# Frequency and Morphological Patterns of Recognition and Thematic Classification of Essay and Full Text Scientific Publications*

**Vladimir Fomin**[1] fomin@mail.ru
**Alexsander Osochkin**[1] osa585848@bk.ru
**Yulia Zhuk**[2] zhuk_yua@mail.ru

[1] Herzen State Pedagogical University of Russia
[2] Saint-Petersburg State Forest Technical University
Saint Petersburg, Russian Federation



**Abstract**

This article discusses the author's classification method for academic documents, scientific materials and articles published in Russian language. The research aims to utilize a maximum possible extent of the educational environment information potential in modern conditions. The task is to develop the tools for data mining of scientific materials allowing to generate a brief description of a particular scientific article, its table of content, as well as a field of science it belongs to. For the purpose of this exercise we selected a scientific repository containing textbooks with commentaries related to ten fields of science, including History, Chemistry, Law, Biology, Medicine, Physics, Philosophy and Economics. The main feature of the proposed classification method involves a use of a minimal theoretical and linguistical set of indicators together with hierarchical algorithm of classification (based on a regression decision trees method), both aimed to identify stable statistical rules of text classification. The procedure for text data transformation into a set of relative indicators is based on a frequency-morphological analysis, which allows to retain unique stylistic features of the texts involved.

There are two experiments presented. In the first experiment it was possible to successfully define some consistent patterns (logical rules and meaningful indicators) of classification allowing to differentiate between a proper publication (a textbook, monography, etc) and a short commentary/annotation, using only relative indicators.

The results of the second experiment confirm the existence of unique thematic and stylistic features common for a specific field of science.

**Keywords:** *NLP, text-mining, nature language, text classification, decision tree, frequency-morphological analysis, classification by science fields, hierarchical classification*

# 1 Introduction

The contemporary educational environment is characterized by the informatization of many aspects of teaching and learning activities that allow you to change the intensity and nature of informational processes involved. Intelligent technologies (IIT), applied to individualized teaching activities, are considered as one of the instruments allowing to create an individualized educational environment. This involves realization of personal informational and educational needs, increase in efficiency of intellectual procedures related to acquisition of knowledge, expansion of diversity of educational activities, organization of independent students' work, increase of educational potential and accessibility of information, realization of personal preferences, etc.

Algorithms of IIT are widely used in the extraction of knowledge from semi-structured text information. Such scientific areas as NLP and data mining appeared for the purpose of dealing with huge arrays of information [Ke, 2007]. Text mining is one of the essential technologies allowing to expand data mining possibilities into the fields of extracting knowledge from e-learning resources [Liu, 2018, Osochkin et al., 2018, Boytcheva, 2015]. It enables an intellectualized search and classification of different scientific areas and educational categories, identification of authors and their respective writing styles, etc.

Part of textual classification tasks comes down to a simple single-level classification, without taking onto account any complex hierarchical structure of information. However, nowadays, what's often required in relation to a specific scientific article, is an acquisition of detailed information in regards to the type of the document, its scientific areas, table of content, commentaries, etc. Moreover, the complexity of researched objects increases. Their information structure is dominated by the hierarchies of classes. Hierarchical multilevel classification (hierarchical multilevel classification (HMC)) of [Cerri, 2015] is a special field of knowledge in machine learning allowing to address these issues.

Use of statistical methods of text data analysis demands the solution of a difficult task, i.e. how to transform text data into numerical indicators. The key to solution is to extract the numerical set of indicators which allows to describe unique features of each text. With the development of TF-IDF method [Jones, 2004, Pontiki et al., 2014], this task seemed to be resolved because the method allowed to compile a list of most frequently words used within a given text, which allowed to classify the data. This approach allowed to classify small texts with high precision. However, in case of more complex texts, using the TF-IDF method turned out to be not so effective [Moraes et al., 2013, Dikoveckij et al., 2010]. Therefore, the TF-IDF method is used only as a supplementing tool for semantic analysis and parsing. The main issue is an identification of significant indicators allowing to fully describe specifics of a particular text, utilizing syntactic or morphological analysis.

This article, within a framework of electronic and educational environment, proposes a solution for classification of scientific texts published in Russian that are related to ten fields of knowledge mentioned above. The ways to search for and identify any reference material are also suggested.

## 2 Purpose

The main purpose of this paper is to apply the author's data classification method to scientific and educational text materials in Russian related to a range of science fields and see if this method allows to distinguish between a referential and full publication, as well as to pinpoint a set of specific indicators unique for a certain science field.

The following objectives have been established to achieve this purpose:

• To isolate, digitize and define a priority and importance of frequential and morphological indicators which can allow to derive a scientifical field for a given text.

• To check existence of the distinctive and generalizing latent characteristics of the texts relating to different types of documents.

• To create two-level hierarchical classification that takes into account the document's type and science field it relates to.

## 3 Problem definition

Researches in the natural language processing field are being actively developed by English-speaking authors, however their works considered only specifics of English morphology. Linguists noted that native speakers use special sentence construction due to their own cultural background and oratory skills; all these factors make machine analysis more complicated.

Among the most remarkable researches in algorithm classification can be named the following: "Investigating classification for natural language processing tasks" [Medlock, 2008] and "Thinking Linguistically: A Scientific Approach to Language" [Honda, 2008]. The classification methods description of text data is provided in their works and the set of answers to key questions in the field of natural language processing and text-mining are given. The main text-mining task in overcoming specifics of a natural language is the semantic uncertainty [Galitsky 2017].

To counter the increasing complexity and cost growth associated with text processing technologies (due to development of resource-intensive neurolinguistic analyzers), a tendency emerged for further research of the potential of classical methods of frequency and morphological analysis [Baoxun et al., 2012, Canuto et al., 2018]. Text search methods and algorithms [Meystre, et al., 2008, Liu, 2018] focus their attention on the attempt to use a minimal set of theoretic-linguistic approaches and favour the formal methods of statistical processing of simplified word forms. The application of these methods in the text-mining field is characterized by a number of issues, including [is inherent in [Boytcheva, 2015, Cerri, 2015, Canuto et al., 2018, Wei-Cheng et al., 2017] :

• The semantic problem limits the possibilities of application of the linguistic and morphological analysis by unique specifics and grammar of each language, etc.

• The procedure of transformation of text data into digital format can lead to a loss of unique characteristics and patterns contained in the text.

• Problem of dimension of theoretic-linguistic set of indicators, their calculation methods.

• Existence of the hierarchical classification demanding change of target clusters during decision tree creation.

# 4 Tools

This article utilizes the author's method of the analysis of texts in Russian [Fomin et al., 2016, Osochkin et al., 2018], combining in itself:

- a set of quantitative indices,

- an algorithm of their extraction from the text (based on a frequency analysis of symbols and the morphological analysis of words),

- the rules of determination of the importance of indicators and procedures of classification by a method of trees of decisions (regression trees).

To solve the problems of text classification, a special software FaM was developed [Fomin et al., 2016], allowing to transform a semi structured text data into a set of indicators. A generalized text analysis algorithm looks as follows: some texts files are allocated to the input of a subsystem for extraction of parameters, a number of numerical characteristics is derived for each file. An array of indicators for frequential evaluation of the text is built.

At the next stage of text analysis the modules of the morphological analysis are utilized. These modules are represented by a set of text processing tools in a natural language and contain two linguistic processor components, which consequently analyse the source text. In other words, once the first linguistic processor component is finished analyzing the text, the next one takes over and continues the analysis.

The first linguistic processor "Solarix Engine-the Dictionary of Russian" is used for identification of a word's prefix, root, suffix, ending and an initial form of a word. The second linguistic processor "AOT" defines a part of speech (noun, verb, adverb, adjective, etc.) and various characteristics of a word (gender, number, case, etc.). The main module of identification of parts of speech is "Solarix Engine". If the "Solarix Engine" module did not manage to identify a word and its features, then "AOT" is used for post-processing.

The algorithm integrated into these programmatic tools derives a set of relative theoretical-linguistic frequential indicators from a text in Russian. The derived set contains 76 indicators. All indicators are relative, which allows to compare texts of different size. The majority of indicators reflect a percentage of a certain part of speech within the total number of words or other parts of speech, as well as a general share of foreign words within a given text. Because of this, the values of these indicators range between 0 and 1.

The analytical part of tools is based on algorithms of classification of data by the method of regression decision trees, including CHAID, exhausting CHAID, CRT, etc. [Official document for IBM SPSS].

# 5 Data base and a set of indicators

The data source for the experiment is represented by an array of of different types of documents belonging to one of the following science fields: Law, Economy, Biology, Chemistry, History, IT, Pedagogics, Medicine, Physics and Philosophy.

The subset «Books» contains a set of textbooks in Russian published between 1930 and 2019. The subset «Essays» is made up of different types of works: commentaries, course works, essays taken from open sources such as «dissers.ru», «referat.ru», as well as from web-site SPBETU "LETI". All documents in «Essays» subset are thematically

related to the selected clusters (science fields) and published after 2013. Information on clusters is presented in table 1.

Table 1: Corpora of text

| Name | Count | Object | Object in one claster | Average words in objects |
|---|---|---|---|---|
| Books | 2000 | 10 | 200 | 142 827 |
| Essays | 900 | 10 | 90 | 2 892 |

# 6 Classification algorithms and their settings

Materials of publications[Meystre, et al., 2008, Canuto et al., 2018, Cerri, 2015] and results of own researches in the field of thematic classification of texts [Osochkin et al., 2018, Fomin et al., 2016] show that, when using the TF-IDF method for the analysis of texts, regression trees of decisions have a number of advantages over other methods of classification. This includes existence of the effective algorithms for interpretation of results (IF-THEN of rules), as well as increased accuracy of classification when using small training selections. Hence, a method of regression trees was chosen for the purpose of this exercise.

Use of regression trees of decisions demands transformation of text data into a set of indicators. The ways of transformation, methods of calculation and size of a set of indicators [Boytcheva, 2015] are important aspects, which influence the final accuracy of classification. In scientific papers dedicated to optimization of the size of a set of indicators [Chen et al., 2016, Liu, 2018] a pattern was established that a reduction of a number of indicators to an optimal minimum increases the accuracy of classification. Based on a number of experiments analyzing the thematic styles, genres, styles of the authors [Fomin et al., 2016, Osochkin et al., 2018], the number of material indicators was reduced from 76 to 35. The narrowing down was based on estimation of the greatest impact of individual indicators on classification of texts in Russian, as well as a number of secondary considerations, such as: the frequency of use of numerals, adverbs, words in Latin and nouns in different cases, etc. In line with these considerations, the choice was also made in favor of building a regression tree using the "exhaustive CHAID" method with the following settings:

- limit the depth of a tree to no more than 30 levels;
- set the level of significance for nodes splitting to 0.005;
- a maximum number of iterations is 1000;
- 70% non-proportional sampling was used for training.

**Identification of distinctive characteristics of documents.** To solve the problem of differentiation of a short (referential) text from a full text, let's undertake an experiment to identify the unique distinctive features of a text, allowing to classify a given text with high degree of reliability as an "Essay" or a "Book". The results of classification are presented in table No. 2

Table 2: Classification by type of document

| | | Predicted | | |
|---|---|---|---|---|
| | | Books | Essays | % correct prediction |
| Training sample | Books | 1017 | 0 | 100% |
| | Essays | 0 | 471 | 100% |
| | % part | 68,30% | 31,70% | 100% |
| Test | Books | 980 | 3 | 99,70% |
| | Essays | 0 | 429 | 100% |
| | % part | 69,40% | 30,60% | 99,80% |

These results show that the test sample was classified with a total accuracy of 99.8%, which can be considered as a successful classification.

Based on the results of this classification, a 3-level decision tree was built utilizing the following indicators:

• First level: Average length of a word. An average length of a word within the books exceeds 6.54 symbols.

• Second level: a lower number of verbs and nouns are used in books than in essays.

• Third level: Unique words within the total number of words.

The «Essays» are a concise retelling of the paper's theme and, as a consequence, the short sentences with a minimal use of auxiliary parts of speech prevail in the «Essays», i.e. sentences mostly contain subjects and predicates.

The indicator of a number of "unique" words within a total number of words shows a percentage of non-repeatable words within a total number of words in a given text. If the essay's author often refers to the formulation of a problem or a research objective, the uniqueness of words in the «Essays» is reduced.

**Identification of general characteristics and specifics of documents.** The purpose of the experiment is to confirm a presence of general characteristics of data in a natural language related to the same science field, despite being written as different types of documents, e.g. "Essays" vs. "Books". The experiment tries to classify an array of texts shown in table No. 1, i.e. 2900 texts in a natural language related to a range on science fields.

When we are classifying the data, the corpora of the "Books" is used as a training sample, and texts from the "Essays" category are used for the test classification. This approach allows the classification of "Essays" according to the rules of classification used for the body of "Books", so you can identify the presence of common characteristics relevant to a particular field of knowledge.

Table No. 3 below shows the results of the experiment with clusters being numbered as follows: 1 – Law, 2 - Economy, 3 - Biology, 4 - Chemistry, 5 - History, 6 - IT, 7 - Pedagogics, 8 – Medicine, 9 – Physics, 10 – Philosophy.

Taking into account the above mentioned conclusions, it can be argued that, with the help of frequency-morphological analysis, the scientific publications can be classified automatically as an essays or books with a higher degree of reliability.

Table 3: Classification by science fields

| | Science fields | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | % correct prediction |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Training sample | 1 | 205 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | 2 | 0 | 207 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 98,1 |
| | 3 | 0 | 1 | 207 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 98,1 |
| | 4 | 0 | 0 | 0 | 198 | 0 | 0 | 0 | 2 | 0 | 0 | 99 |
| | 5 | 0 | 5 | 2 | 0 | 177 | 0 | 0 | 0 | 0 | 0 | 96,2 |
| | 6 | 0 | 0 | 0 | 0 | 0 | 204 | 0 | 0 | 11 | 0 | 94,9 |
| | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 207 | 0 | 30 | 0 | 100 |
| | 8 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 216 | 3 | 0 | 97,7 |
| | 9 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 198 | 0 | 99 |
| | 10 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 190 | 99,5 |
| | % part | 10 | 10 | 8,3 | 10,1 | 10,1 | 8,9 | 10,1 | 10,5 | 10,4 | 9,3 | 98,2 |
| Test | 1 | 85 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | 2 | 0 | 70 | 0 | 1 | 0 | 1 | 3 | 0 | 3 | 0 | 88,6 |
| | 3 | 0 | 2 | 70 | 2 | 2 | 1 | 4 | 0 | 0 | 0 | 88,6 |
| | 4 | 0 | 0 | 0 | 84 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | 5 | 5 | 0 | 1 | 3 | 94 | 0 | 0 | 0 | 0 | 0 | 88,7 |
| | 6 | 0 | 1 | 0 | 1 | 0 | 69 | 0 | 0 | 5 | 0 | 92 |
| | 7 | 2 | 0 | 0 | 0 | 0 | 0 | 79 | 0 | 2 | 0 | 95,2 |
| | 8 | 0 | 0 | 1 | 0 | 4 | 0 | 0 | 64 | 0 | 0 | 92,8 |
| | 9 | 3 | 2 | 2 | 0 | 3 | 7 | 2 | 0 | 80 | 0 | 88,9 |
| | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 98 | 99 |
| | % part | 11 | 8,5 | 8,9 | 10,5 | 10,3 | 12,1 | 10,7 | 7,6 | 10,8 | 11,6 | 93,3 |

From the results presented in table No. 3 it is obvious that the objects belonging to the same science fields were correctly classified in 93.3% of cases. This allows to confirm the existence of common characteristics for the objects belonging to the same science field, but being technically different types of documents. The main indicators involved in classification of 10 science fields were as follows:

- Percentage of adverbs from the total number of words
- Percentage of Latin characters from the total number of characters
- Average word length
- Percentage of Nouns from the total number of words (in different cases)
- Percentage of Verbs from the total number of words.

The total percentage of predictions shows that most of the inaccurate classifications are related to Economics and Biology (11.4%). The main problem with the Economy is the heterogeneity of its content. Some documents contain theoretical reflections, philosophizing about the economic organization of the state and world's economy in general, thus making the science field of Economy similar to History and Philosophy. Part of the material on Economics is mathematical calculations and formalized methods, which increases its similarity with Physics, IT, etc. There is a similar problem with Biology, i.e. different styles of data presentation. The emphasis on Chemistry and Physics as opposed to a description of general structure or anatomy of living organisms sometimes makes Biology publications similar to Medicine and Philosophy, sometimes closer to other areas such as Physics, etc.

**Hierarchical classification with definition of document type and science field** The works of [Otero et al., 2010, Cerri et al., 2015, Cerri and Carvalho, 2010,

Table 4: Hierarchical classification

| | Science fields | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | % correct prediction |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Training sample | 1 | 200 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | 2 | 0 | 188 | 5 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 94 |
| | 3 | 0 | 1 | 189 | 0 | 1 | 1 | 0 | 3 | 1 | 0 | 94,5 |
| | 4 | 0 | 0 | 0 | 198 | 0 | 0 | 0 | 2 | 0 | 0 | 99 |
| | 5 | 0 | 0 | 0 | 0 | 198 | 0 | 0 | 0 | 0 | 0 | 99 |
| | 6 | 0 | 0 | 2 | 1 | 2 | 195 | 0 | 0 | 2 | 0 | 97,5 |
| | 7 | 1 | 2 | 0 | 0 | 0 | 0 | 195 | 0 | 2 | 0 | 97,5 |
| | 8 | 1 | 0 | 0 | 2 | 0 | 3 | 0 | 195 | 0 | 0 | 94,5 |
| | 9 | 2 | 1 | 4 | 0 | 0 | 3 | 0 | 1 | 189 | 0 | 99 |
| | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 200 | 100 |
| | % part | 10,2 | 9,6 | 10 | 10,2 | 10,1 | 10,2 | 9,8 | 10,3 | 9,9 | 10 | 97,4 |
| Test | 1 | 90 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | 2 | 0 | 88 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 97,8 |
| | 3 | 0 | 0 | 84 | 0 | 1 | 1 | 0 | 3 | 1 | 0 | 93,3 |
| | 4 | 0 | 0 | 0 | 88 | 0 | 0 | 0 | 2 | 0 | 0 | 97,8 |
| | 5 | 0 | 0 | 0 | 1 | 89 | 0 | 0 | 0 | 0 | 0 | ?98,9 |
| | 6 | 0 | 0 | 1 | 1 | 0 | 87 | 0 | 0 | 1 | 0 | 96,7 |
| | 7 | 0 | 1 | 0 | 0 | 0 | 0 | 89 | 0 | 0 | 0 | 98,9 |
| | 8 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 87 | 0 | 0 | 96,7 |
| | 9 | 2 | 1 | 1 | 0 | 1 | 2 | 0 | 1 | 83 | 0 | 92,2 |
| | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 90 | 100 |
| | % part | 10,3 | 9,4 | 9,8 | 10,1 | 10,2 | 10,2 | 9,7 | 10,6 | 9,7 | 10,00% | 96,4 |

Bi and Kwok, 2015], demonstrated the effectiveness of hierarchical classification with the use of trees of decisions. Two approaches stand out:

- treelike identification (top-down), when at first the object's cluster is identified, and then the same is done for a sub-cluster;

- local classification that identifies a cluster and sub-cluster simultaneously. We use the first approach (treelike classification), which enhances rules of classification by adding several levels, thus allowing to increase the general accuracy of classification.

Let's undertake a comparative hierarchical classification of texts' corpora where at the first level we identify the type of the document (books, papers), with the science fields (ten clusters) identified on the second level. The results of classification are shown in table No. 4:

As a result of classification, a 7-tiered tree of "if then" rules was built, allowing to identify science fields and a types of the document with an accuracy of 96.4%.

Identification of the type of a document is achieved within levels from 1 to 3 of a tree. These rules correlate with the rules established earlier when we carried out a classification by the type of the document.

Levels from 4 on 7 utilize distinctive features relevant to a particular science field; rules of classification for hierarchical classification and classification of text's corpuses without taking into account the documents' types are identical.

The above mentioned indicators were used in 80% of cases for identification of science fields. The key indicators for each tree level are presented in table No. 5 below

Table No. 5 shows only those indicators that were used as rules for classification of only 75% of objects within the same level.

Two indicators are used at the fourth level: "Latin symbols within the total number

Table 5: Features used at certain tree levels

| Tree level depth | Features |
|---|---|
| 1 Level | Average word length |
| 2 Level | Nouns within the total number of words<br>Verbs within the total number of words |
| 3 Level | Unique words within total number of words<br>Number of Latin symbols |
| 4 Level | Latin symbols within the total number of symbols<br>Adverbs within the total number of words |
| 5 Level | Nouns within the total number of words<br>Verbs within the total number of words |
| 6 Level | Average length of a word<br>Nouns in different cases |
| 7 Level | Unique characteristics of science fields |

of symbols" and "Adverbs within the total number of words". The use of the “Latin symbols within total number of symbols” as an indicator allowed to separate science fields into several subgroups:

1) History, Physics, IT, Economy, Chemistry Medicine, Philosophy, Medicine ;

2) Biology, Law, Philosophy, Pedagogics;

Also, adverbs are most commonly used outside of natural sciences.

At the fifth level the indicators "Quantity of nouns within the total number of words" and "Verbs within the total number of words" allowed to separate the objects from adjacent groups, e. g. Physics from Chemistry, where in Physics verbs were more often used than in Chemistry, with a higher degree of accuracy.

The sixth level utilizes a lot of indicators, but “Nouns in different cases” are most commonly used. For example, nouns in a genitive case meet twice as often in Economics than in Pedagogics.

The seventh level of classification encompasses the indicators specific for a particular science field. The most relevant indicators for identification of science fields are listed in table No. 6.

# 7 Comparative classification

Let’s try to compare the author’s tools with the two most relevant similar tools allowing to extract theoretically-linguistic indicators from text: NLTK4Russian [Shaik, 2017, Official document for IBM SPSS] is a specialized environment for automatic processing of Russian language texts. NLTK4Russian is based on a number of algorithms and instrumental methods of a modern computational linguistic (NLTK, Pattern, GenSim, etc.).

PyMorphy2 [Korobov, Official document for IBM SPSS] is a morphological analyzer for Russian language based on NLTK algorithms with several modifications for morphological modules.

Table 6: Indicators used to identify specific science fields

| Science field | Indicators |
|---|---|
| IT | Use of nouns in a masculine gender is more often than in other science fields; frequent use of participial phrases, Latin symbols and numerals. |
| History | The share of Latin symbols and numerals within total number of symbols is slightly lower than for Medicine and IT; “Average length of words” and “Number of words” in one sentence is lower than for a majority of other science fields. |
| Chemistry | Prevalence of Latin symbols, special symbols, numerals, participles of perfective aspect are more rarely used. |
| Law Average | length of words is less than for a majority of other science fields. At the same time, longer sentences prevail with the frequent use of punctuation marks. |
| Biology | Frequent use of adjectives and verbs in a text. Latin symbols and complicated sentences are rarely used. The balanced use of other parts of the speech in texts. |
| Medicine | Frequent use of Latin symbols, numerals (including numerals written in words), imperfect verbs, balanced sentences. Texts are often similar with Biology, Chemistry, Philosophy |
| Pedagogics | Frequent use of nouns of a feminine gender and nouns in an instrumental case. Texts contain long sentences. |
| Physics | Prevalence of Latin symbols, numerical symbols. Frequent use of verbs and adverb phrases is common. A rare use of phrases like “Usually, ...”, “As a rule, . . . ”, “Periodically, . . . ” etc. |
| Philosophy | The lowest use of numerals; the longest words; frequent use of punctuation marks. Economy The balanced texts without any obvious distinctive features. |

Comparative analysis of various morphological modules and parsers for the Russian language [Lashevskaja 2014, SpbTU] showed that NLTK4Russian and PyMorphy2 are quite accurate and reliable modules for the analysis of the Russian language.

The indicators extracted through PyMorphy2 and NLTK4Russian are calculated with the use of the same formulas as FaM’s indicators in the analysis of the text corpuses presented in table No. 1. The comparative results of accuracy of classification by science fields are presented in table No. 7.

The results of a comparative experiment show that the highest accuracy utilizing the use of a method of regression trees was achieved by using a set of the indicators extracted through "FaM". The application of FaM allowed to receive the highest accuracy of 96.4% for hierarchical classification and 93.3% for classification by science fields without a reference to a type of the document.

The efficiency of "FaM" is due to the distinctive features of its architecture. All three tools use a two-level morphological analysis of words. The first level is a word normalization (a reduction of a word to its initial form) and comparison with the dictionary by

Table 7: Comparative classification

| Classification | FaM | NLTK4 Russian | PyMorphy2 |
|---|---|---|---|
| Science fields (not HMC) | 93,3% | 88,4% | 83,3% |
| Science fields (HMS) | 96,4% | 95,8% | 89,7% |

A.A. Zaliznyak. If a word hasn't been found within the dictionary or a full similarity with the word's form hasn't been established, then a risk of a mistake is calculated. If this risk is high enough, then another auxiliary dictionary is used. In addition to this, this auxiliary dictionary is actually a module that was built based on a modified system of the French-Russian machine translation. This system was specifically designed to derive the initial form of a word, as well as to perform a context-dependent analysis of phrases or even whole sentences.

# 8 Conclusions

As a result of this research, an extensive array of textbooks and essays was digitized and transformed into a set of theoretical-linguistic indicators. This set of indicators reflects the specifics of texts involved, which is confirmed by a high precision of classification by science fields.

Normally, a transformation of a semi-structured text information into a set of theoretical-linguistic indicators leads to a loss of many linguistic, semantic and other text features. However, it was demonstrated that the proposed approach manages to preserve a range of crucial text characteristics, thus still allowing to cross-reference the documents with their respective science fields.

The hierarchical classification employed in this research demonstrated its effectiveness by helping to increase the accuracy of classification. The majority of the mistakes made throughout this classification indicate a very wide heterogeneity of stylistics in articles involved, which somewhat lowered the general accuracy.

The results of experiments confirmed the effectiveness and good potential of the proposed integrated technique (based on algorithms of frequency-morphological analysis and regression trees) in application to the tasks of locating the referential materials and textbooks, deriving their table of content and a field of science they relate to. The utilization of a method of classification by regression trees of decisions allows to perform the semantic analysis of the text, thus helping to maintain a logical connection between derived indicators and their respective data clusters.

In the upcoming articles we plan to expand our experiments into a classification of bigger arrays of texts, as well as to develop the ways to identify the genres of literature, authors' styles, tables of contents, etc.

# 9 Acknowledgements

The research was supported by the Ministry of Science and Higher Education of the Russian Federation (project No. 2.2939.2017/4.6).

# References

[Ke, 2007] Ke W. (2007) Collaborative classifier agents: studying the impact of learning in distributed document classification // J. Mostafa, Y.Fu. - In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. ACM, New York, JCDL '07. Pp. 428–437

[Moraes et al., 2013] Moraes R., Valiati J.F., and Gavião Neto W.P. (2013) Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 2013, No. 40. Pp. 621–633

[Pontiki et al., 2014] Pontiki M., Galanis D., Pavlopoulos J., Papageorgiou H., Androutsopoulos I., Manandhar S. ( 2014 ) Task 4: Aspect based sentiment analysis. //The 8th Intern. Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland. 2014. Pp. 27–35

[Chen et al., 2016] Chen K. et al. (2016) Turning from TF-IDF to TF-IGM for term weighting in text classification. //Expert Systems with Applications, vol. 66. Pp. 245–260

[Dikoveckij et al., 2010] Dikoveckij V.V., Shishaev M.G. (2010) NLP models for web-search system [Obrabotka tekstov estestvennogo yazyka v modelyah poiskovyh sistem] //The journal "proceedings of the Kola science centre of RAS [Jurnal «Trudy Kol'skogo nauchnogo Centra RAN»]. Pp. 29 – 34 (In Rus.)

[Boytcheva, 2015] Boytcheva S., G.Angelova1 , Z.Angelov , D.Tcharaktchiev (2015) «Text Mining and Big Data Analytics for Retrospective Analysis of Clinical Texts from Outpatient Care» Cybernetics and information technologies Vol.4 (15). Pp. 58–77

[Fomin et al., 2016] Fomin V.V., et al. (2016) Classification of texts based on frequency and morphological analysis using data-mining algorithms[ Klassifikaciya tekstov na osnove chastotnogo i morfologicheskogo analizov s primeneniem algoritmov data-mining] //Informatization of education and science, Institute of information technologies and telecommunications [Gosudarstvennyj nauchno-issledovatel'skij institut informacionnyh tekhnologij i telekommunikacij] (Moscow), Vol. 3. Pp. 137-152 (In Rus.)

[Meystre, et al., 2008] Meystre, S. G. Savova, K. C. Kipper-Schuler, J. F. Hurdle. (2008) Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research», Yearbook of Medical Informatics. Pp. 128-144

[Canuto et al., 2018] Canuto, S. et al. (2018) «Thorough Evaluation of Distance-Based Meta-Features for Automated Text Classification.» IEEE Transactions on Knowledge  Data Engineering, vol. 30, № 12, pp. 2242–2256

[Liu, 2018] Liu L. (2018) An effective dimensionality reduction method for text classification based on TFP-tree. //Journal of Intelligent  Fuzzy Systems. vol. 34,No 3.Pp. 1893–1905

[Osochkin et al., 2018] Osochkin A.A, et al. (2018) Text-minig experiments on the classification of texts in the framework of the problems of personalization of the educational environment [Eksperimenty text-minig po klassifikacii tekstov v ramkah zadach personalizacii obrazovatel'noj // Informatization of education and science [Informatizaciya obrazovaniya i nauki]. Vol. 2 (38). 2018. Pp. 38-50 (In Rus.)

[Galitsky 2017] Galitsky B. (2017) «Learning Noisy Discourse Trees» Oracle Corp Redwood Shores CA USA, pp. 89-102 avable at:
http://www.dialog-21.ru/media/3982/dialogue2017_v1.pdf

[Wei-Cheng et al., 2017] Wei-Cheng C. et al. (2017) Deep Learning Approach for Extreme Multi-label Text Classification // Microsoft Research EURASIP Journal on Wireless Communications and Networking. Language Technologies Institute. 8-th Dec. 2017 avable at:
http://nyc.lti.cs.cmu.edu/yiming/Publications/jliu-sigir17.pdf

[Baoxun et al., 2012] Baoxun X., Xiufeng G. et al., (2012) An Improved Random Forest Classifier for Text Categorization //Journal of computers , vol. 7 (12). Pp. 2913-2920.

[Medlock, 2008] Ben W. Medlock (2008) Investigating classification for natural language processing tasks. Cambridge: University Cambridge. - 138 p.

[Shaik, 2017] Shaik J. Machine Learning, NLP: Text Classification using scikit-learn, python and NLTK. Towards Data Science 102 avable at:
https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a

[SpbTU] Official site departmetns mathematical linguistics Saint Petersburg State University avable at: http://mathling.phil.spbu.ru/node/160

[Korobov] Korobov M. Official docs for «Pymorphy2» avable at:
https://pymorphy2.readthedocs.io/en/latest/user/index.html

[Lashevskaja 2014] Lashevskaja O.N Evaluation of automatic text analysis methods: morphological parsers of the Russian language avable at: http://www.dialog-21.ru/evaluation/2010/morphology/

[Official document for IBM SPSS] Official document for IBM SPSS avable at:
http://www.math.uni-leipzig.de/pool/tuts/SPSS/IBM%20SPSS%20Decision%20Trees.pdf

[Cerri, 2015] Cerri R.(2015) An extensive evaluation of decision tree based hierarchical multilabel classification method and performance measures //Computational Intelligence, Volume 31, No 1, 2015 avable at:
https://ru.scribd.com/document/333057486/An-Extensive-Evaluation-of-Decision-Tree-Based-Hierarchical-Multilabel-Classification-Methods-and-Performance-Measures

[Honda, 2008] Honda M.Thinking Linguistically: A Scientific Approach to Language. Wiley, 2008. - 253p.

[Jones, 2004] Jones, K.S. A (2004) Understanding Inverse Document Frequency: On theoretical arguments for IDF. Microsoft Research UK (and City University, London, UK) avable at:
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.2284&rep=rep1&type=pdf

[Bi and Kwok, 2015] Bi and J.V. Know (2015) «Bayes-Optimal Hierarchical Multilabel Classification» IEEE transactions on knowledge and data engineering vol. 27.No 11, November 2015.

[Cerri and Carvalho, 2010] Cerri and Carvalho (2010) «New top-down methods using SVMs for Hierarchical Multilabel Classification problems» Conference: International Joint Conference on Neural Networks, IJCNN 2010, Barcelona, Spain, 18-23 July, 2010.

[Cerri et al., 2015] R.Cerri, C.B.Rodrigo, J.Wehrmann (2015) «Hierarchical Multi-Label Classification Network» Proceedings of Machine Learning Research avable at:
http://proceedings.mlr.press/v80/wehrmann18a/wehrmann18a.pdf

[Otero et al., 2010] F.B. Otero, C. Johnson , A. Freitas (2010) «A hierarchical multilabel classification ant colony algorithm for protein function prediction» Memetic Computing vol. 3, No 2, September 2010. Pp. 165-181 .