How Do People Read Science Fiction and Why is it Popular: Common Tendencies and Comparative Analysis

Karina Agafonova[1], Kseniya Bekisheva[1], Olga Drabenia[1], Arina Petrova[1], Krassimira Ivanova[2]

[1] *ITMO University, St. Petersburg, Russia*
[2] *Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences, Sofia, Bulgaria*

Abstract

The article presents a study in the field of digital humanities, dedicated to the genre of science fiction, which is one of the most popular in the world and Russia. The assumption is that there should be trends or similarities between highly rated books and also that certain descriptive characteristics could be particular to a specific subgenre. The comparative analysis of the chosen subgenres is made and some tendencies in acquired data from statistical analysis of reader groups and their preferences and topic modelling of the annotations of the books are outlined.

Keywords

Digital Humanities, Comparative Analysis, Data Mining, Topic Modelling, Latent

1 Introduction

The digital transformation of society in recent years has led to a qualitatively new leap in the development of social sciences and opens new horizons for making social analyses. Nowadays, the cooperation between applied mathematics, informatics and information technologies, and another science or scientific discipline gives rise to new disciplines, such as "digital humanities", which is in our focus here. By the words of Brett Bobley of the National Endowment for the Humanities [1] "digital humanities (DH) is just an umbrella term – a term of convenience – that refers to a whole bunch of activities happening where the humanities interact with technology".

Digital humanities are at the leading edge of applying computer-based technology in the humanities, combining methodologies from traditional humanities with computer science, opening up new possibilities for data collection and visualization, information retrieval, data analysis, and data mining (DM). Nowadays, the application of data analysis and knowledge extraction methods is expanding and conquering the humanitarian field, which usually was considered as unanalysable due to the unstructured nature of the materials and to the more blurred distinction between objects and phenomena in the observed processes.

Literature is one of the main cultural artefacts, which is based on creative imagination and is able to keep and spread a cultural code relevant to a social group, to which literature belongs to. Moreover, literature represents language and, in this regard, – a specific mindset that is unique to a certain group of people. Additionally, literature is not only a product of imagination but also a research field that is studied for educational purposes.

In our research, we have concentrated on one genre – Science Fiction (SF). We have chosen to research it in both personal and academic interests, the latter being solidified by previous research done, for instance, by Science Fiction Foundation [2]. Moreover, judging by the amount of Science Fiction

Education and Research in the Information Society, September 27–28, 2021, Plovdiv, Bulgaria
EMAILS: agafonovakarin@gmail.com; morinberk0@gmail.com; o.drabenia@yandex.ru; petrova.arina.s@yandex.ru; kivanova@math.bas.bg
ORCID: 0000-0001-5813-9732; 0000-0003-2105-4238; 0000-0003-3058-4863; 0000-0002-6462-3050; 0000-0001-5056-7513

archives, we assume that this genre is popular among both readers and academics [3], [4], [5]. We have chosen "FantLab" for our research – a digital archive, following the tradition of preserving science fiction books' metadata. It is the most extensive archive of this kind in Russia, and it is open for users' contribution: they are free to rate books, add books' characteristics to descriptions, and leave comments. All of this creates a unique dataset: the archive consists of the best samples of the genre, selected by literary scholars or book editors, and data on many less famous and notable texts. This brings us closer to the method of distance reading developed by the Italian literary historian Franco Moretti [6]. The essence of the method lies in the study of a large corpus of texts, which include the established literary canons and other much less well-known texts. Even though our tasks do not study the texts themselves, we can get some conclusions only based on metadata and annotations. Here, however, our research is in close contact with digital anthropology since the choice of users determines the metadata of books.

It is worth stopping here and noting that our study is akin to two others, which also examine large corpora of science fiction texts. The first one is dedicated to the research of "Bob Gibson anthologies of speculative fiction". The author of this collection "harvested a wide range of science-fictional materials for his more than 890 anthologies from primarily English-language magazines published from the 1840s onwards". [7] The researchers, in turn, tried to explore the boundaries of the subgenres of fiction using the visualization method and relying on a system of symbols developed by Gibson himself. The second paper, "Science fictions, cultural facts: A digital humanities approach to popular literature," reflects the idea that science fiction is the mirror of human culture in general. The dataset was "popular SF magazines written in English, following the first publication of Amazing Stories in 1926." [8] The researcher tried to define the boundaries of the genre by analysing many sociological surveys of the target audience of magazines and classified magazine covers using digital methods.

From the above, it becomes clear that our research, like the other two, strives to explore the genre framework of SF based on a large corpus of texts. However, both datasets – with anthologies and journals, despite the representativeness of the samples, have a common flaw – anglocentric. Our archive, in turn, has a bias towards Russian-language works and does not include unpopular examples of the genre that have not been translated into Russian. We assume that geographic and cultural traits have a significant impact on the archive content; hence we would like to contribute to the study of the genre from the paradigm of those aspects that characterize our country and culture in particular and do not apply the results of our research to the entire genre.

In addition, our research does not set itself the task of investigating the historical process of the formation of SF subgenres, just as it does not investigate the culture through the archive's data. So, we pursued the following goals:

- to look at a portrait of a reader for each genre by analyzing users' sex and age to get some idea of the target audience of the site and the archive as a whole,
- detect tendencies in acquired data using statistical analysis,
- perform a comparative analysis of the chosen subgenres,
- create a predictive model of subgenre based on descriptions of the books, which can be used as a supporting tool for automatizing ingestion in the repository.

The next part of the article is organized as follows: Chapter 2 explains the used methodology of the research, Chapter 3 presents the dataset and describes preprocessing works done over the data, Chapter 4 shows the achieved results and possible explanation of some correlations. Chapter 5 is devoted to deeper text analysis of the annotations of the books. Finally, some conclusions and possible future work are pointed out.

2 Methodology

To conduct our research, we used quantitative and comparative analysis as well as some data mining and machine learning techniques. The quantitative method is applied by gathering data using structured search instruments, the results of which are based on a monumental sample that is representative of a certain reader group. Finally, we applied a quantitative method by generalizing the research concept and investigating the relationships within the concept. The comparative method is applied by investigating the similarities and differences within selected genres and between them. We compared

average scores for the books in general, by gender and age. Moreover, we compared general characteristics of the books and looked into dependency between an average score and the characteristics properties. Also, we have applied some data mining techniques for revealing some hidden relationships between attributes. Speaking of machine learning techniques, we used topic modelling (gensim library in particular) to gather more information for further analysis. Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents [9]. For our research, the Latent Dirichlet Allocation (LDA) method is used as a method for topic modelling of the annotations of books by subgenres.

The results of the analysis are compared with domain knowledge.

3 Dataset

As a source of data, we have used the FantLab resource (https://fantlab.ru/), which contains one of the biggest collections of metadata for fiction books across the Internet in Russia. We concluded that the FantLab resource would be suitable to achieve our goals due to its open API [10] and broad collection of books. Moreover, its genre-thematic classified metadata and the scores' system provide information that is relevant to our research question and hypothesis and allow us to perform gender, age and characteristics analysis.

We have used the Web Scrapper extension [11] to collect the main body of the books' metadata by the following categories: *title*, *author*, *year of publication*, *language of the publication*, *main characteristics*, *genre*, *subgenre*, *place of action*, *time of action*, *plot goal*, *plot linearity*, *recommended age*, *annotation*, *total voters*, *total votes*, *average rating*, *gender histograms*, and *age histograms*. However, due to some technical issues with the extension we were able to obtain just a portion of data. Nevertheless, it was a greater part of the entire set.

Overall, we have extracted data regarding 3709 records for books in the "Fantastic" section from the following subgenres: *Hard SF*, *Soft SF*, *Space Opera*, *Planetary Fantastic*, *Utopia*, *Dystopia*, *Cyberpunk*, *Timepunk*, *Temporal Fiction*, *Post-apocalyptic*, and *Catastrophe*. The extraction was made with corresponded requests of the subgenre, so, we have established a new *subgenre-label* attribute based on which request the given record was received from (as far as in the original attribute subgenre only the main subgenre is recorded, but some books belong to more than one subgenre). Also, since the *main characteristics* contained more than one value (from a controlled vocabulary), we broke this attribute into 22 attributes indicating whether or not there was a corresponding storyline value, such as: Adventure, Anti-war, Humorous, Parody, Religious, etc. We have ignored 5 values because of the minor presence.

After data cleaning (removing some outlier records and refining some attribute values, but keeping duplicate books that are pointed to different subgenres) the final dataset consists of 3676 records, distributed in 11 subgenres as is shown in Table 1.

Table 1. Examined Dataset, by subgenres

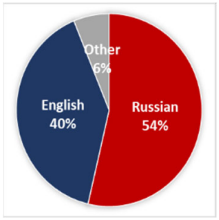
subgenre	number of records
Catastrophe	40
Cyberpunk	301
Dystopia	418
Hard SF	534
Planetary Fantastic	378
Post-apocalyptic	459
Soft SF	441
Space Opera	475
Temporal Fiction	442
Timepunk	95
Utopia	93
Total	3676

In order to support the DH community, we have shared the used instruments for data gathering in GitHub open repository [12] and the final dataset – in Kaggle repository [13].

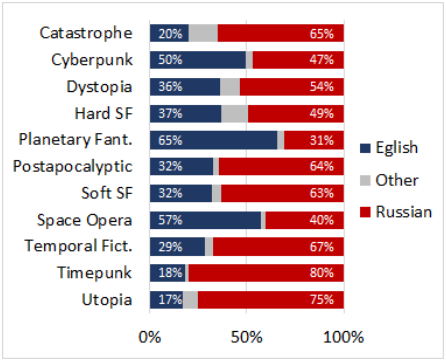
4 Analysis of the interconnections between subgenres and attributes extracted from the descriptions of the books

We have made a statistical analysis of the original languages of the observed works. The spread between languages was from Russian (54%) and English (40%), followed by Polish, French, German, Ukrainian, Bulgarian, etc. (30 languages total), but all other languages together have a minor presence against Russian and English (Figure 1a). Because of this result, we have focused further analysis on these two languages.

From the language frequency analysis, we have observed that from all of the analysed Science Fiction subgenres, English is the most common language for *Planetary Fantastic* and *Space Opera*, while for *Timepunk* Russian prevails significantly (Figure 1b). This may be connected with the fact that the FantLab library (as a Russian-language resource) has more books in Russian than in other languages. However, it centres on popular English-language books too. The results for *Space Opera* can be explained with help of the *Language – Year of publication* correlation.



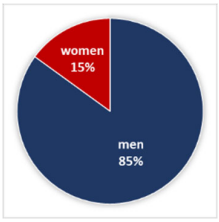
(a)



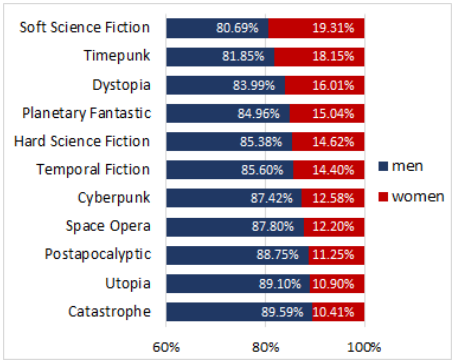
(b)

Figure 1. Percentage of original languages of the observed works:
a) total distribution; b) by subgenres.

Moreover, we have analysed the correlation between subgenres and reading auditorium by sex and age groups. For this purpose, we relied on the information about the voters for the respective books in the library. We are aware that the activity of readers to give feedback is different, but we assume that within the entire readership, active users can be considered as a representative sample of all readers.



(a)



(b)

Figure 2. Percentage of men' women' votes:
a) total distribution; b) by subgenres.

To that end, the total percentage of male readers is about 5 times more than female readers, which might mean that such kind of literature is preferred by men (Figure 2a). On this background, *Soft SF* and *Timepunk* are more preferred by women than other subgenres, while the presence of men in the fan groups of *Post-apocalyptic*, *Utopia*, and *Catastrophe* is even stronger (Figure 2b). Nevertheless, we do not imply that this correlation is universal or generalized, as the results might be determined by the target group of the website.

Looking at the distribution of scores (votes are scaled from 1 to 10) we can see that there is no big difference between different subgenres (Figure 3). This can be explained by the fact that good authors are not only in one specific domain – it depends on the author's mastery to captivate the reader, regardless of the place and time of the action and the plot linearity. The only *Catastrophe* has a significantly low assessment compared to the others. Looking at the distribution of votes between different age groups (Figure 4) we found the tendency of more exciting ratings in the young age group, which slowly decreases to older years. Here again, the *Catastrophe* is distinctively lower against the other ones, but keeps the same tendency.

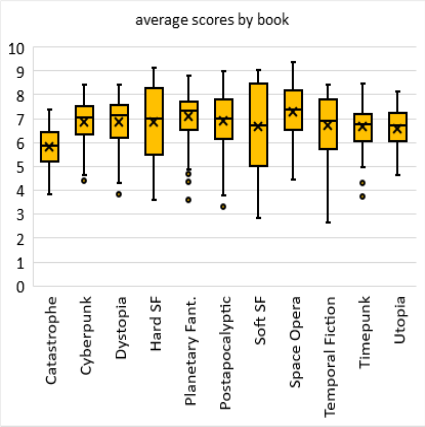


Figure 3. The distribution of scores by subgenres

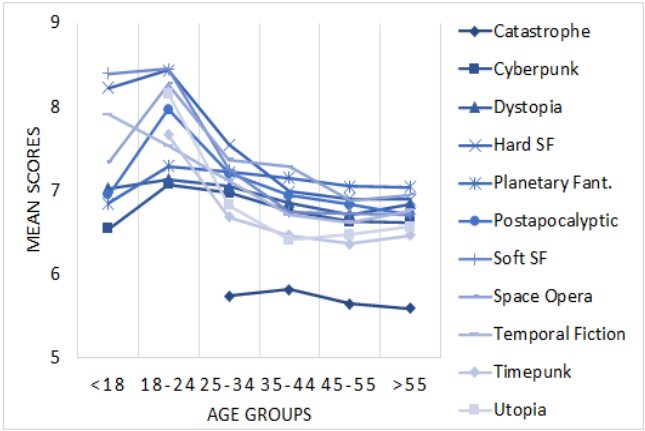


Figure 4. Mean scores for subgenres, by different age groups

We have made several experiments with a dataset, formed on the base of the primary one, which contains only some of the book descriptors, such as: *place of action*, *time of action*, *plot goal*, *plot linearity*, and reformed 22 attributes that describe *main characteristics*, using a *subgenre-label* as a class label. After preliminary tests of the performance of different classification models (Naïve Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, Support Vector Machines), we stopped our attention on the Support Vector Machines, which has achieved highest accuracy for this dataset.

Applying 5-fold cross-validation using WEKA's SMO (representative of Support Vector Machines in the Waikato Environment of the Knowledge Analysis) we have achieved 51.58 % accuracy. This is not enough for creating a reliable classification model, but taking into account the number of class labels (11) it showed some stable dependencies between chosen attributes and corresponding classes. Also, our other task was to see where the confusion matrix shows the mixing between classes.

According to the results, presented in Table 2, we can see that *Space Opera*, *Timepunk*, and *Dystopia* are more distinctive than the others. There is a high dependency between *Catastrophe* and *Post-apocalyptic*, which can be explained with the minor presence of *Catastrophe* in the dataset and similar storyline with *Post-apocalyptic*. Also, there is a label misplacing also in the prediction model of *Utopia*, replacing prediction mostly with *Dystopia* and *Hard SF*.

Table 2. Confusion matrix of the SVM applying 5-fold cross-validation presented as percentage of correctly (on the diagonal) and incorrectly recognized instances per class

belongs to: \ classified as:	Catastrophe	Cyberpunk	Dystopia	Hard SF	Planetary Fantastic	Post-apocalyptic	Soft SF	Space Opera	Temporal Fiction	Timepunk	Utopia
Catastrophe	15.00	0.00	17.50	7.50	0.00	40.00	7.50	2.50	10.00	0.00	0.00
Cyberpunk	0.00	35.88	17.94	4.65	3.65	21.59	3.32	9.97	2.66	0.33	0.00
Dystopia	0.00	7.18	64.35	3.35	1.91	10.77	5.50	1.44	2.15	1.44	1.91
Hard SF	0.56	1.50	5.06	50.00	9.74	5.43	9.18	8.99	8.24	0.37	0.94
Planetary Fantastic	0.00	1.32	1.85	11.90	36.77	0.79	3.70	41.53	1.32	0.00	0.79
Post-apocalyptic	0.22	6.10	16.12	5.66	0.44	58.82	5.01	1.96	4.79	0.87	0.00
Soft SF	0.45	2.49	8.16	13.15	7.03	7.03	36.96	10.88	13.38	0.23	0.23
Space Opera	0.00	1.26	0.63	5.05	12.42	2.74	4.63	72.42	0.84	0.00	0.00
Temporal Fiction	0.00	1.13	4.30	10.18	1.13	10.86	9.05	2.71	59.05	1.13	0.45
Timepunk	0.00	0.00	9.47	7.37	0.00	4.21	0.00	4.21	7.37	65.26	2.11
Utopia	0.00	2.15	26.88	23.66	5.38	5.38	7.53	4.30	11.83	5.38	7.53

This gives us the confidence to look closer to the correlation of the description attributes and class labels. We focus our attention to: *place of action*, *time of action*, *plot goal*, and *plot linearity*, observing the distribution of the values of corresponded attributes, presented with more than 5% in the dataset.

To that end, according to the results of *place of action* (Table 3), there is a very high correlation between *Virtual reality* and *Cyberpunk* (more than 80%). Also, the place of action *Outside the Earth* or *Another word* often is connected with *Space Opera*.

Table 3. Distribution of major values of the attribute *place of action*

subgenres: / **place of action:**	Catastrophe	Cyberpunk	Dystopia	Hard SF	Planetary Fantastic	Post-apocalyptic	Soft SF	Space Opera	Temporal Fiction	Timepunk	Utopia
Our world (Earth)	1.7	9.9	16.8	15.0	0.9	20.3	13.0	1.2	18.0	0.4	2.9
Outside the Earth	0.1	4.0	1.9	16.4	27.6	1.0	11.9	34.2	1.7	0.0	1.1
An alternative history of our world (Earth)	0.6	1.7	19.3	1.7	0.0	8.3	1.7	1.1	20.4	36.5	8.8
Another world	0.0	3.6	14.3	1.8	26.8	0.0	1.8	39.3	1.8	10.7	0.0
Parallel world/universe	0.0	0.0	4.1	22.4	4.1	20.4	18.4	0.0	2.0	24.5	4.1
Virtual reality	0.0	80.9	4.3	0.0	0.0	0.0	0.0	4.3	6.4	4.3	0.0

The correlation between *time of action* and subgenres is not so distinctive (Table 4), although there are some tendencies in the correlations – *Distant future* usually is used in *Planetary Fantastic* and *Space Opera*, while *20th century* as a starting point of temporal travel is also often used.

Table 4. Distribution of major values of the attribute *time of action*

subgenres: / **time of action:**	Catastrophe	Cyberpunk	Dystopia	Hard SF	Planetary Fantastic	Post-apocalyptic	Soft SF	Space Opera	Temporal Fiction	Timepunk	Utopia
Distant future	0.1	4.3	4.3	16.6	20.5	5.3	13.0	31.0	3.4	0.2	1.3
Near future	1.2	18.9	25.0	10.6	3.0	26.5	8.0	0.5	4.4	0.1	1.9
20th century	0.9	0.0	5.8	23.8	7.3	3.3	16.4	1.7	30.9	5.4	4.4
21th century	5.3	9.9	12.5	9.2	1.3	18.1	16.4	4.3	15.5	3.0	4.6
Indefinite time	1.7	6.4	13.3	5.8	12.1	16.8	9.2	15.6	9.2	9.8	0.0

The correlation of plot goal and subgenres (Table 5) is most distinctive in the pair *Travelers (hitmen)* and *Temporal Fiction*. It is no surprise also that *Inventions and research* are often connected with *Hard SF*, and *Artificial Intelligence* with *Cyberpunk*.

Table 5. Distribution of major values of the attribute *plot goal*

subgenres: / plot goal:	Catastrophe	Cyberpunk	Dystopia	Hard SF	Planetary Fantastic	Post-apocalyptic	Soft SF	Space Opera	Temporal Fiction	Timepunk	Utopia
Becoming/growing a hero	0.4	5.9	21.9	6.9	11.9	10.0	11.5	15.2	10.6	2.4	3.3
Travel to a special destination	0.0	12.3	7.3	8.3	16.5	20.6	3.8	17.5	9.3	4.0	0.5
Inventions and research	0.5	6.6	7.3	37.5	2.8	3.5	13.7	1.0	16.2	5.3	5.6
Genetic experiments, mutations	0.4	14.8	13.0	12.2	8.7	27.4	10.9	11.7	0.9	0.0	0.0
Contact	0.9	0.0	2.7	28.7	17.9	1.8	27.4	13.5	4.9	0.4	1.8
Artificial intelligence	0.0	36.5	7.1	15.2	4.6	8.1	11.7	15.7	0.5	0.0	0.5
Travelers (hitmen)	1.0	0.0	2.1	0.0	11.9	4.1	4.1	1.5	71.1	1.5	2.6

Concerning attribute *plot linearity* (Table 6) there are some similarities between *Hard SF* and *Soft SF* as these genres both have approximately equal percentages in such plot goals as *contact* and *artificial intelligence*, although they also have a difference in the topic of *inventions and research*. Moreover, according to the table, we can assume that *Space Opera* has multiple plot goals due to its entertainment and purposes.

Table 6. Distribution of major values of the attribute *plot linearity*

subgenres: / plot linearity:	Catastrophe	Cyberpunk	Dystopia	Hard SF	Planetary Fantastic	Post-apocalyptic	Soft SF	Space Opera	Temporal Fiction	Timepunk	Utopia
Linear	1.2	7.4	10.8	16.4	11.2	10.4	15.3	10.5	13.1	1.9	1.9
Linear with excursions	0.9	9.2	13.9	12.2	8.6	19.0	7.0	11.4	9.9	3.9	3.9
Linear parallel	1.1	11.2	10.1	10.1	11.2	10.5	4.6	25.7	8.6	3.6	3.2

5 Annotation analysis

Moreover, we have performed text analysis of the annotations of the books. So, some interesting similarities and contrasts were observed, which found their explanation in the domain knowledge.

The first one concerns *Utopia* and *Dystopia* word clouds (Figure 5 a&b). Even from the first glance, we can see that there is not much difference between them: the most frequent words for both are “war”, “human”. It might be explained in two ways. There is quite a literary explanation: every *Utopia* is a *Dystopia* at its very core. There cannot be an ideal society and over rational attempts to build one have something evil wrong in them. Moreover, *Dystopia* is a reversed *Utopia*, and both subgenres touch the same topics: society, state structure, ideals, ethics, human rights, love, and so on.

All of the above-mentioned subgenre word clouds are quite representative: they consist of words, deeply connected with topics, associated with these subgenres. Although *Soft SF* and *Hard SF* seems very similar, there is “human” at the centre of the *Soft SF*, which logically underlines its main concern – not future, not planets, not aliens for the sake of science success representation as for *Hard SF*, but future, planets, aliens as the source of decorations and unusual circumstances to put people in, to test their ethics, to explore human nature (Figure 5 c&d).

(a) Utopia (b) Dystopia (c) Soft SF (d) Hard SF

Figure 5. Annotation analysis

Additionally, we performed topic modelling, which proved to be a critical part of the annotations' analysis. As we have stated before, we used the Gensim library to fulfil the task and more specific information on the process can be found in the "Methodology" section. Upon finishing the topic modelling, we have found out that some of the genres share a great number of similarities, such as *Hard SF* and *Soft SF*, *Space Opera* and *Temporal Fiction*, and *Dystopia* and *Utopia* which we labelled as "couples".

```
[(0, '0.012*"которыи" + 0.011*"человек" + 0.010*"свои"
+ 0.010*"планета" + 0.006*"земля" + 0.006*"время"
+ 0.006*"год" + 0.005*"становиться" + 0.005*"корабль"
+ 0.005*"космическии"'),
(1, '0.020*"новыи" + 0.014*"робот" + 0.006*"оставаться"
+ 0.005*"никто" + 0.005*"идти" + 0.004*"любои"
+ 0.004*"находиться" + 0.004*"способныи" + 0.003*"причина"
+ 0.003*"раса"'),
(2, '0.016*"будущее" + 0.006*"начинаться" + 0.004*"сердце"
+ 0.004*"кольцо" + 0.003*"повелитель" + 0.003*"зонд"
+ 0.002*"закрывать" + 0.002*"уносить" + 0.002*"деятельность"
+ 0.002*"сражаться"')]
```

(a) Hard SF

```
[(0, '0.011*"свои" + 0.010*"которыи" + 0.010*"человек"
+ 0.008*"планета" + 0.005*"жизнь" + 0.005*"год"
+ 0.004*"время" + 0.004*"мир" + 0.004*"становиться"
+ 0.004*"космическии"'),
(1, '0.011*"оставаться" + 0.006*"смочь" + 0.005*"главныи"
+ 0.005*"лес" + 0.005*"будущее" + 0.004*"обнаруживать"
+ 0.004*"идти" + 0.003*"против" + 0.003*"увидеть"
+ 0.003*"солнце"'),
(2, '0.004*"ставить" + 0.004*"переворот"
+ 0.002*"непонятныи" + 0.001*"хвост"
+ 0.001*"цетагандиискии" + 0.001*"спонсор"
+ 0.001*"гаррисон" + 0.001*"бледнотик" + 0.001*"хлопотныи"
+ 0.001*"обманывать"')]
```

(b) Soft SF

```
[(0, '0.011*"которыи" + 0.010*"мир" + 0.008*"свои"
+ 0.008*"жизнь" + 0.006*"год" + 0.006*"новыи"
+ 0.005*"земля" + 0.004*"становиться" + 0.004*"жить"
+ 0.004*"будущее"'),
(1, '0.045*"человек" + 0.010*"век" + 0.006*"решать"
+ 0.005*"дом" + 0.004*"понимать" + 0.004*"служба"
+ 0.004*"цивилизация" + 0.004*"свободныи" + 0.004*"вернуться"
+ 0.003*"попадать"'),
(2, '0.004*"равныи" + 0.004*"компания" + 0.004*"большинство"
+ 0.003*"флэшбэк" + 0.002*"служащии" + 0.002*"охранять"
+ 0.002*"просить" + 0.002*"завтра" + 0.001*"стандарт" + 0.001*"дух"')]
```

(c) Dystopia

```
[(0, '0.009*"свои" + 0.008*"год" + 0.007*"мир"
+ 0.007*"человек" + 0.006*"время" + 0.005*"воина"
+ 0.005*"которыи" + 0.004*"новыи" + 0.004*"страна"
+ 0.003*"герои"'),
(1, '0.010*"земля" + 0.010*"жизнь" + 0.006*"будущее"
+ 0.006*"россия" + 0.004*"молодои" + 0.003*"освоение"
+ 0.003*"против" + 0.003*"второи" + 0.002*"семья"
+ 0.002*"мечта"'),
(2, '0.002*"смерть" + 0.001*"исаакович" + 0.001*"слабыи"
+ 0.001*"бункер" + 0.001*"кардинальныи" + 0.001*"греитинг"
+ 0.001*"договорпullцикл" + 0.001*"стенаисчезать"
+ 0.001*"вернуться" + 0.001*"бульверлиттон"')]
```

(d) Utopia

Figure 6. The results of topic modelling for two observed pairs subgenres

The first "couple" (Figure 6 a&b) shares an interest in space with similar keywords such as *"солнце"* ("sun"), *"космический"* ("cosmic") for *Soft SF* and *"планета"* ("planet"), *"земля"* ("Earth"), *"космический"* ("cosmic") for *Hard SF*. These results are probably connected to the genres' characteristic aspects, although it was anticipated that *Soft SF* would deal with the topic less due to its general scientific inaccuracy. Another similarity is the concept of future with related keywords such as *"становится"* ("become"), *"новый"* ("new"), *"будущее"* ("future") in Hard fiction and *"будущее"* ("future"), *"идти"* ("advance") for Soft fiction. This may indicate an accustomed tendency to look forward to the unknown and anticipate the future. Another two universal concepts for this "couple" are time and humanity. As for the first concept, the keywords seen in the annotations are identical in both genres: *"год"* ("year") and *"время"* ("time"). The concept of humanity and its perseverance is unfolded with the help of the following keywords: *"человек"* ("human"), *"раса"* ("race"), *"деятельность"* ("endeavour") in Hard fiction and *"человек"* ("human"), *"жизнь"* ("life"), *"смочь"* ("manage") in Soft fiction. However, there is a topic that is exclusive to *Hard SF*: technological progress. The predominant keywords are *"робот"* ("robot") and *"корабль"* ("spacecraft"), which are not seen in *Soft SF*. This may be connected with the fact that *Hard SF* is usually more scientifically accurate and is related to so-called "hard" sciences such as physics, mathematics, engineering etc. *Soft SF*, on the other hand, is generally related to so-called "soft" sciences such as sociology or psychology.

To summarize, we may say that *Hard SF* and *Soft SF* share a vast amount of similarities and interconnections, which can be perceived as both a sales pitch (considering that we researched the annotations, which often make one's mind when purchasing) and a shared literary pattern. [14], [15]

The second "couple", which consists of *Space Opera* and *Timepunk*, was labelled as such due to one but major similarity: humanity and machines. Both share similar keywords such as *"машина"* ("machine"), *"человек"* ("human"), *"война"* ("war") for *Space Opera* and *"война"* ("war"), *"машина"* ("machine"), *"человек"* ("human"), *"солдат"* ("soldier") for *Timepunk*. This may have to do with the fact that both genres are generally connected to glorious adventures and space warfare, which usually involve humans and other species (often machines or robots). Apart from this notable similarity, there are none other, so it is reasonable to talk about the genres separately at this point. [16], [17]

Space Opera exhibits rather restless behaviour, which unfolds with the help of the following keywords: *"война"* ("war"), *"победить"* ("win"). *"легион"* ("legion"), *"победа"* ("victory") and *"грабить"* ("rob"). This may be connected with the above-mentioned militaristic and adventurous tendency of the genre, which was anticipated. *Timepunk*, on the other hand, seems entirely disconnected from the concept of time or timepunks, so the genre does not live up to general expectations considering its name. *Timepunk* is a derivative of *Steampunk*, which examines retrofuturistic aesthetics and is not related to time or timepunk whatsoever. [18], [19]

To sum it up, the two genres share a major similarity – militarism, which is a common thread in both of them. However, *Space Opera* is generally more adventurous and militaristic, whilst *Timepunk*, being derived from *Steampunk*, is more about advanced technologies and their aesthetics.

At times one may feel like the name and the essence of an object are reasonably interconnected but there is no obvious way to prove it. Topic modelling has proved to be useful when resolving such uncertainties. For instance, *Cyberpunk* lives up to its name and follows the topic of digital surrealism, which unfolds with the help of the following keywords: *"виртуальный"* ("digital" / "virtual"), *"реальность"* ("reality"), *"мир"* ("world"), *"становится"* ("become") and *"игра"* ("game"). Moreover, we have found out that the digital world of *Cyberpunk* is often complicated and disturbing: *"пересадка"* ("transplantation"), *"секретный"* ("secret"), *"утечка"* ("exposure"), *"прознавать"* ("find out"), *"смертоносный"* ("deadly"), *"убийца"* ("murderer"). This may be connected with the fact that *Cyberpunk* often examines conflicts between hackers and corporations, as well as bits of artificial intelligence. [20]

Likewise, *Cyberpunk*, *Catastrophe* and *Timepunk* science fiction live up to their names. The *Catastrophe* books' annotations contain keywords such as *"катастрофа"* ("catastrophe"), *"аберрация"* ("aberration"), *"поиск"* ("search") and *"жизнь"* ("life"). As for *Timepunk*, we have found the following keywords: *"время"* ("time"), *"год"* ("year"), *"прошлое"* ("past"), *"будущее"* ("future") and *"история"* ("history"). This is of course connected with the fact that *Timepunk* fiction centres on transtemporal travel and its consequences. [21]

To summarize, all three of the above-mentioned genres are indeed what they seem like and due to topic modelling we were able to prove it statistically and not just through perception.

The analysis of *Utopian* and *Dystopian* annotations proved to be the most interesting (Figure 6 c&d). Both genres are often perceived as opposites, however, we were able to detect more similarities than differences. Both genres examine the questions of existence and future: *"цивилизация"* ("civilization"), *"человек"*, ("human"), *"будущее"* ("future"), *"решать"* ("decide"), *"понимать"* ("understand") in *Dystopia* and *"будущее"* ("future"), *"время"* ("time"), *"век"* ("century"), *"человек"* ("human") and *"становится"* ("become") in *Utopia*. Furthermore, the concept of freedom and hope is relatable for both genres too: *"свободный"* ("free") and *"дух"* ("spirit") in *Dystopia* and *"молодой"* ("young") and *"мечта"* ("dream") in *Utopia*. Finally, the concept of general well-being and greater good is similar in both genres as well: *"земля"* ("Earth"), *"охранять"* ("defend") in *Dystopia* and *"земля"* ("Earth"), *"освоение"* ("exploration") in *Utopia*.

Nevertheless, there are visible differences between the two genres. For instance, the concept of coexistence is seen antagonistically in both genres: *"мир"* ("peace") in *Dystopia* with no mention of *"война"* ("war") and strong presence of *"война"* ("war") in *Utopia*. Another curious difference between *Dystopia* and *Utopia* is that the concept of *"смерть"* ("death") is seen in *Utopia* exclusively.

To summarize, the similarities and differences of *Dystopia* and *Utopia* may be connected to the common thread of these genres, which is of course society. The issue is that society is viewed differently in each genre, however, they both share the same endeavours and hopes for its future – prosperity and life, even though in opposite ways. [22], [23]

In conclusion, we may say the annotation analysis proved to be enlightening in terms of applied concepts and ideas. Moreover, we have found out that the coined "couples" share a great number of characteristics, which at times were unexpected as in the case with *Dystopia* and *Utopia*, for instance.

6 Conclusions & Future work

Thus, in the course of the study, we came up with the following conclusions. Firstly, as we suspected, the archive is biased towards works in Russian or translated into Russian. In addition, a significant proportion of the users who contributed to the formation of metadata are men. Machine learning analysis of metadata has shown that automatic classification can be applied to at least some genres. The most suitable characteristics for classifying subgenres are location, and plot goal. The time of action and the linearity of the plot, however, turned out to be insufficient to determine the belonging of work to a particular subgenre. An analysis of the frequency of words in annotations, in turn, showed an uncertain result: for some subgenres, such as *Hard SF* and *Cyberpunk*, the frequency of words provides sufficient grounds for their classification; for others, such as *Utopia* and *Dystopia*, on the contrary, it does not clearly distinguish between subgenres.

At the same time, topic modelling has shown promising results.

In the future, we would like to expand our research towards the study of user behaviour to see the tone of annotations, how the rating of the work is related to metadata, and how the gender and age of commentators correlate with books' metadata.

In addition, it would be interesting for us to conduct a similar analysis of the Fantasy genre since the FantLab also stores an extensive collection of data on this genre.

Furthermore, our research can serve as a basis for creating an automatic classifier of fiction books based on metadata and annotation texts.

Acknowledgements

The research was inspired by the ideas discussed in the course "Introduction to Cultural Data Mining and Visualization" within the Master program "Data, Culture, and Visualization" at ITMO University.

References

[1] Scott, R.: Digital Humanities and Open Access: An Interview with Brett Bobley of the National Endowment for the Humanities. Oct 24, 2014, http://www.righttoresearch.org/blog/digital-humanities-and-open-access-an-interview-wi.shtml (last accessed: 01.09.2021)

[2] Science Fiction Foundation (SF Foundation), URL: https://www.sf-foundation.org/ (last accessed 03.09.2021)

[3] The Internet Speculative Fiction Database, URL: http://www.isfdb.org/ (last accessed 03.09.2021)

[4] The Encyclopedia of Science Fiction (SFE), URL: http://www.sf-encyclopedia.com/ (last accessed 03.09.2021)

[5] NooSFere, URL: https://www.noosfere.org/ (last accessed 03.09.2021)

[6] Moretti F., Conjectures to World Literature, in F. Moretti, Distant Reading, Verso, London & New York, 2013, pp. 43-62.

[7] Forlini, S., Hinrichs, U., Moynihan, B.: The Stuff of Science Fiction: An Experiment in Literary History, 2016. URL: http://www.digitalhumanities.org/dhq/vol/10/1/000228/000228.html

[8] Menadue, C.B.J.: Science Fictions, Cultural Facts: a Digital Humanities Approach to a Popular Literature. PhD thesis, James Cook University, 2019, https://doi.org/10.25903/5ef01cb4754df

[9] Blei, D.M.: Probabilistic Topic Models. Communications of the ACM, 55(4), 2012, pp. 77-84, https://doi.org/10.1145/2133806.2133826

[10] Github – FantLab API Description. URL: https://github.com/FantLab/FantLab-API (last accessed: 01.09.2021)

[11] Web Scraper – The #1 web scraping extension, https://www.webscraper.io/ (last accessed: 01.09.2021)

[12] Bekisheva, K.: GitHub DH Project: Files for a data mining project 'How do people read science fiction and why it is popular: common tendencies and comparative analysis', 2021, URL: https://github.com/morin-berk/DataMining

[13] Agafonova, K.: Kaggle – FantLab Library, 2021,
URL: https://www.kaggle.com/karinaagafonova/fantlab-library

[14] Science Fiction Definition – The Encyclopedia Britannica,
URL: https://www.britannica.com/art/science-fiction (last accessed: 31.08.2021)

[15] Stableford, B.M.: Science Fact and Science Fiction: an Encyclopedia. Routledge – Taylor & Francis, 2006, p. 227

[16] Pringle, D.: What is this Thing Called Space Opera? Space and Beyond: The Frontier Theme in Science Fiction (ed. G. Westfahl), 2000, pp. 35-47.

[17] Daller, J.X.: Dierchomai Dystopia. Dunkle Visionen & Lichte Bilder. GRIN Verlag, 2011, pp. 29-119.

[18] Clarke, I.F.: Future-War Fiction: The First Main Phase, 1871-1900. Science Fiction Studies, 24(3), Nov. 1997, pp. 387-412.

[19] Latham, R.: The Oxford Handbook of Science Fiction. Oxford University Press, 2014, p. 439.

[20] Graham, St.: The Cybercities Reader. Routledge: New York, 2004, p. 389.

[21] Nahin, P.J.: Time Machines: Timepunk in Physics, Metaphysics, and Science Fiction. (2nd ed.), Springer, 2001, 628 p.

[22] Gier, E.: Between “Utopia” and “Dystopia”. OpenEdition Journals, Vol. 17, 2020, URL: https://doi.org/10.4000/nrt.7496 (last accessed 09.09.2021)

[23] Macleod, G., Ward, K.: Spaces of utopia and dystopia: landscaping the contemporary city, Geografiska Annaler: Series B, Human Geography, 84(3-4), 2002, pp. 153-170, URL: https://doi.org/10.1111/j.0435-3684.2002.00121.x (last accessed 09.09.2021)