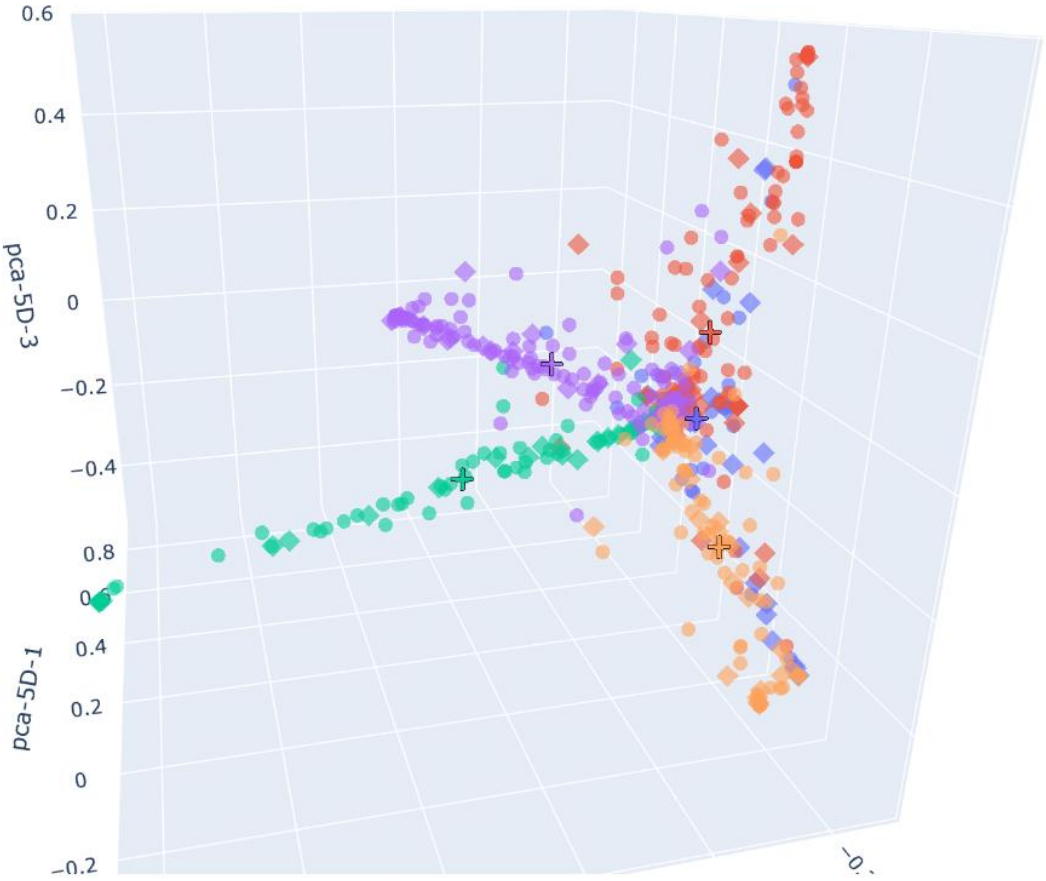
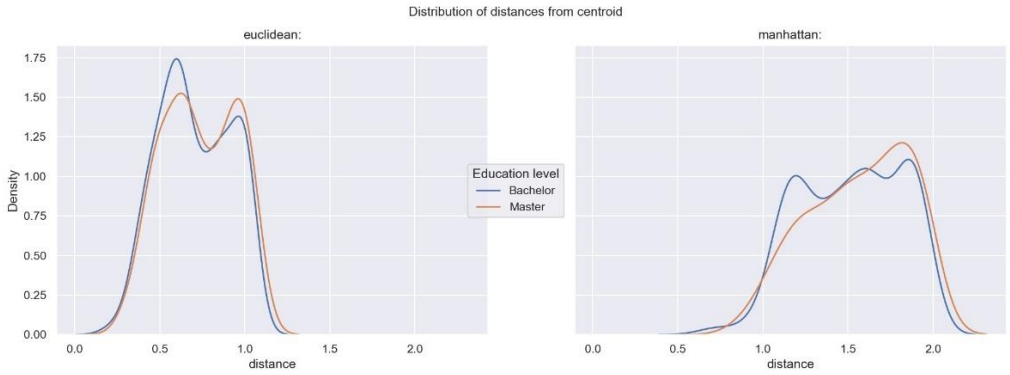


Figure 13. Centroids in 3D space

It should be noted that Lines from the Fig. 8 are not approximations of any sort, they are just lines from the center of the shape to the furthest point in the class. I used them while experimenting with possible measures (in this case – distance from a point to a "central" line) of its class). A more reasonable approach would be to use lines that go through the centroid, but they do not provide for a visually pleasing image.

The next step is to explore described above measures in relation to Schools, Programs and Levels of education. Firstly, mean values and distributions of all the measures show no significant differences in relation to the Level of education, as can be seen in Fig. 14:



Level	
Master	0.735165
Bachelor	0.707921

Level	
Master	1.564588
Bachelor	1.516337

Figure 14. Distances to the Centroid vector vs Level

However, if compared for every School individually, it can be noticed, that Master works from School of Philology tend to be further from the Centroid vector than Bachelor works (Fig. 15):

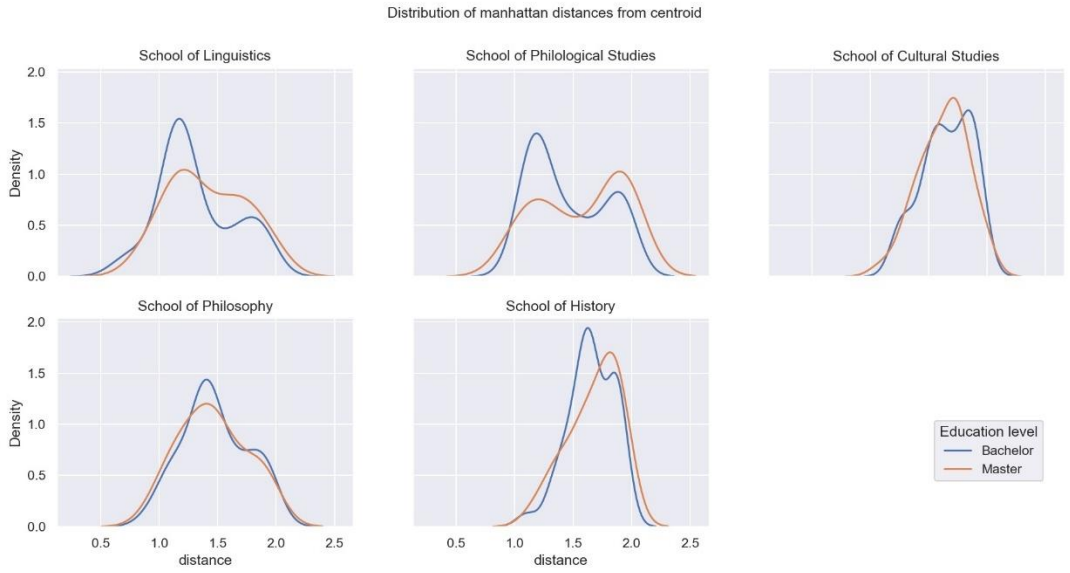


Figure 15. Manhattan distances from the Centroid vector.

The explanation for that can be seen in Table 1. Programs with the largest mean distance are *Cultural and Intellectual History: Between East and West* and *Language Policy in the Context of Ethnocultural Diversity*, both Mater programs affiliated with School of Philology:

Level	School	Program	mean_eu	mean_manh	mean_cos
Master	School of Philological Studies	Cultural and Intellectual History: Between Eas...	1.045577	1.958092	0.966634
Master	School of Philological Studies	Language Policy in the Context of Ethnocultura...	0.867468	1.871222	0.923424
Bachelor	School of History	History of Arts	0.908440	1.803367	0.797697
Master	School of History	Medieval Studies	0.774296	1.744502	0.817227
Master	School of History	History of Artistic Culture and the Art Market	0.827604	1.736141	0.770253
Bachelor	School of Cultural Studies	Cultural Studies	0.737796	1.654699	0.679956
Master	School of Cultural Studies	Applied Cultural Studies	0.725004	1.630560	0.684335
Master	School of Cultural Studies	Visual Culture	0.753487	1.622218	0.630326
Bachelor	School of History	History	0.726447	1.604963	0.576930
Master	School of History	Historical Knowledge	0.719664	1.578078	0.565792
Master	School of History	History of Knowledge and Social History	0.724772	1.572979	0.682198
Master	School of Philosophy	Philosophical Anthropology	0.737951	1.506769	0.537864
Master	School of Linguistics	Language Theory and Computational Linguistics	0.760072	1.484554	0.603018
Bachelor	School of Philosophy	Philosophy	0.691588	1.477208	0.546577
Master	School of Philological Studies	Comparative Studies: Russian Literature in Cro...	0.690031	1.473639	0.511735
Bachelor	School of Philological Studies	Philology	0.674639	1.453105	0.491407
Master	School of Linguistics	Computational Linguistics	0.698316	1.404707	0.521863
Master	School of Philological Studies	Russian and Comparative Literature	0.612970	1.390477	0.387374
Master	School of Philosophy	Philosophy and Religious Studies	0.641267	1.369237	0.466711
Master	School of Philosophy	Philosophy and History of Religion	0.685898	1.346549	0.431413
Master	School of Linguistics	Linguistic Theory and Language Description	0.614943	1.325430	0.560560
Bachelor	School of Linguistics	Fundamental and Computational Linguistics	0.651519	1.312485	0.415364
Master	School of Linguistics	Russian as a Foreign Language in Cross-Linguis...	0.511419	1.253240	0.298244

Table 1. Programs, sorted by mean Manhattan distance.

Mean Manhattan, Euclidean and Cosine distances all show the same tendencies (Table. 2), especially for the top and the bottom ends of this scale.

	mean_eu	mean_manh	mean_cos
mean_eu	1.000000	0.917252	0.902283
mean_manh	0.917252	1.000000	0.952446
mean_cos	0.902283	0.952446	1.000000

Table 2. Correlation between mean measures for Programs

Top three Programs with the **highest** mean distances are:

1. Cultural and Intellectual History: Between East and West
2. Language Policy in the Context of Ethnocultural Diversity
3. History of Arts

Top three Programs with the **lowest** mean distances are:

1. Fundamental and Computational Linguistics
2. Russian as a Foreign Language in Cross-Linguistic and Cross-Cultural Perspective
3. Philosophy and Religious Studies

It also seems like there is some correlation between the School and the position in Table 1. It can be verified by providing same analysis for each School, see Fig. 16, Table 3 and Table 4:

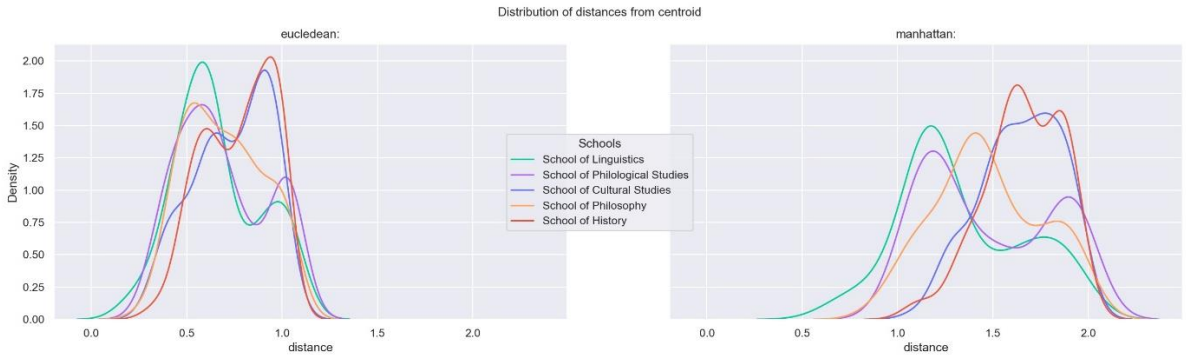


Figure 16. Distances to the Centroid vector by School

School	mean_eu	mean_manh	mean_cos
School of Cultural Studies	0.739132	1.644561	0.670475
School of History	0.771551	1.657841	0.647778
School of Philosophy	0.695610	1.473173	0.538748
School of Philological Studies	0.687201	1.476848	0.512658
School of Linguistics	0.661373	1.336168	0.441604

Table 3. Programs sorted by mean Cosine distance

	mean_eu	mean_manh	mean_cos
mean_eu	1.000000	0.961217	0.936966
mean_manh	0.961217	1.000000	0.985188
mean_cos	0.936966	0.985188	1.000000

Table 4. Correlation between mean measures for Schools

Analysis of the same measures but in 3D space yielded all the same results.

Besides working with single measures, I also explored the idea of a combined measure. The intuition behind this is that distance and cosine represent different types of relations to the centroid. For example, both marked dots in Fig. 17 have the same distance to the centroid, but the cosine of the left one is much larger than the cosine of the right one. Left thesis shows more interdisciplinarity.

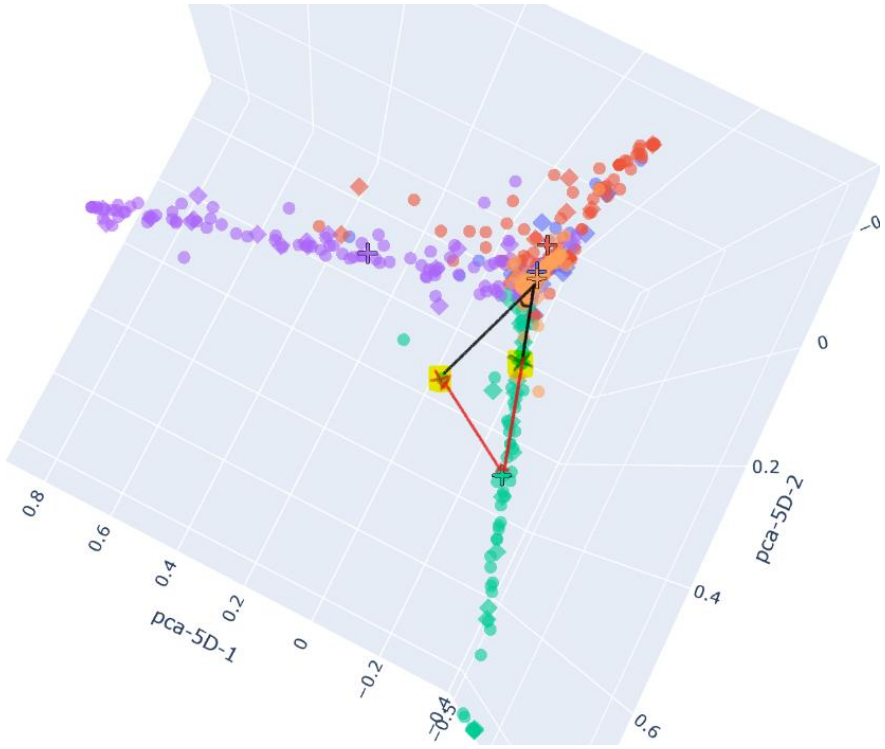


Figure 17. Distance in red and vectors + angle in black

However, the cosine of the lowest green dot is also very small, while the distance to the centroid is larger than the distance of both marked points, but this work shows the least amount of interdisciplinarity. It means that in order to capture theses with the most interdisciplinarity, we need to maximize both cosine and distance. One way to do that is to fit a curve on the cosine and distance data and use the resulting function as the most significant interdisciplinarity measure (Fig. 18).

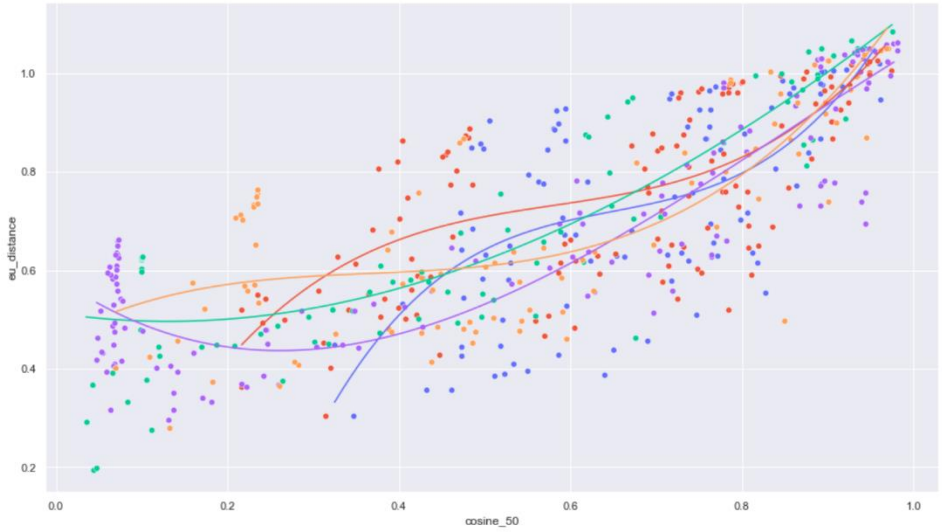


Figure 18. Curves, fitted on each School's distance and cosine measures with objective of cubic equation.

This combined measure produces results, very similar to my qualitative analysis (Fig. 19).

```
School of Cultural Studies
        Poetics of the Social Criticism in the Russian Rock- and Rap-Poetry
        Literature as an Alternative Writing Experience: the Case of Fedor Stepun
        London Text of Russian Literature in the Second Half of the XIXth century
        The Influence of Socio-cultural Environment on the Emergence and Development of Certain Kinds of Sport (Illustrated by The Olypmic and Paralympic
        History Teaching in Soviet Schools (1917-1958), the case of the Tatar ASSR
School of History
        Students, Scientists and Tricksters in the Works of Geoffrey Chaucer: Literary Traditions and Historical Context
        Generational Synthesis Principle and the Problem of Reflections on Serfdom among Russian Peasantry after 1861
        Savoraim in Mesopotamia and Midrash's Penetration into the Aramaic Exegetics
        Personification of Death in "The Husbandman and Death" by Johannes von Tepl
        Practices of Judical Settlement of Conflicts Between Nobles and Serfs in the 1760s–1770s in Russia
School of Linguistics
        Developing Linguistic Ontology for the Intellectual Property of Pharmaceutical Drugs
        The Distribution of the Variants of the Preposition a/ab in Classical Latin
        Linguistic Features of Manipulation in Case of Gricean Maximas Violation in Modern Political Discourse
        Поле глаголов эмоций русского жестового языка в типологическом освещении
        Language Situation in the Republic of Karelia: Current Ethnolinguistic Dynamics
School of Philological Studies
        The Concept of "World Literature" in Soviet Culture of the Thirties
        The Olfactory and Auditory Characteristics of the City in the Works of James Joyce and John Dos Passos
        Poetics of Space in W. Faulkner's Short Stories
        The Transformation of Fictional Space in A. Huxley's Novels "Brave New World", "Ape and Essence", "Island"
        Dynamics of Ethnolinguistic Identity of the Russian-Speaking Minority in Latvia
School of Philosophy
        L.N. Tolstoy and Russian Intelligentsia: a Philosophical Anallysis of Historical Conflicts
        The Ethical Problems of the Tranformaion of War
        Criticism and Development of D. Davidson's Theory of Meaning in Contemporary Analytical Philosophy
        The Theory of Truth and the Theory of Sense in "Tractatus Logico-Philosophicus"
        The concept of State in the Context of Eurasianism – the Russian Philosophical  Thought of the XX Century
```

Figure 19. Top 5 most interdisciplinary theses per School according to the combined measure.

Conclusions

The above analysis leads to several conclusions.

Firstly, Topic Modeling Approach, proposed in this work, does indeed capture the interdisciplinarity of academic texts, written by students from HSE Faculty of Humanities. Following the proposed pipeline for the analysis, one can apply methods of Topic modeling to completely different corpora or to the same type of data, but on a larger scale.

Secondly, the best measure of interdisciplinarity is yet to be discovered, but for now, distance and cosine distance to the centroid both produce decent results, while also capturing different aspects of interdisciplinarity presented in a given text. Besides, a well-fitted curve from those measures, in my opinion, has a lot of potential in this field of research.

Thirdly, only the second hypothesis was proven to be true by arranging all Programs on an interdisciplinarity scale. No significant differences between Bachelor and Master theses were found. As an additional conclusion, Schools themselves are arranged on a scale from Linguistics to Cultural Studies, with latter combining the most topic or theme variation.

Last but not least, now there is a corpus of processed academic texts in Russian, written by bachelor and master's students, as well as a pipeline for its processing. Clean and formatted academic corpora of this volume (over 8 million tokens) can be useful for solving different NLP problems and for transferring existing instruments into scientific domain and/or Russian language.

My approach can be further improved by implementing other NLP methods, such as text2vec models, and including their outcomes into the comparison.

I hope this study will provide the research and educational community with a better understanding of the scientific structure of Humanities Faculty of HSE University

References

Älgå, A., Eriksson, O., & Nordberg, M. (2020, 11). Analysis of Scientific Publications During the Early Phase of the COVID-19 Pandemic: Topic Modeling Study. *Journal of Medical Internet Research, 22,* e21559. doi:10.2196/21559

Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications, 6*. doi:10.14569/ijacsa.2015.060121

Asmussen, C. B., & Møller, C. (2019, 10). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data, 6*. doi:10.1186/s40537-019-0255-7

Bakarov, A., Kutuzov, A., & Nikishina, I. (2018). Russian Computational Linguistics: Topical Structure in 2007-2017 Conference Papers. doi:10.13140/RG.2.2.28391.75687

Belousov, K. I., Baranov, D. A., & Erofeeva, E. A. (2018). Thematic and paradigm models of the concept system of science. *Epistemology & Philosophy of Science, 55*, 184–203. doi:10.5840/eps201855116

Blei, D. M., & Lafferty, J. D. (2007, 6). A correlated topic model of Science. *The Annals of Applied Statistics, 1*. doi:10.1214/07-aoas114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993–1022.

Braam, R. R., Moed, H. F., & van Raan, A. F. (1991, 5). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science, 42*, 233–251. doi:10.1002/(sici)1097-4571(199105)42:4<233::aid-asi1>3.0.co;2-i

Deerwester, S. (1988). Improving information retrieval with latent semantic indexing.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990, 9). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391–407. doi:10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9

Dumais, S. T. (2005, 9). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*, 188–230. doi:10.1002/aris.1440380105

Dutta, B. (2008, 8). Classification of Keywords Selected from Research Articles on Physics and Development of a Quantitative Subject Access Tool. *IFLA World Library & Information Congress (WLIC) 2013 Singapore.*

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP \textquotesingle08.* Association for Computational Linguistics. doi:10.3115/1613715.1613763

Han, X. (2020, 10). Evolution of research topics in LIS between 1996 and 2019: an analysis based on latent Dirichlet allocation topic model. *Scientometrics, 125*, 2561–2595. doi:10.1007/s11192-020-03721-0

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009). Detecting topic evolution in scientific literature. *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM \textquotesingle09.* ACM Press. doi:10.1145/1645953.1646076

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI1999).*

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning, 42*, 177–196. doi:10.1023/A:1007617005950

Lauscher, A., Fabo, P. R., Nanni, F., & Ponzetto, S. P. (2016, 12). Entities as Topic Labels: Combining Entity Linking and Labeled LDA to Improve Topic Interpretability and Evaluability. *Italian Journal of Computational Linguistics, 2*, 67–87. doi:10.4000/ijcol.392

Mahmood, A. (2013). Literature Survey on Topic Modeling.

Nichols, L. G. (2014, 5). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics, 100*, 741–754. doi:10.1007/s11192-014-1319-2

Pandur, M. B., Dobša, J., & Kronegger, L. (2020, 10). Topic Modelling in Social Sciences: Case Study of Web of Science.

Paul, M., & Girju, R. (2009, 9). Topic Modeling of Research Fields: An Interdisciplinary Perspective. *Proceedings of the International Conference RANLP-2009* (pp. 337-342). Borovets: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/R09-1061

Porter, A. L., & Rafols, I. (2009, 4). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics, 81*, 719–745. doi:10.1007/s11192-008-2197-2

Rafols, I., & Meyer, M. (2009, 6). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics, 82*, 263–287. doi:10.1007/s11192-009-0041-y

Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. doi:10.13140/2.1.2393.1847

Small, H. (1973, 7). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*, 265–269. doi:10.1002/asi.4630240406

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., . . . Börner, K. (2011, 1). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics, 5*, 14–26. doi:10.1016/j.joi.2010.06.004

Zhu, M., Zhang, X., & Wang, H. (2016). A LDA Based Model for Topic Evolution: Evidence from Information Science Journals. *Proceedings of 2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016).* Atlantis Press. doi:10.2991/msota-16.2016.12