Languages Cities and People Speak

Anatoly Kozlov[♠] (kozlov.aa@phystech.edu),
Dmitry Rodin[♠] (rodin.dv@phystech.edu),
Alexey Petrov[♠] (petrovalex_99@mail.ru),
Alexey Karpov[♠] (alexeykarpov98@gmail.com),
Han Nguen[♠] (nguen.thi@phystech.edu),
Ilya Raskin[♠] (Raskin.II@phystech.edu),
Alexandr Ryashchikov[♠] (SanhiseR@mail.ru),
Nikolay Pak[♠] (pak_kolya98@mail.ru),
Dmitry Ivanov[♠] (dmitriy250998@gmail.com),
Anastasia Moiseeva[♠] (moiseeva.aa@phystech.edu),
Lyudmila Rvanova[♠] (rvanova.lyu@phystech.edu),
Tatiana Shavrina[†,♡] (rybolos@gmail.com),

[♠]Moscow Institute of Physics and Technology, Moscow, Russia
[†]Sberbank, Moscow, Russia
[♡]National Research University Higher School of Economics, Moscow

The web is a rich platform for collecting natural language data and corpus construction, but still, web corpus resources remain quite complex for a wide range of users. To make the use of the corpus more suitable for a wide audience we provide a new project, based on the data from the General Internet Corpus of Russian Language. The interface provides users to explore the variety of word distribution - by age, gender and regions of the world, to look up word usage and trends, check their texts on different kinds of authorship features and share the visualizations. We have also conducted experiments on texts classification by age, gender and location using data from the social media segment of the corpus. All of the algorithms are integrated into one clear web interface.

Key words: machine learning, corpus linguistics, age classification, gender classification, region classification

Языки Городов и Людей

Анатолий Козлов[♠] (kozlov.aa@phystech.edu),
Дмитрий Родин[♠] (rodin.dv@phystech.edu),
Алексей Петров[♠] (petrovalex_99@mail.ru),
Алексей Карпов[♠] (alexeykarpov98@gmail.com),
Хань Нгуен[♠] (nguen.thi@phystech.edu),
Илья Раскин[♠] (Raskin.II@phystech.edu),
Александр Рящиков[♠] (SanhiseR@mail.ru),
Николай Пак[♠] (pak_kolya98@mail.ru),
Дмитрий Иванов[♠] (dmitriy250998@gmail.com),
Анастасия Моисеева[♠] (moiseeva.aa@phystech.edu),
Людмила Рванова[♠] (rvanova.lyu@phystech.edu),
Татьяна Шаврина[†,♡] (rybolos@gmail.com),

[♠]Московский физико-технический институт, Москва, Россия
[†]Сбербанк, Москва, Россия
[♡]НИУ Высшая Школа Экономики, Москва, Россия

Интернет позволяет собирать большое количество информации о естественных языках и строить корпуса, но все еще большинство корпусных ресурсов остаются слишком

сложными для большого числа пользователей. Для того чтобы сделать корпус более удобным для широкого круга людей мы представляем свой проект, основанный на Генеральном Интернет Корпусе Русского Языка. Интерфейс предоставляет возможность наблюдать множество словесных распределений - по полу, возрасту и местоположению, искать словоупотребления и словесные тренды, проверять тексты на различные авторские признаки и визуализировать полученные данные. Мы также провели эксперименты по классификации по полу, возрасту и региону по данным из сегмента корпуса, относящегося к социальным сетям. Все полученные алгоритмы интегрированы в один понятный интерфейс.

Ключевые слова: машинное обучение, корпусная лингвистика, классификация по возрасту, классификация по полу, классификация по региону

1 Introduction

Corpora have proved to be a very useful instrument in natural language analysis. And with the presence of the Internet and the social network, it has become possible to obtain large amounts of annotated data and to constantly keep it up-to-date. A common problem with the Web Corpora is that they usually have complicated APIs and search interfaces, so some of the researchers may have problems with using the data to its full potential, and some specialists from related fields of study - social media specialists, marketers, interested web search amateurs - thus are not considering corpora as a proper instrument. In this paper, we are trying to solve some of such problems by creating a web service capable of making requests to General Internet Corpus of Russian (GICR) and applying machine learning algorithms to the data obtained.

2 Project Goals and Functionality

The main goal of our project "Languages of Cities and People"[1] is to make obtaining corpus statistics less complex, so a wider range of people can use the natural language data for their researches, and to make a clear picture of word frequency distribution available to the general public. To achieve this we created a web service with a simple interface which does not require fine-tuning but is still able to provide reliable results. The frequency dictionaries with differential features can help to estimate the most used word and even help predict social trends.
The secondary goal of our project is to estimate authorship attribution by classifying text by the author's gender, region, and age. For this purpose, we have collected data from the social media segment of the GICR and conducted several experiments described in chapter 5. All of the results obtained are integrated into the interface, so they can be used by the general public to get the authorship features of their texts.

[1] https://int.webcorpora.ru/yalg/

Figure 1: Search example for regional frequency statistics

3 The Data

The project is based on the data provided by General Internet-Corpus of Russian [1], including main Russian social media sources: VKontakte[2](VK) and Live Journal[3](LJ). During the project creation, two databases containing word frequency distribution on every combination of parameters (authors' location, gender and age) were obtained on the whole amount of data available in the corpus (10 and 9 billion words each) – see example 1.
Example 1.

- *2 проживание N:"gender=M""genrei=blog""loc=Оренбург""loc=Оренбургская область""loc=Россия""month=10""rule=vk post""source=vk""year=2014"*
- *1 санька N:"gender=M""genrei=blog""loc=Архангельская область" "loc=Мирный""loc=Россия""month=08""rule=vk post""source=vk""year=2014"*

For such tasks as the classification and analysis of trends over time, we also needed original text materials - a fragment of the Vkontakte segment (100 million words) was used for the tasks of authorship attribution - age, gender and region classification.
Social media data needs to be updated frequently enough to capture major trends and changes in it, while corpus projects are usually subject to the reverse trend - due to the large volume, needs of the technological chain and labor-intensiveness of re-indexing, they often form a frozen cast of language over time. The GICR data had to be updated, and therefore the original corpus was supplemented with additional texts, collected from the most prominent group of authors: popular bloggers and original content writers.

3.1 Data Preprocessing

The texts of a large web corpus often cause many problems caused by a large number of accumulated "garbage" - technical tokens, automatically generated text, typos, preprocessing errors, very rare words. Because of this, the distribution of the law digit can be violated [3] In the case of our material - databases of word frequencies by city, gender and age - the situation with rare tokens and errors becomes critical: most of our frequencies are very rare - such a frequency dictionary does not at all comply with Zipf's law [5]- see Figure 2

[2]https://vk.com/
[3]https://www.livejournal.com/

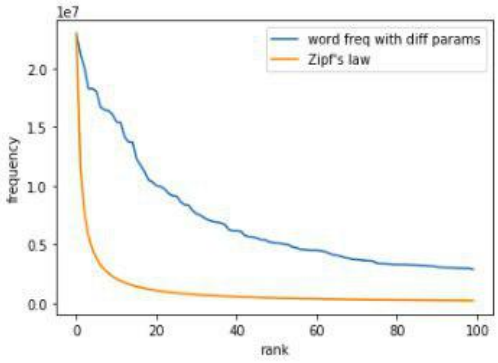

Figure 2: Zipf law distribution vs distribution of words with differential features (see example 1)

To avoid the visualization of insufficient data, the following measures were taken to filter the data:

- filtering from NaN values
- filtering from invalid toponyms

The procedure for the unification of place names was as follows: for each place tag the nearest string in lookup database was found, with the help of geonames library this string was converted to the proper unified region and city name.
Authors' gender and age was taken for granted, which, however, created some difficulties (see chap 4).
In addition to working with frequencies on the whole corpus, we also needed to solve the problem of data cleansing for the original texts collected by us from the aforementioned sources. For these purposes, deduplication was carried out through Onion pomikalek2011removing, and the texts were lemmatized using the pymystem3 python library[4].

4 Frequency Distribution of Russian Words

We proceed from the hypothesis that the distribution of words according to differential features can tell both about the language as a whole and be used to study the generalized features of social groups. The corpus must be large enough, and the distribution of important parameters in it should not be contrary to common sense. In our data, the age distribution of authors is close to normal - see Figure 3. It is easy to see that there are no authors aged less than 12, but there is a peak about of this age - according to the rules, users under 12 are not allowed to use the resource, and they set the minimum possible age to register. There are also outliers considering authors who set themselves as extremely old.

[4] https://github.com/nlpub/pymystem3

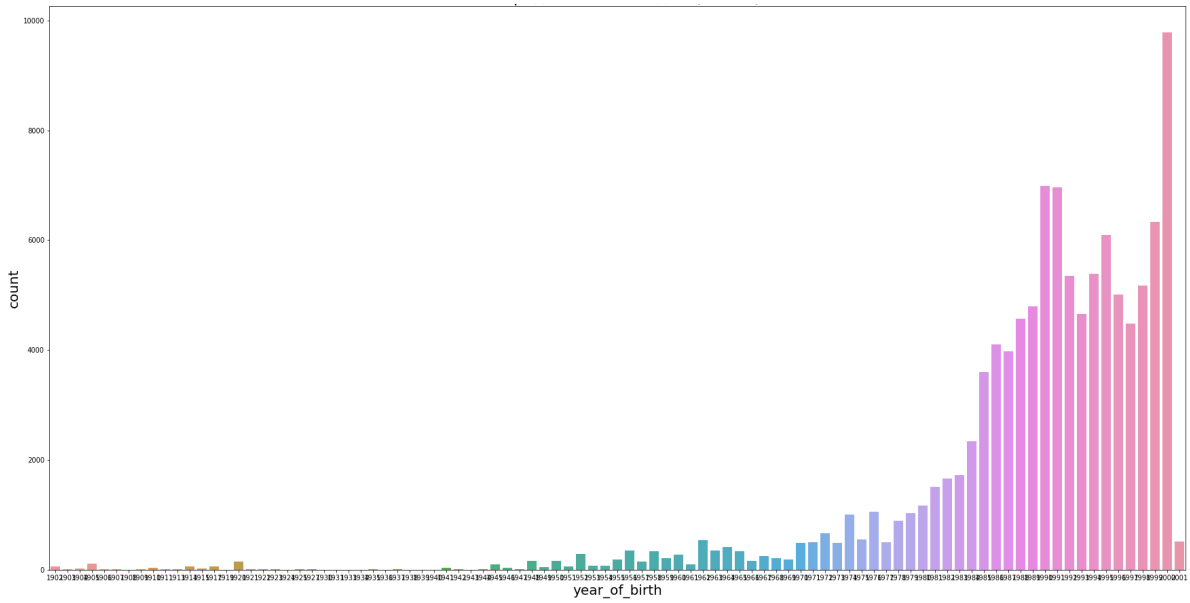

Figure 3: Text by year of author's birth distribution

To obtain the words that can be considered reliable features for the tasks of authorship attribution, we we conducted a frequency analysis of our materials using the following procedure:

- the word and its frequency distribution by a parameter (gender, age, region) is obtained;
- it can be assumed that the frequency distribution of the word in the parameter should be uniform, but in reality, there may be outliers;
- using Pearson's criterion, one can find words for which the uniform distribution hypothesis should be rejected;
- by the largest difference between the observed and critical values, we can sort the words with the strongest unevenness in the non-uniform distribution.

Using this simple test, heavily based on the work of [2] we can observe rather interpretable results - see Table 1 and 2 for gender and age features.

Table 1. Most significant age features according to Pearson's criterion

Group	**Key words**
Age < 17	хлопнуть, коментарий, джефф, сканер, бали, спускаться, гдз, лера, сургут, пту, проигравший, майнкрафт, дебил, яна, хлопчик, грузчик, арина, монстер, настроить, даша, ржать, мега-низкий, фокси, прикольный, билетик, разрешать, кругом, вормикс, опубликованный, противоправный, полина, настя, наручный, вже, диана, конвенция, граффити, аниматроник
Aged 18-30	лимузин, фотосессия, студийный, казино, свадебный, клин, юбилей, ростов, интерфейс, элитный, оформление, биг, зеленоград, опрос, шар, съемка, свадьба, мгновенно, спб, конкуренция, оформлять, курортный, организовать, удобство, прокат, ленинград, оператор, очаг, таковой, сниженный, недорогой, текила, круглосуточный
Aged >30	достаток, таможенный, пропорциональный, родовой, справочник, регулирование, подготовить, задолженность, подкормка, гравитационный, сформулированный, менеджмент, геймер, эйнштейн, плотность, настойка, физик, простота, савченко, залог, посев, якорь, дебальцево, петиция, сруб, утверждение, кандидат, коломойский, антимайдан, специальность, кайрат, отопление, привилегия, перемирие, захарченко, фашист

Table 2. Most significant gender features according to Pearson's criterion

Group	Key words
Male	автомат, автомобиль, аккаунт, алексей, аллах, армия, аэропорт,боец, борьба, босс, браузер, бугор, возраст, вообще, вопрос, вспомнить, встроенный, гоблин, жим, заем, заинтересованный, заценивать, защитник, заявка, звено, одержать, окоп, ополченец, оружие, очки, панфиловец, партнерский, пассив, перешедший, пиар, победить, прийнявый, прогресс, противник,процентный, работать, рад, разрешать, разрешение, рана,ремонт, репутация, речь, розыск
Female	блин, блюдо, бог, богатство, бровь, бросать, букет, булгаков, булочка, бульон, бумага, бутик, вареный, великолепный, вера, верность, вернуться, верхний, вес, веселый, весенний, вечерний, вечный, вещество, вещь, взаимный, взамен, вздыхать, вид, животный, жизнь, заболевание, забота, забывать, забыть, зависеть, зависимость, завтра, завтрак, зодиак, золотистый, зона, зря, зубчик, игрушка, идеально

5 Classification Experiments

This section provides information about the experiments conducted: prototypes of these solutions will be included in the functionality of the service for checking arbitrary texts.

5.1 Gender Classification

Two classification experiments were conducted: on freshly collected texts and on full corpus data. For each word we calculated statistics of usage in male and female texts using word keyness on corpora[5]. In both cases logistic regression classifier with no regularization was used. Best results on Vkontake data show 87% accuracy on binary gender classification.

5.2 Location Classification

The task of classifying locations is traditionally reduced to several levels: the classification of a country, a region, and a city. In the case of VKontakte data, it turned out to be much easier to determine exactly the city, since within the CIS countries and the world the Russian language does not show as much variation on our data as it does in the regions of Russia (the exception is Ukraine, which is well defined). Resulting from the classification of author city by text, we achieved an accuracy of 40% on 52 classes.

5.3 Age Classification

Authors' age classification appeared to be one of the most laborious task – as younger people show the most fruitful online behavior writing posts, the only successful solution was to divide all authors into three isometric categories - children (under 17), young (18-30) and middle-aged (after 30). Children and the elderly differ quite strikingly from the whole category of middle age, whereas middle age authors, perhaps due to their superior amount, show a very large variety of topics and word usage, making smaller age strata problematic to classify.

[5] https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf

6 Future work and discussion

This project for the first time makes it possible to get a clear frequency picture of the Russian language to the general public - we can tell that this project can be explored both for entertainment purposes: visual trend analysis, social media research, simple authorship attribution hypothesis ;generalization, etc, and both for the primary checks that do not claim 100% academic accuracy, testing for such simple hypotheses as gender bias of a word, a certain age category bias, and even regionality of vocabulary.
As part of further work, a more accurate study of various signs of statistical bias is planned - combinations of locations and gender, gender and age group, the number of authors and the number of texts from the region, and so on. A more thorough study of trends in social networks is also planned, on an extended sample of LJ and other sources.
Also, one of the areas of research may be the search for jointly biased phrases and expressions that occur in the same authors' conditions.
Summing up the results of studies on classification, it is worth noting a rather strong coherence between the themes of the corpus and the genders of the authors that correlate with them. Potentially, such a model conclusion may interfere with the full-fledged work on the analysis of transmitted texts. The quality of age classification turned out to be quite low - the most distinctive groups on the basis of their language behavior are children and the older generation, while the texts of people of young and middle age are more homogeneous. Regional classification remains one of the promising areas of work - additional research is needed on the clustering of texts from close regions, however, the first 3 classification predictions give the 40% accuracy in such a complex task.
We invite all interested - linguists, developers, as well as philologists and just enthusiasts - to try out our new tool in their research and visit our Github. [6]

References

[1] Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S. (2013), *Big and diverse is beautiful: A large corpus of Russian to study linguistic variation*. In Proc. Web as Corpus Workshop (WAC-8).

[2] Kuratov Yu., Lagutin M., Kopylov N. (2016) Statistical processing of Search results in differential corpora. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" Moscow, June 1–4, 2016

[3] Moreno-Sánchez I, Font-Clos F, Corral Á (2016) Large-Scale Analysis of Zipf's Law in English Texts. PLoS ONE 11(1): e0147073. https://doi.org/10.1371/journal.pone.0147073

[4] Pomikalek J.(2011) Removing boilerplate and duplicate content from web corpora, Ph.D. thesis, Masaryk University Faculty of Informatics, Brno

[5] Zipf G.K.(1949) Human Behavior and the Principle of Least Effort. — Addison-Wesley Pres. — c. 484-490. — 573 c.

[6] https://github.com/Anat37/yalg2019