Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

"Национальный исследовательский университет "Высшая школа экономики"

Факультет Гуманитарных Наук

Школа Лингвистики

КУРСОВАЯ РАБОТА

Компьютерно-лингвистическое исследование
научно-популярных текстов

Студент группы № МКЛ 171

Лапидус Анна
Кузнецова Анастасия
Коломенская Юлия
Самойленко Ксения

Научный руководитель
канд. филол. наук,
доц. Б.В. Орехов

Москва, 2018г.

1. Введение

Начиная с 2010-ых годов в русскоязычном медиа-пространстве наблюдается рост интереса аудитории к научно-популярным и просветительским проектам. Помимо научных разделов на традиционных новостных сайтах[1] появляются самостоятельные ресурсы, которые становятся альтернативой традиционным («журнальным») научно-популярным изданиям («Вокруг света», «National Geographic», «Популярная механика», «Техника — молодёжи» и др). Кроме того, научно-популярные проекты выходят за пределы интернет-среды: растет количество просветительских мероприятий и лекториев[2], успехи в популяризации отмечаются специальными премиями, в том числе государственной «За верность науке», а в академической среде и просветительском сообществе ведутся дискуссии о методах популяризации, ее задачах и перспективах[3].

Будучи заметным общественным явлением, популяризация имеет и свое языковое измерение. У читателей и исследователей появился доступ к большому числу текстов на русском языке, посвященных разным областям научного знания, но не представляющих собственно академический жанр. Насколько справедливо будет говорить об особом языке популяризации и стилевых особенностях научно-популярных текстов? Мы попробуем ответить на этот вопрос методами компьютерной лингвистики.

Цель исследования — проанализировать корпус научно-популярных текстов на русском языке при помощи актуальных инструментов компьютерной лингвистики. Мы выбрали четыре направления работы:

1) Поиск и извлечение научных терминов, встречающихся в научно-популярных текстах, и составление словаря на основе извлеченных понятий. Наличие такого словаря позволяет оценить, о чем чаще всего говорят в научно-популярных текстах, а также частоту встречаемости сложных терминов — это может стать важной характеристикой “популяризаторского” стиля, отличающего его от академического, публицистического и разговорного жанров.
2) Поиск и извлечение именованных сущностей-имен ученых. Это позволит составить список наиболее часто цитируемых и упоминаемых ученых в научно-популярной среде и понять степень значимости таких упоминаний для текстов научно-популярного жанра.

---

[1] Lenta.ru: https://lenta.ru/rubrics/science/, РИА Новости https://ria.ru/science/, Газета.ру https://www.gazeta.ru/science/

[2] Приведем не претендующий на полноту список: лектории АРХЭ и “Курилка Гуттенберга”, фестивали GeekPicnic, Science slam, Pint of Science, Ученые против мифов, лекции, организуемые премией “Просветитель”, издательствами научно-популярной литературы и университетами

[3] В качестве примера можно привести статью социолога Виктора Вахштайна, вышедшую на портале Индикатор после его выступления на “Слете просветителей” https://indicator.ru/article/2017/12/13/viktor-vakhshtayn-slyot-prosvetiteley-2017/

3) Определение тематической близости текстов разных жанров и из разных источников. Научно-популярные порталы по-разному делят свои материалы по принадлежности к какой-либо области знания; мы попытались разработать универсальную рубрикацию и сверить тематическую близость текстов, отнесенных к различным темам.
4) Измерение удобочитаемости текстов. Научно-популярные тексты бывают очень неодинаковыми по уровню сложности: какие-то понятны и людям, не окончившим школу, а какие-то требуют специального образования в данной области. Нам было интересно понять, насколько разнятся по сложности научно-популярные тексты и разработать способ их классификации по уровням удобочитаемости и понятности.

Поставленные задачи мы планируем решать посредством создания размеченного корпуса, его обработки и написания соответствующего программного обеспечения на языке программирования Python. Мы предполагаем, что данное исследование важно не только для исследования феномена научно-популярных текстов, но и может быть полезным в решении прикладных задач для самих научно-популярных ресурсов и послужить их улучшению и оптимизации.

1.2. Данные для корпуса

Мы собрали тексты из шести научно-популярных русскоязычных ресурсов. При создании корпуса мы обращали внимание на жанры материалов: нам было важно получить расшифровки лекций как образцы прямой речи ученых, тексты статей, интервью, текстов блогов и новостей. Мы отказались от включения отрывков из книг, так как они часто были переводными, а значит, не вполне соответствовали нашим целям, а также игровых и микро-форматов, так как они обычно являются разработкой редакторов и не несут интересующей нас информации.

«N+1»
Ресурс специализируется на научных новостях. Создан в 2015 году. В основном освещает вопросы точных и естественных наук, иногда рассказывая о лингвистических и археологических исследованиях. Основатель и издатель — Андрей Коняев, кандидат физико-математических наук, преподаватель мехмата МГУ им. М. В. Ломоносова.[1]

«ПостНаука»
Интернет-журнал, публикует лекции и статьи ученых, освещает все области научного знания. Создан в 2012 году. Сооснователь и издатель — Ивар Максутов, кандидат философских наук.

«Чердак»

Новостной портал, создан на базе информационного агентства ТАСС при поддержке Министерства образования и науки Российской Федерации в 2014 году. [2] Издание специализируется на публикациях о точных науках, современных технологиях и медицине. Руководитель проекта — Егор Быковский, в прошлом — главный редактор и редактор журналов «Наука в фокусе», «Вокруг света». [3]

«Geektimes»
Раздел проекта «Хабр», специализирующийся на новостях и статьях о современных технологиях. Выделился в 2014 году.[4]

«Полит.ру»
Новостное СМИ, создано в 1998 году. На сайте есть отдельный раздел PRO Science, где публикуются научные новости и расшифровки лекций ученых.

«Индикатор»
Новостной портал, специализирующийся на новостях российской науки, вопросах организации науки и взаимодействия в рамках научного сообщества. Создан в 2016 году. Главный редактор — Николай Подорванюк, астрофизик, кандидат физико-математических наук. [5]

Всего корпус составил **30 000** текстов. Для каждой из задач данные были необходимым образом предобработаны и размечены. Инструменты и принципы разметки будут описаны ниже в каждом соответствующем разделе.

1.3 Рубрикация текстов

Поскольку рубрики на взятых нами ресурсах очень различаются, а иногда рубрикации нет совсем, нам необходимо привести весь корпус научно-популярных текстов к единой рубрикации наук.

Существуют различные классификации областей науки, например, Библиотечно-библиографическая классификация (ББК), Универсальная десятичная классификация (УДК), Государственный рубрикатор научно-технической информации (ГРНТИ). Каждый из этих классификаторов предоставляют полную иерархическую систему областей знаний, однако для рубрикации научно-популярных текстов эти классификации оказались слишком подробными. При этом в собранном корпусе некоторые области науки представлены шире, чем другие, что также хотелось учесть в рубрикации текстов. В поисках классификации, которая больше соответствовала бы тематической структуре корпуса, был проведен анализ рубрикаций на разных научно-популярных ресурсах. Большинство источников не предоставляют единую

рубрикацию наук, а используют множество меток для указания тематики статьи. Наиболее полная классификация представлена на ресурсе ПостНаука (https://postnauka.ru/), она была взята за основу для рубрикации корпуса. Мы переработали классификацию с учетом особенностей корпуса: были добавлены рубрики *Технологии*, *Науки о Земле*, *Computer Science* и *Футурология*, “Медицина” и “Мозг” были объединены под рубрикой *Физиология человека,* некоторые рубрики были переименованы (“Астрономия”, “Право”). В результате нами была предложена следующая классификация научно-популярных текстов по областям:

1. Космос
2. Психология
3. Язык
4. Экономика
5. Биология
6. Политология
7. Культура
8. Философия
9. История
10. Социология
11. Математика
12. Химия
13. Физика
14. Физиология человека
15. Технологии
16. Computer Science
17. Науки о земле
18. Футурология (рубрика для текстов, посвященных описанию различных аспектов будущего).

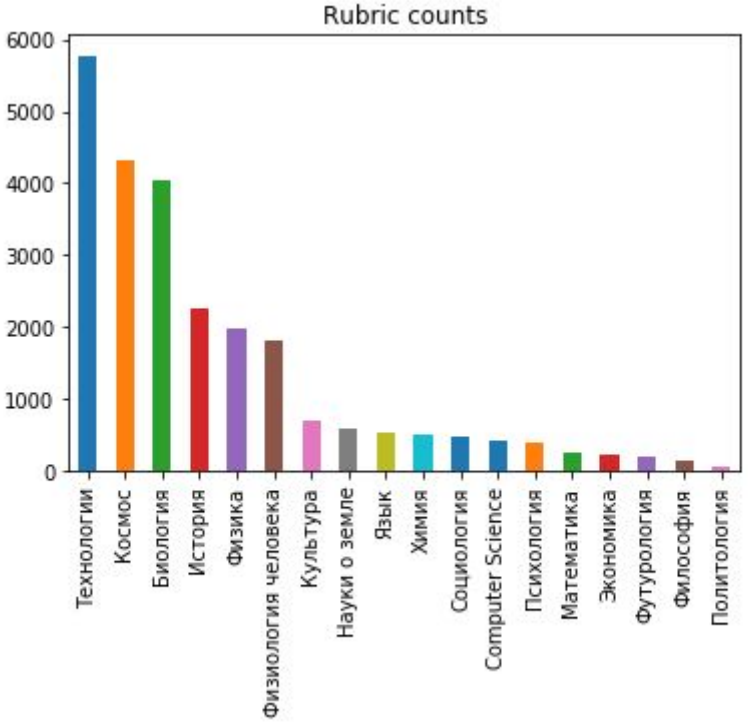
С помощью сопоставления тематических меток, присвоенных текстам на интернет-ресурсах, и предложенных рубрик мы получили таблицу соответстветствий для приведения текстов к единой рубрикации. Далее каждому тексту корпуса была автоматически присвоена рубрика, к которой относится большинство меток этого текста. Если в результате такого сопоставления по большинству меток для текста

оказались равнозначными несколько рубрик, то для выбора единственной рубрики применялись дополнительные правила соответствия.

Правила соответствия включали в себя правила, основанные на взаимосвязи некоторых наук. Например, *Физиология человека* является частью *Биологии*, следовательно для такого сочетания рубрик было задано правило: если тексту соответствуют рубрики *Физиология человека* и *Биология*, то присвоить тексту рубрику *Физиология человека*. Другим примером могут послужить сочетания различных рубрик с рубрикой *Космос*: т.к. исследования космоса связаны с разными областями (*Физика*, *Химия*, *Технологии* и т.д.), то в случае таких сочетаний задавались правила присваивания текстам рубрики *Космос*.

Некоторые правила являются специфичными для заданного корпуса текстов. Например, в результате анализа текстов корпуса было выявлено, что тексты с сочетанием рубрик *Химия* и *Физиология человека* в большинстве случаев посвящены химическим процессам, протекающим в организме человека, следовательно в качестве единственной рубрики выбирается *Физиология человека*.

В результате с помощью таблицы соответствия меток с интернет-ресурсов рубрикам и дополнительных правил каждому тексту была сопоставлена единственная рубрика из предложенной классификации научно-популярных текстов. Полученное соотношение рубрик по количеству текстов представлено на диаграмме ниже.



2.1 Изучение цитирования ученых в научно-популярных текстах

Исследовательский вопрос

Одна из задач нашего исследования -- понять, кого из ученых больше всего упоминают авторы научно-популярных текстов, узнать, в каких областях науки их имена употребляются больше, а в каких меньше, для каких жанров текста характерно большее количество упоминаний ученых. Мы попытаемся составить рейтинг ученых на основе собранного корпуса научно-популярных текстов, а также сравнить получившиеся результаты с официальными рейтингами цитирования.

Однако для того чтобы исследовать этот вопрос, нам необходимо научиться извлекать из текстов имена ученых. В настоящем тексте мы приведем лишь краткий обзор задачи извлечения именованных сущностей.

Извлечение именованных сущностей (*Named Entity Recognition)* -- классическая задача компьютерной лингвистики, которая подразумевает выделение в тексте имен собственных: названий организаций, имен людей, названий географических объектов. В настоящее время в области распознавания именованных сущностей появляются все новые методы, что связано, в том числе, и с развитием нейронных сетей, алгоритмов машинного обучения, широким использованием различных типов векторного представления слов.

State of the art в распознавании именованных сущностей считается NER-теггер, разработанный Lample et al. в Стенфордском университете. Изначально теггер работал на выбранных вручную признаках *(features)*, затем система была улучшена CRF и LSTM и обучена на неразмеченных корпусах четырех крупных языков: английского, французского, испанского и немецкого. Стендфордская LSTM-CRF модель работает с символьно-векторным представлением слов, полученных при обучении на размеченных корпусах текстов.

Главными источниками данных для разметки именованных сущностей являются размеченные и неразмеченные корпуса текстов: Википедия, Web, CONLL, Onto, Tweet и др.

2.3 Подготовка данных

Чтобы выделить имена ученых из текстов корпуса, было необходимо выяснить, в каких контекстах они встречаются. Вручную было размечено 167 текстов. В размеченных данных обнаружилось 3 тыс. контекстов, где встречаются имена ученых. Далее вручную из этих контекстов отбирались наиболее характерные и наиболее частотные. Например:

```
# X (и У) предположил(и) (что)
# ученые X и У  | занимались
# X и У сыграли роль в … / изучали (изучили)
# разработал (ADJ) специалист из (…) X
# имя X часто вспоминают
```

X дал (ADJ) оценку У

исследование X показало, что CONJ

Всего для дальнейшей было отобрано 120 подобных типов контекстов.

2.4 Структура парсера

Решение задачи извлечения именованных сущностей в нашем проекте осуществлялось без использования нейронных сетей, поскольку контексты, в которых употребляются имена ученых, ограничены и специфичны, размеченных данных было слишком мало для обучения сети. Мы воспользовались Томита-парсером, разработанным компанией Яндекс[4]. Он предполагает создание пользовательских контекстно-свободных грамматик, позволяющих искать в текстах определенные шаблоны, извлекать факты (в нашем случае это имена ученых). При выделений контекстов для дальнейшего написания правил в общем виде нам было важно определять не только то, в окружении каких слов находится имя, но и какими грамматическими характеристиками они обладают.

Грамматика состоит из правил вида:

```
Scholar -> (Adj<gnc-agr[1]>) Status<gnc-agr[1], gram='sg'>
Person<gnc-agr[1], rt>;
```

где может быть описан грамматический контекст сущности, которую мы хотим выделить: граммемы, которые ее окружают, главное слово, с которым будут согласованы другие слова выделяемой цепочки. При необходимости в состав правила могут быть включены регулярные выражения:

```
Name -> Word<wfm=/[А-Я][а-я-]+/>;
```

Далее с помощью интерпретатора выделенные цепочки преобразуются в факты, которые могут быть представлены в отладочном выводе html-таблицей, в xml и других форматах для дальнейшей работы с ними.

Фрэнк Дайсон	астроном
Эддингтон	профессор
Эндрю Кроммелина	астроном

Оказалось, что грамматика выделяет слишком много слов, не относящихся к именам ученых. Чаще всего это были названия географических объектов или просто имена нарицательные, которые случайно захватывались правилами.

[4] Ссылка на документацию Томита-парсера:
https://tech.yandex.ru/tomita/doc/tutorial/concept/about-docpage/

Однако нам не удалось справиться с некоторыми проблемами в работе парсера на уровне построения грамматики. Мы получили ложные срабатывания правил на словах, стоящих в начале предложения, начинающихся с заглавной буквы. Например:

Западе
Александр Смирнов
Исследования
Рапп
Рестораны
Владимира Клименко
Палеоклиматология

Поскольку само имя выделялось при помощи регулярного выражения (см. выше), под шаблон попадали первые слова в предложении. Чаще всего это были топонимы либо просто словарные слова. Одним из решений Томита-парсера был запрет выделения первых слов в предложении, захватываемых регулярным выражением с помощью пометы ~<fw> (not first word). В таком случае точность выделения имен парсером бы повысилась, но мы бы потеряли в полноте. Еще одно решение, возможное с применением средств Томита-парсера -- выделение имен с помощью специальных помет gram='имя', gram='отч', gram='фам' внутри регулярного выражения, но в таком случае мы могли бы выделить только русские имена и фамилии, что связано с особенностями Томиты. Чтобы избавиться от ненужных слов, мы добавили дополнительные фильтры.

Для удаления словарных слов мы использовали словари А. М. Шведова и Д. Н. Ушакова, а для удаления топонимов -- словарь географических названий, взятый в Wikilivres[5]. Это помогло избавиться от лишних слов и улучшить качество разметки. Некоторых слов в используемых нами словарях не оказалось, поэтому из финального списка их пришлось удалять вручную. Далее следует фрагмент общего списка, где можно увидеть слова, оставшиеся после фильтрации через словари.

Многие
Кто-то
Часто
Нобелевской
Сталин
Эйнштейн
Аристотель
Платон
Декарт
Дарвин

2.5 Результаты

[5] Словарь географических названий Wikilivres.
http://wikilivres.ru/Словарь_географических_названий

В результате работы парсера на 167 размеченных вручную текстах мы получили точность разметки (accuracy) 13%. Значение получилось таким низким, поскольку контекстов, в которых используются имена ученых, слишком много, и мы смогли формализовать и написать правила только для небольшой их части. Чтобы написать правила, которые бы охватывали все контексты, необходимо гораздо больше времени. Поэтому работа над правилами Томиты будет продолжена.

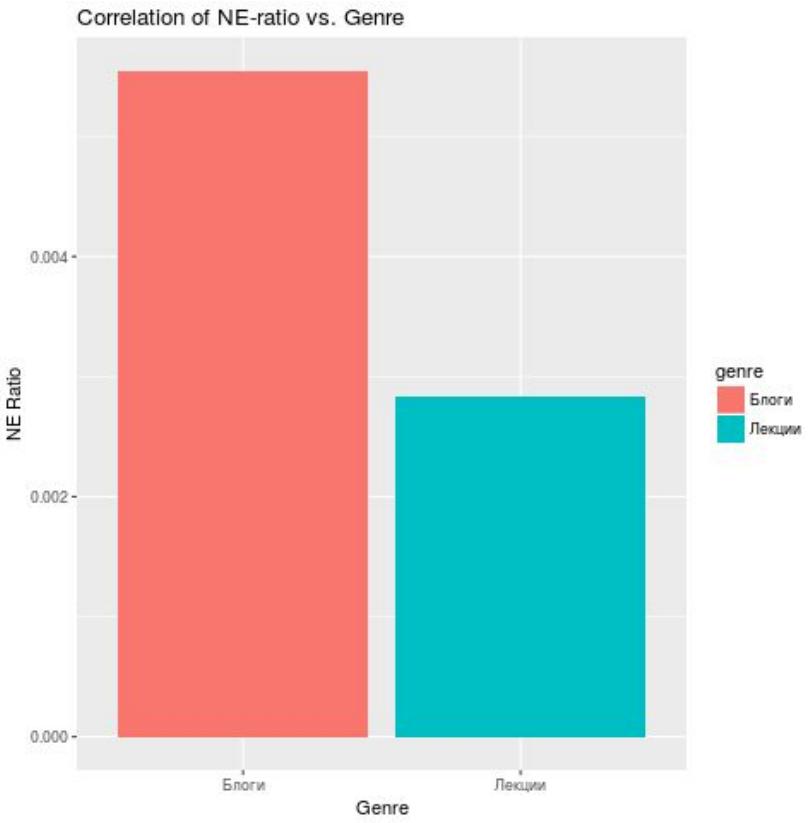
Для составления рейтинга ученых мы решили взять сегмент корпуса (6068 текстов), в который вошли блоги и лекции, в 2665 из них наш модуль выделил имена ученых. Такой выбор жанра не случаен, так как контексты упоминания имен ученых в блогах и лекциях отличаются от упоминаний, например, в новостях. Упоминания ученых в новостном жанре имеет вид: “астроном Х открыл новую звезду”. Такой контекст упоминания не указывает на общественную значимость этого ученого, поскольку имя используется в статье только один раз, и не является цитированием или отсылкой на идеи ученого. Многократные упоминания Аристотеля или Эйнштейна в лекциях, на наш взгляд, говорят о значимости этих людей в истории и общественном сознании. Также язык лекций и блогов представляется наиболее простым и разговорным, что является характерной чертой научно-популярных текстов. Далее представлена сводная таблица упоминания ученых по всем жанрам и рубрикам текстов из взятого сегмента. Стоит принять во внимание, что это не вполне точный список с учетом качества работы парсера и отсутствия инструмента для лемматизации[6] имен.

Имена ученых	**Количество упоминаний**
Аристотель	24
Эйнштейн	23
Платон	22
Декарт	26
Дарвин	19
Ницше	18
Карл Шмитт	17
Галилей	16
Ирвинг Гоффман	15
Зигмунд Фрейд	15

[6] Еще одна непростая задача, которую нам предстоит решить в дальнейшем, -- лемматизация имен собственных. В рамках этого проекта пока не удалось добавить решение для лемматизации.

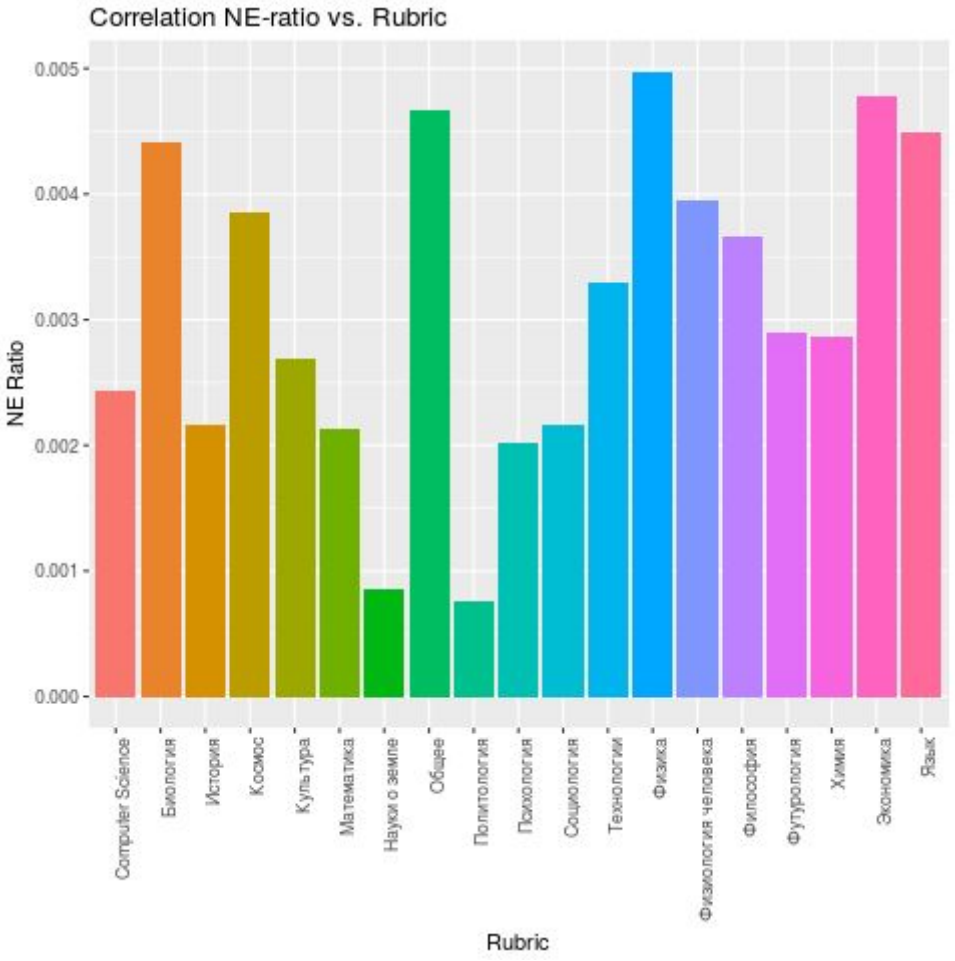
Томас Гоббс	13

Всего в подкорпусе мы выделили 5523 имени, куда попали в том числе выделенные парсером имена людей, не принадлежащих к числу ученых (*Путин, Достоевский, Булгаков* и др.).



В ходе исследования мы получили следующее распределение имен ученых по жанрам и рубрикам: оказалось, что в блогах имена ученых употребляются больше, чем в лекциях. Больше всего имен упоминается в текстах, принадлежащих рубрикам “Физика”, “Экономика”, “Язык”, “Общее”. К тематике “Общее” относятся тексты, которые рассказывают о премиях, присваиваемых выдающимся ученым (Нобелевская

премия), поэтому часто в таких текстах упоминается много имен лауреатов.



Далее в таблицах представлены наиболее часто употребляемые имена по рубрикам.

Computer Science

Шеннон	5
Чистяков	5
Ник Бостром	4
Курцвейл	4
Тьюнинг	3
Цермело	3
Хэмминг	2
Сергей Марков	2
Миронов	2

Пайк	2

Биология

Дарвин	16
Холдейн	4
Тимонов	4
Синъя Яманака	4
Патрик Хаус	4
Теркер	4
Вальтер Флемминг	3
Ратгерский	3
Кингдон	3
Брайант	3

История

Гумбольдт	12
Фукидид	11
Артем Ефимов	10
Галилей	9
Аристотель	6
Данилевский	6
Геродот	5
Гоббс	5

Космос

Эйнштейн	11
Пифагор	6

Артур Эддингтон	5
Стивен Хокинг	3
Артем Елмуратов	3
Вальтер Бааде	3

Математика

Колмогоров	17
Перельман	6
Апу	4
Мариам Мирзахани	3
Семереди	3
Бурнаев	3
Риман	3

Психология

Дональд Винникотт	8
Зигмунд Фрейд	6
Маслоу	3
Татьяна Карягина	3
Фредерик Бартлетт	3

Социология

Ирвинг Гоффман	15
Бруно Латур	10
Стивен Вулгар	8
Альфред Шюц	7
Рем Колхас	7

Физика

Декарт	26
Эйнштейн	16
Дж. Максвелл	10
Пол Дирак	6
Паули	5
Питер Хиггс	4
Эрнест Резерфорд	4
Ричард Фейнман	4
Лев Ландау	3

Философия

Фридрих Ницше	10
Аристотель	9
Платон	7
Джеймс Мэдисон	5
А. Ф. Лосев	5
Сократ	4
Никколо Макиавелли	4

Экономика

Джон М. Кейнс	9
Адам Смит	5
Сергей Гуриев	4
Рональд Коуз	3
Бен Бернанке	3

Язык

Людвиг Витгенштейн	8
А. М. Пешковский	4
В.А. Плунгян	4
Стивен Пинкер	3
Сыма Цянь	3
Ноам Хомски	3

Наиболее цитируемыми учеными[7] с индексом Хирша более 100, которые попали и в наши списки стали Зигмунд Фрейд, Ноам Хомски, Чарльз Дарвин, Альберт Эйнштейн, Стивен Хокинг, Лев Ландау.

Выводы

Результаты сравнения нашего рейтинга и рейтинга цитирования Google Scholar оказались вполне ожидаемыми. В наши списки наиболее цитируемых ученых попали самые цитируемые и авторитетные ученые в своих областях. Практически во всех рубриках можно наблюдать, что в топ попали ученые, принадлежащие к конкретной области науки. Исключением стал, например, А. Эйнштейн, результаты трудов которого используются не только физиками но и астрономами.

Удивительно, что наибольший индекс Хирша -- у З. Фрейда, однако, он не попал в сводную таблицу по всем рубрикам. Нужно отметить, что многие античные ученые и ученые XVII - VIII веков не значатся в списке цитирования Google Scholar, поэтому их значимость мы можем оценить только на наших рейтингах. Античные историки, философы, математики (Пифагор) имеют большой вес в современных научно-популярных текстах.

[7] Согласно Google Scholar Citation Profiles. http://www.webometrics.info/en/node/58.

3.1 Извлечение терминов

Введение

Одна из задач нашего комплексного исследования заключалась в анализе наполнения терминами русскоязычного сегмента научно-популярных текстов.

При создании алгоритма извлечения терминов из текста учитываются лингвистические и статистические характеристики термина (Ahrenberg 2013). К лингвистическим характеристикам относятся:

1) Синтаксическая структура. На основе анализа этой структуры пишутся специальные правила, формализующие характерные для употребления терминов синтаксические конструкции, например:
((Adj|Noun)+|((Adj|Noun)*(NounPrep)?)(Adj|Noun)*)Noun для английского языка. Такие правила подаются на вход программам-парсерам, дающим на выходе слова, удовлетворяющие полученным правилам.
Для использования подобных правил необходим корпус с частеречной разметкой или POS-тэггер. Как правило, подобные правила помимо термина захватывают много нерелевантных кандидатов, для их отсеивания используются дополнительные фильтры.

2) Морфологическая структура (специфические аффиксы в составе термина); реализуется в алгоритме на основе списка специфических символьных n-грамм.

3) Типичная структура контекстного окружения; также как и синтаксическая структура реализуется в виде специальных формальных правил, учитывающих характерный для терминов контекст (например, с какими словами термины часто употребляются в словосочетаниях). Такие правила подаются на вход программе-парсеру (также как и при извлечении именованных сущностей, описанном выше). Недостатком их работы, как правило, также является большое количество нерелевантных кандидатов на выходе.

Статистические характеристики включают в себя частотность и совместную встречаемость отдельных слов внутри терминологического словосочетания (co-occurence) (для мультиграмм).

Оценка работы алгоритма по извлечению терминов сопряжена с рядом трудностей, так как отсутствует золотой стандарт, и в лучшем случае приходится использовать готовые глоссарии или тезаурусы. Для нашей задачи этот метод применить довольно сложно, так как списки терминов составляются, как правило, для конкретной области знания, а не для разных тематик.

Еще одним способом валидации отобранных кандидатов является проверка экспертами вручную.

Для англоязычных текстов традиционно имеется большое множество готовых инструментов, решающих эту прикладную задачу (например, SynchroTerm, SDL MultiTerm Extract, Taas, Lexterm и т.д.). Эти инструменты основываются на статистических и лингвистических принципах. Основной минус работы большинства подобных инструментов - большое количество нерелевантных кандидатов.

Для русского языка примером можно привести использованную нами библиотеку Rutermextract, речь о которой пойдет ниже.

Исследовательская гипотеза

Наша исследовательская гипотеза заключалась в предположении, что для того, чтобы быть понятными широкому кругу читателей, научно-популярные тексты не должны быть перегружены терминами. При этом при выборе терминов для научно-популярного текста автор должен ограничиваться только ключевыми и самыми необходимыми для раскрытия конкретной темы.

Наше исследование включало в себя:

- сравнение терминов из текстов различных тематических рубрик, а также жанров (например, существуют ли термины универсальные для всех рубрик), построение графа совместной встречаемости для наиболее частоупотребимых терминов;
- анализ наполненности научно-популярных текстов терминами (по сравнению с узкоспециальными научными текстами), сравнение плотности терминов в текстах разных тематик;
- изучение контекстного окружения терминов;
- изучение семантических особенностей терминов, построение word2vec модели для частотных терминов-униграмм.

Разметка

Для ручной разметки тексты отбирались таким образом, чтобы в размеченном корпусе были представлены все рубрики и жанры. В итоге одним аннотатором было размечено 72 текста. Для маркирования границ терминов в тексте, также как и для именованных сущностей, использовались символы “&” и ”&!”; так как термины и имена ученых размечались на разных текстах, при дальнейшей обработке путаницы не происходило.

Основной сложностью при разметке текстов являлась субъективность аннотатора, особенно когда речь шла о терминах, которые начинают проникать в повседневный язык и подвергаются процессу так называемой детерминологизации (Кругосвет). Так как при ручной разметке решение о том, является ли слово термином или нет, принималось аннотатором на основании собственного мнения из-за отсутствия четких однозначных критериев определения “термин - не термин”, при проверке размеченных текстов другими исследователями возникали споры на тему того, является ли то или иное слово собственно термином или уже не является. Полемику вызывали такие слова, как “криминология”, “девайс”, “атмосфера” - то есть слова, которые в силу определенных экстралингвистических причин (частое употребление в сми, распространение предметов и явлений, обозначаемых данными терминами, в

повседневной жизни и т.д.) утратили свою строго ограниченную принадлежность к узкоспециальной области знания и стали широкоупотребимыми (причем зачастую такие слова все еще могут маркироваться в словарях специальными пометами, однако при этом уже активно использоваться в повседневном языке).

В большинстве источников и толковых словарей понятие “термин” имеет четкое и однотипное определение. Возьмем, к примеру, энциклопедию “Кругосвет”: *“Термин (лат. terminus 'граница, предел, конец') – это специальное слово или словосочетание, принятое в определенной профессиональной сфере и употребляемое в особых условиях”*. По иронии в контексте нашего исследования понятие “термин” не имеет четко обозначенной границы с общеупотребительной лексикой. Это объясняется, с одной стороны, спецификой исследовательского поля, - как следует из самого названия, научно-популярные тексты объединяют в себе особенности и научного, и газетно-публицистического стиля. С другой стороны, подобная размытость границ понятия “термин” обусловлена тематическим и жанровым разнообразием исследуемого корпуса текстов, в то время как большинство исследований по извлечению терминов на русскоязычном материале проводилось на корпусах текстов, принадлежащих к конкретной отрасли знания, что вполне логично (например, тексты по корпусной и компьютерной лингвистике (Митрофанова, Захаров 2009, Соколова, Семенова 2011) или тексты технической направленности (Клышинский, Кочеткова 2014).

Анализ размеченных текстов

Всего из размеченного вручную корпуса было извлечено 1135 терминов; из них 558 униграмм, 464 биграммы, 75 триграмм и 38 мультиграмм (4 и более слов).

Частеречный анализ словных униграмм выявил значительный перевес в сторону существительных (532 слова из 558) с очень редким включением глаголов (16 слов из 558) и прилагательных (10 слов из 558) (см. приложения 1 и 2). Интересно заметить, что некоторые лингвисты рассматривают термины исключительно как номинативные единицы (существительные и именные группы) и в принципе исключают глаголы из терминологического поля (Гринев-Гриневич 2008).

Таким образом, результаты обработки небольшого корпуса текстов показывают преобладание среди терминов словных униграмм и биграмм, а также исключительное преобладание существительных над другими частями речи.

Для анализа контекстов употребления после компьютерной обработки список терминов выводился в формате KWIC (key word in context) с контекстным окном в 5 слов справа и слева от термина. При этом отдельно выделялось последнее слово левого контекста для облегчения сортировки и анализа.

Left context	Left word	Term	Right context
в тот день когда я	я	*синтезировал ДНК*	и присоединил ее к сгустку
причиной расхождения в датировках является	является	*пресноводный резервуарный эффект*	ПРЭ Он проявляется в мнимом
силой противящейся движению поезда является	является	*аэродинамическое сопротивление*	Предыдущий рекорд скорости был установлен
огромной опорой именно на язык	язык	*эмоциональной рефлексии*	об этом пишет польско австралийский

В итоге из чуть более 1000 употреблений мы смогли выделить 34 устойчивых контекста (см. приложение 3). Их можно разделить на следующие группы:

- контекстное существительное + термин в родительном падеже;
- контекстный глагол + термин в косвенном падеже;
- контекстный глагол “являться” + термин в именительном падеже (и иногда наоборот);
- вводное слово “например” + термин в именительном падеже;
- “под названием” + термин в именительном падеже.

Однако хотя вероятность встретить термин в выявленных выше паттернах высока, они не являются специфическими только для терминов из-за высокой частоты употребления данных конструкций в научно-популярных текстах.

Основываясь на этих выводах, мы решили, что использование Томита-парсера будет малоэффективным из-за отсутствия специфических для терминов контекстов, а также из-за невозможности писать большое количество правил. Методы машинного обучения также пришлось исключить из-за недостаточного для обучения размера размеченного корпуса.

Таким образом, было решено строить алгоритм извлечения терминов, основываясь на особенностях их морфологии, а также на частоте употребления.

Алгоритм извлечения терминов

Первичный отсев кандидатов

Для первичного отбора кандидатов из текста мы использовали библиотеку Rutermextract[8]. Эта библиотека позволяет выделять ключевые слова и словосочетания для конкретного текста; они выводятся списком от более значимых к менее значимым в лемматизированном виде. В основе работы Rutermextract лежат правила,

[8] https://github.com/igor-shevchenko/rutermextract

морфологический анализ производится при помощи модуля Pymorphy2. Как указывает сам автор, основные недостатки этой библиотеки - неполные правила (так, не выделяются словосочетания с предлогами, например, “сопротивление в цепи”) и неоднозначность при морфологическом разборе.
В результате теста на размеченных текстах мы получили полноту равную 86,5% (153 несовпадения из 1135). При этом Rutermextract в качестве кандидатов выводит все термины-униграммы, которые были выделены при ручной разметке. Также он хорошо выделяет би- и триграммы, и даже распознает такой сложный термин как 'рандомизированные двойные слепые плацебо-контролируемые клинические исследования'. Некоторые термины из ручной разметки не совпали с аналогичными кандидатами из извлеченных Rutermextract из-за того, что были по-разному лемматизированы. Некоторые словные тетраграммы, выделенные при ручной разметке, Rutermextract выделяет как отдельные биграммы.

Таким образом, библиотека Rutermextract обеспечивает хорошую полноту при извлечении необходимых нам терминов, но при этом также выделяет много слов и словосочетаний, которые являются ключевыми или значимыми для конкретного текста, но при этом не являются терминами. В связи с этим, следующий шаг - это отсев неподходящих кандидатов. Как мы писали выше, выбранные критерии для отсева - частота употребления, а также морфологическая структура терминов.

Проверка частотности
Отсеивание кандидатов по критерию частотности производилось при помощи “Частотного словаря современного русского языка” (Ляшевская, Шаров, 2009).
Стратегия отсева заключается в отборе слов, не попавших в состав частотного словаря, а также слов с низким показателем F(imp). Для определения порогового значения F(imp) мы проанализировали эти значения для терминов, выделенных нами при ручной разметке и попавших в частотный словарь. Полученные значения F(imp) варьировались в пределах от 119.59 для слова 'функция' до 0.4 для слова 'синекура'. В итоге после анализа кандидатов, попавших в этот интервал, пороговым значением было выбрано $F(imp) < 20$, при котором хорошо отсеивались нерелевантные слова и сохранялась полнота.

Анализ морфологической структуры
Следующий шаг в нашем алгоритме для выделения терминов основан на анализе морфологической структуры слова. Этот анализ заключается в проверке наличия символьных n-грамм, специфичных для терминов, в составе кандидата. Для получения списков символьных n-грамм из множества выделенных нами вручную терминов, дополненного терминами, взятыми из Википедии, после процедуры стемминга (отсекания флексий у словоформ), извлекались символьные би-, три- и тетраграммы из начальной и конечной частей слова. Далее из полученных списков n-грамм удалялись

соответствующие n-граммы, являющиеся высокочастотными для русского языка (список частотных n-грамм был сформирован на текстах из Википедии). Важно отметить, что при получении специфических для терминов n-грамм использовался именно стемминг, а не лемматизация, чтобы можно было выделить суффиксы и не захватывать при этом флексии.

Результаты работы алгоритма

После первой итерации алгоритма было выделено 901090 униграмм. Предварительная оценка полученных результатов показала, что среди кандидатов содержится много имен собственных и географических названий. В качестве фильтра для этих нерелевантных кандидатов были использованы список имен ученых, полученный при извлечении именованных сущностей, а также список географических названий, выкачанный из Википедии. Помимо этого часто попадаются слова на английском языке, либо слова с латинскими символами или цифрами, такие слова были удалены.
После этой обработки полученные термины были ранжированы по частоте употребления.

Анализ наиболее часто употребляемых слов показал, что среди нерелевантных кандидатов очень часто встречаются:

- географические названия (не попавшие в начальный стоп-лист);
- имена собственные, отчества и фамилии (также не попавшие в начальный стоп-лист);
- названия национальностей и рас;
- названия животных, насекомых, реже растений;
- названия религий и принадлежности к ним;
- заимствования с нехарактерной для русского языка морфологической структурой (“сюрприз”, “резюме”, “экскурсия”, “жюри” и т.д.);
- названия планет;
- слова, типичные для научно-популярного и публицистического жанра, а также слова, обозначающие абстрактные понятия (“распространенность”, “современность”, “многообразие”, “основоположник”, “осмысление” и т.д)
- аббревиатуры;
- неправильно лемматизированные слова (“даба” - “дабы”, “чаща” - “чаще”, “кривой” - “кривая”, “светова” - “световой”, “церер” - “церера” и т.д.)

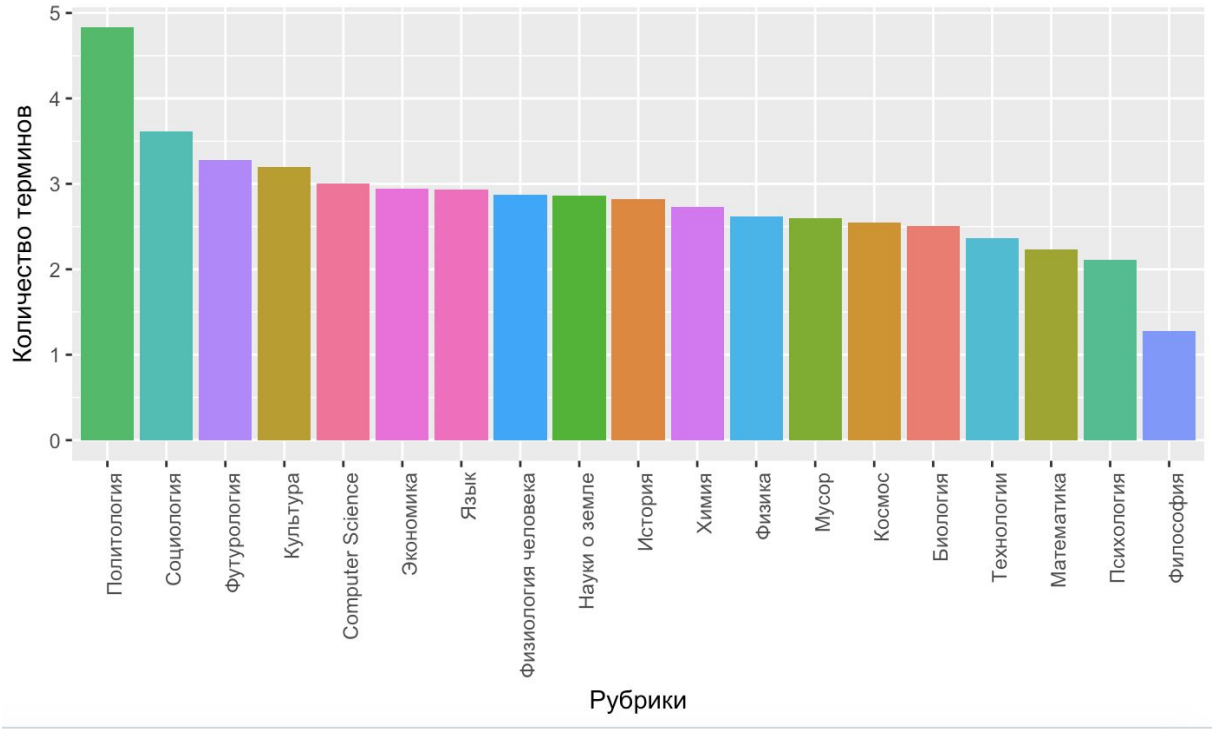
Среди редко употребляемых кандидатов часто встречаются:

- слова с орфографическими ошибками;
- неологизмы;
- неправильно лемматизированные слова (особенно для существительных, определенных как глаголы, и неправильно лемматизированных субстантивов).

Список стоп-слов был дополнен наиболее часто встречающимися нерелевантными кандидатами, выявленными на основе приведенного выше анализа, и его размер сейчас составляет 19149 слов. После применения расширенного фильтра количество извлеченных униграмм сократилось более чем в два раза, а именно до 432 514 слов.

Предварительные результаты

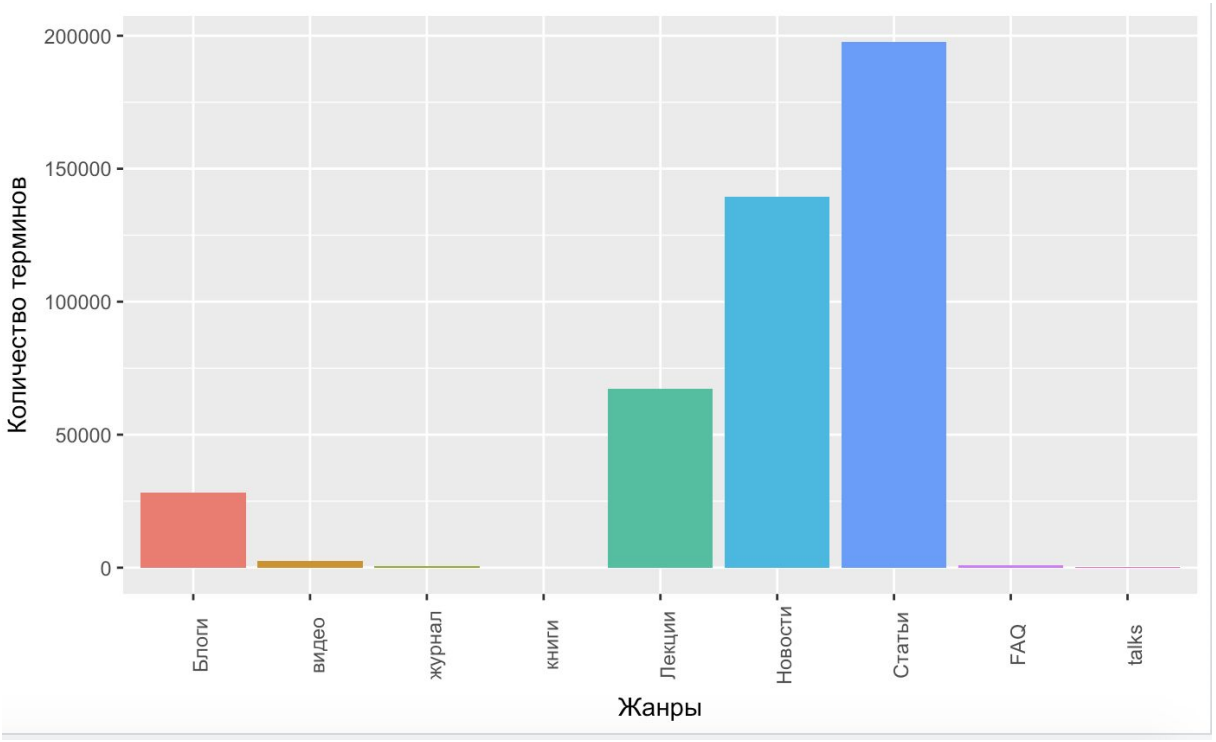
Наполненность текстов терминами
Наполняемость текстов терминами по рубрикам и жанрам для униграмм выглядит следующим образом (с учетом все еще значительного количества нерелевантных кандидатов):



(наполняемость = $\log\frac{N}{K}$, где N - количество терминов, употребленных в конкретной рубрике, К - количество текстов в рубрике)

В целом, термины распределены по рубрикам достаточно равномерно, с заметным перевесом использования терминов в текстах по политологии; меньше всего терминов употребляется в текстах по философии.

Анализ распределения терминов по жанрам показал самую высокую наполняемость для статей и самую низкую - для блогов. В лекциях ораторы используют меньше терминов, чем в новостях или статьях (то есть в письменных текстах). В научно-популярных новостях термины встречаются регулярно, но в целом реже, чем в статьях.



Частотный словарь

Всего из извлечнных нами 432 514 слов было выделено 70 693 уникальных термина, из которых был сформирован словарь, ранжированный по частотности и разбитый по рубрикам.[9]

Анализ терминов для каждой рубрики показал, что, как правило, среди самых частоупотребимых встречаются (в порядке убывания):

- названия профессий или исследователей (*разработчик, программист, биолог, палеонтолог, археолог, генетик, геофизик, метеоролог*);
- объекты исследования и их составляющие (*контент, бактерия, днк, нейрон, рецептор, митохондрия, артефакт, фотон, гравитация, социум, константа, фермент, тренд, морфема, импорт, синтаксис*);
- названия исследовательской деятельности (*распознавание, визуализация, активизация, выработка, интерпретация, полемика, скрининг*);
- действия, производимые объектами исследования (*адаптация, деление, распад, выброс, отклонение, коммуникация, дискурс, коммуникация*);
- инструменты и устройства (*микроскоп, телескоп, планшет, ускоритель, радар, генератор, аккумулятор*);
- результаты исследовательской деятельности (*вакцина, инновация, антибиотик, прототип, плацебо*).

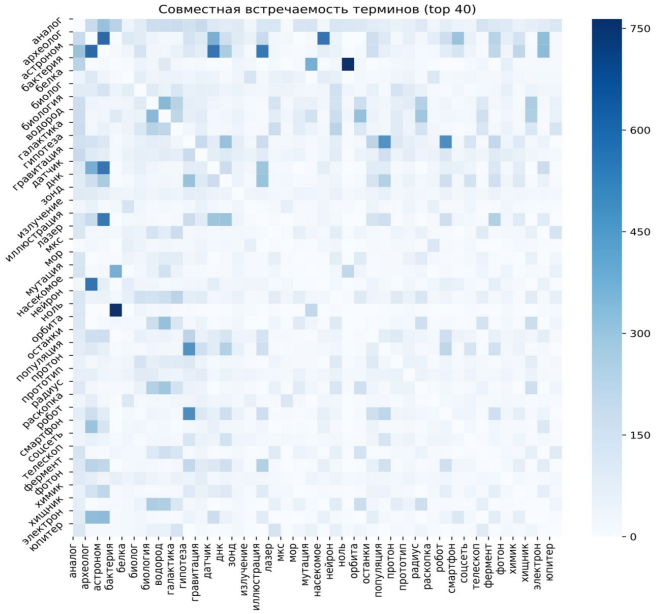
9 https://github.com/ana-kuznetsova/Popular-Science-Texts-Compling-research/tree/master/term_extraxtion/terms_sorted

Заметно выделяются рубрики “Физика” и “Химия”, в которых наполненность узкоспецифическими терминами значительно выше, чем в других рубриках (*электрон, фотон, протон, нейтрон, вакуум, нейтрино, кварк, антиматерия, фермент, катализатор, катализ, аминокислота*). Это объясняется спецификой объектов исследования этих научных отраслей, без упоминания которых не обойтись даже в доступном научно-популярном тексте.
Самую низкую наполненность терминами показала рубрика “Философия” (всего 23 слова, каждое из которых встречается в рубрике 1-2 раза). При этом в рубрике отсутствуют термины, специфичные для этой науки, и также много шума. Возможно, это объясняется, тем, что в корпусе очень мало текстов этой тематики и он для нее нерепрезентативен (всего 143 текста из более чем 30 000).

Универсальными терминами для всех рубрик оказались слова “гипотеза” и “аналог”, что вполне объяснимо.

Граф совместной встречаемости

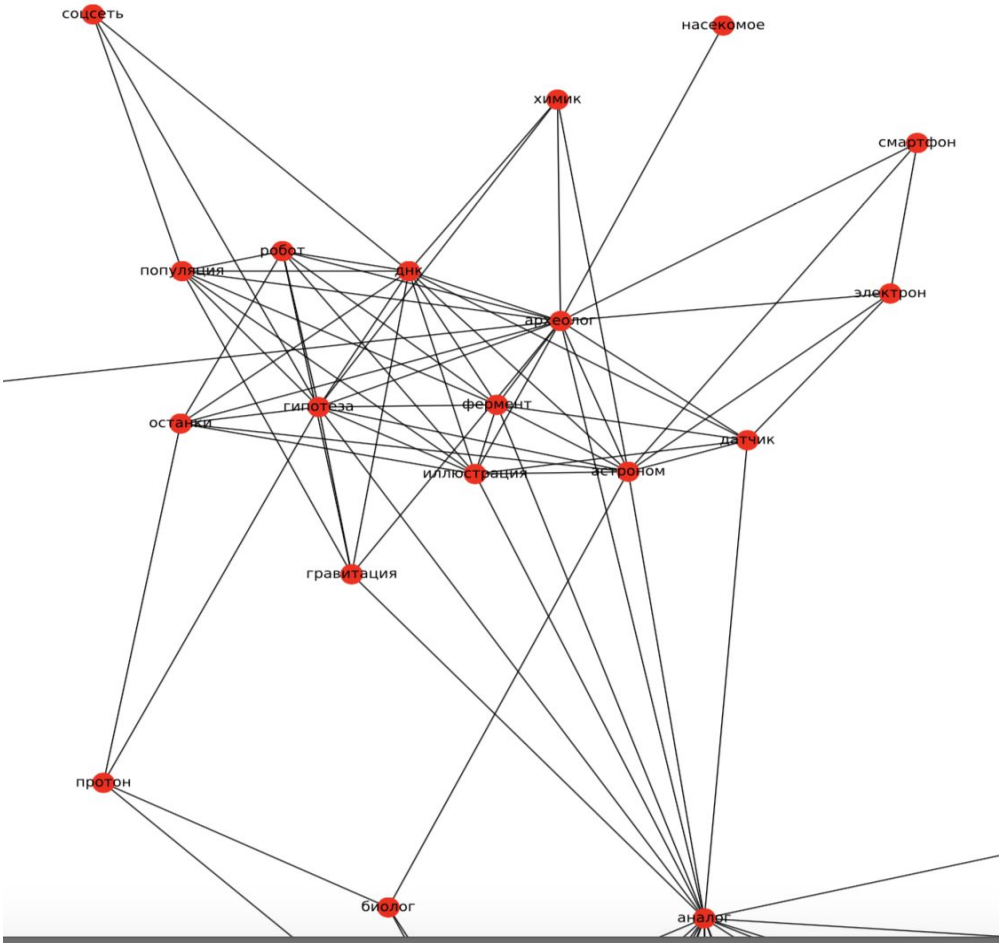
Анализ совместной встречаемости первых 40 самых частоупотребимых терминов на целом корпусе текстов показал довольно равномерное распределение этого показателя:



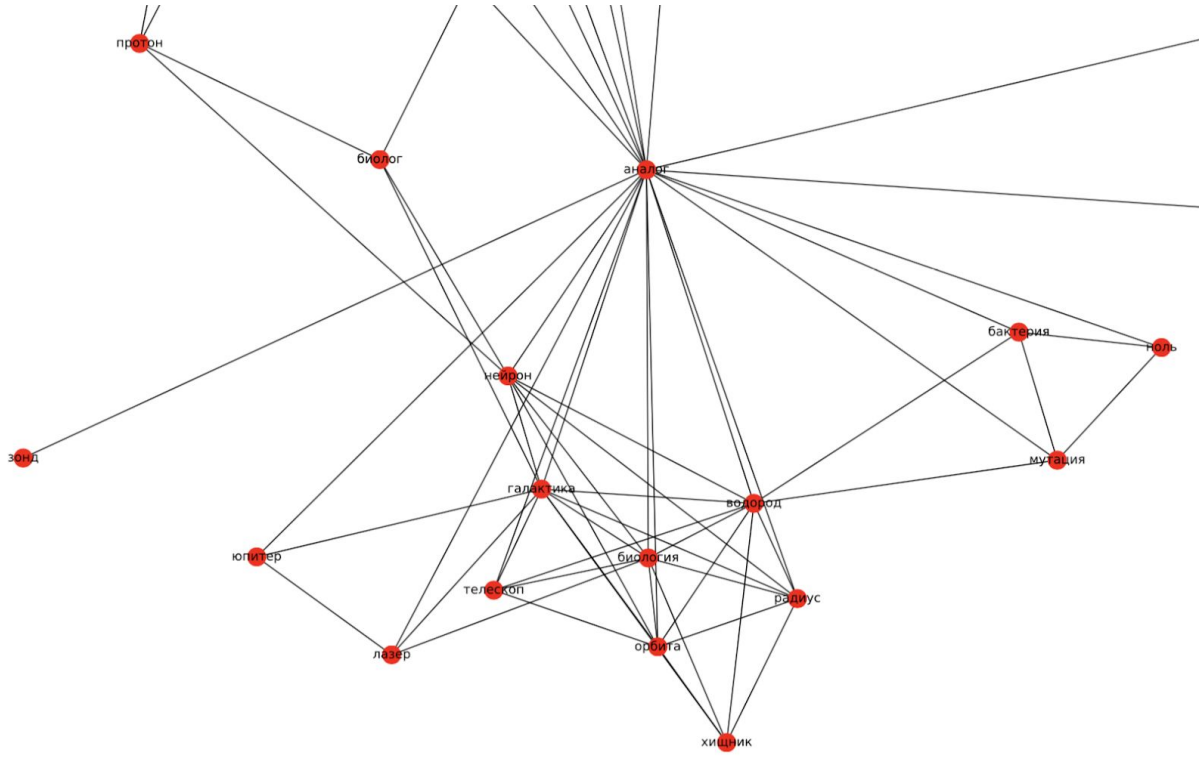
На основе приведенной выше матрицы мы построили граф совместной встречаемости, отфильтровав пары терминов, встречающиеся между собой менее 100 раз[10]. На графе хорошо прослеживаются два кластера:

10

https://github.com/ana-kuznetsova/Popular-Science-Texts-Compling-research/blob/master/term_extraxtion/cooccurence_graph/top40_graph_filter100.pdf

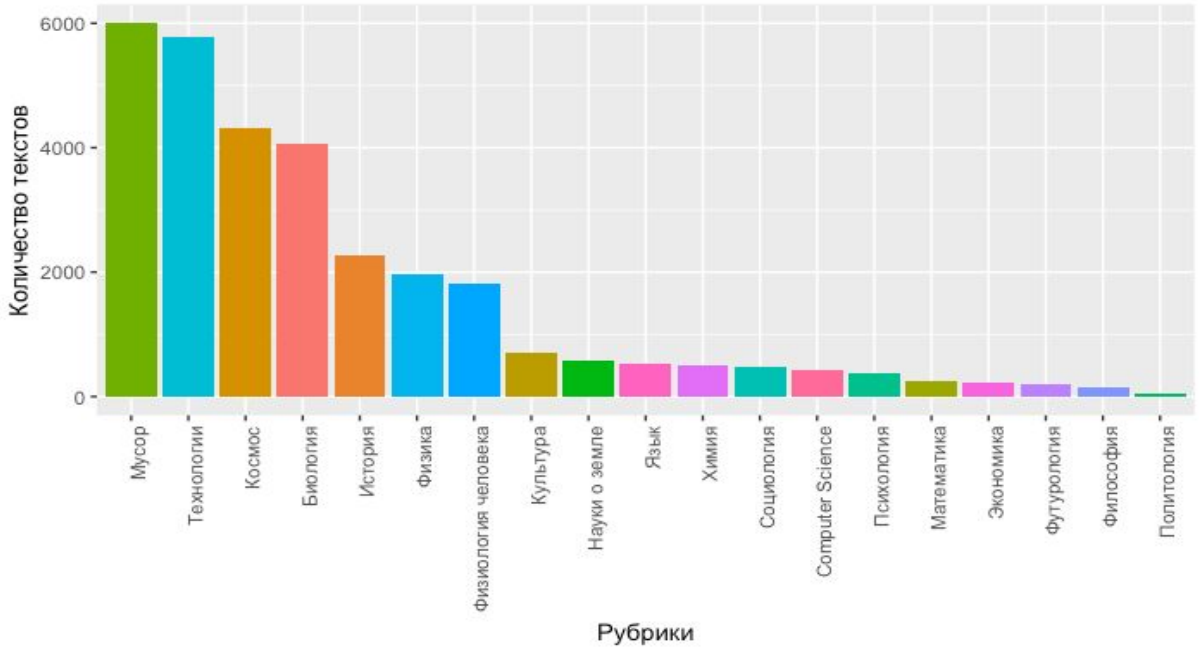


Первый кластер объединяет в себе термины, характерные для текстов про историю, космос, физику и биологию. Это объясняется значительным количеством междисциплинарных текстов в корпусе, например, об исследовании физических явлений или поведения животных, бактерий и растений в космосе, или изучения останков человека, животных и даже насекомых инновационными методами. Слово “археолог” соседствует со словом “астроном” также не случайно: это результат присутствия в корпусе текстов про археоастрономию - науку, изучающую представления древних людей об астрономии.



Второй кластер показывает связь терминов, характерных для рубрик “Космос”, “Химия” и “Биология”, что также говорит нам о значительном количестве текстов, отражающих устойчивую взаимосвязь этих областей знания друг с другом.

Эти кластеры отражают сильную взаимосвязь и междисциплинарность перечисленных выше наук, но также такая картина может быть следствием значительного перевеса в корпусе текстов о технологиях, космосе, биологии и истории:



Word2vec

Для изучения семантических особенностей терминов на основе нашего корпуса была натренирована модель word2vec[11]. Анализ ближайших соседей для первых 20 самых частотных терминов показал, что для названий профессий ближайшими к ним будут узкоспециализированные профессии и объекты исследования данных областей: *астроном - астрофизик, планетолог; археолог - египтолог; биолог - зоолог, этномолог, генетик* и тд. Для названий частиц - другие частицы: *электрон - фотон, протон, позитрон, ион; фотон - квант, электрон, позитрон; протон - кварк, мюон, позитрон* и тд. Для объектов исследования - их составные части или наоборот, другие объекты, частью которых они являются: *нейрон - синапс, дендрит, аксон, мозг; белка (=белок) - фермент, пептид, аминокислота; днк - геном, нуклеотид, рнк*. Для явлений - объекты, с которыми они взаимодействуют: *излучение - луч, радиоволна, спектр, гамма; мутация - аллель, фенотип, ген, хромосома*.

Вывод

В заключении можно сказать, что наша изначальная гипотеза о том, что научно-популярные тексты не должны быть перегружены терминами, подтвердилась. Анализ частотности показал, что на вершине списка находятся слова, легко понимаемые почти любым носителем языка, а узкоспецифичные термины встречаются достаточно редко (часто с показателем частотности ниже 10) и только тогда, когда без них действительно не обойтись, за исключением текстов по физике и химии. Помимо этого можно сказать об очень высокой степени взаимопроникновения различных областей знания и, как следствие, большого количества междисциплинарных текстов и общих терминах в них.

11 https://github.com/ana-kuznetsova/Popular-Science-Texts-Compling-research/blob/master/term_extraxtion/word2vec/w2v_results.txt

4. Текстовая близость

4.1. Исследовательский вопрос

Научно-популярные тексты представляют собой интересный объект с точки зрения анализа близости текстов разных тематик и внутри одной тематики, т.к. в этом жанре часто освещаются междисциплинарные вопросы и встречается большое количество текстов на стыке разных наук.

В своей работе мы попробовали провести тематический анализ научно-популярных статей с помощью различных методов компьютерной лингвистики. Цель исследования состояла в том, чтобы понять, какие тексты оказываются близкими с точки зрения автоматической обработки и как эта близость соотносится с тематикой текстов: для каких научных областей статьи сильно различаются, или наоборот считаются похожими. Кроме того, интересно выяснить, будут ли различаться оценки близости, полученные с использованием различных методов.

4.2. Методы

Для анализа текстовой близости в соотношении с тематикой текстов мы рассмотрели следующие группы методов:

1. Методы оценки схожести текстов (text similarity).
2. Кластеризация текстов.
3. Методы снижения размерности.
4. Тематическое моделирование.

Text similarity

Текстовая близость в корпусе статей изучалась на 24611 текстах. Для анализа схожести текстов внутри рубрик и между различными рубриками была построена матрица попарной близости текстов корпуса.

Существуют различные методы для измерения схожести текстов:

1. Подход на основе лексической близости -- включает методы, измеряющие близость текстов по составу термов (*term-based*).
2. Корпусные подход -- методы, оценивающие семантическую близость текстов с использованием корпуса (*corpus-based*).
3. Подход на основе семантических сетей -- методы, определяющие степень сходства текстов через семантические отношения между словами (*knowledge-based*) (Gomaa et al. 2013).

В своем исследовании мы использовали *term-based* и *corpus-based* методы.

Из лексических методов мы рассмотрели *меру сходства Жаккара* (Gomaa et al. 2013). Мера Жаккара для двух текстов определяется как отношение числа термов, входящих в оба текста, к общему числу уникальных термов двух текстов (отношение объема пересечения словарей двух текстов к объему их объединения).

На рассматриваемом научно-популярном корпусе значения меры Жаккара получались очень низкими, т.к. в корпусе представлены тексты большого объема, соответственно пересечение двух текстов по словам оказывается значительно меньше, чем общий объем слов для двух текстов. Таким образом, мера Жаккара не давала показательной информации для оценки схожести текстов.

Для того, чтобы компенсировать влияние объема текстов, мы обратились к корпусным методам, которые для оценки близости текстов вместо сравнения полных составов слов для двух текстов используют сравнение текстов по значимым словам. Один из таких методов -- *Explicit Semantic Analysis* использует в качестве меры схожести текстов косинусную меру близости векторов TF-IDF представлений текстов (Gabrilovich, Markovitch 2007). Значение TF-IDF слова в тексте отражает значимость слова для заданного текста, т.к. оно учитывает частоту слова в тексте и частоту слова в корпусе. Если слово имеет высокую частоту в тексте, и при этом редко встречается в корпусе, то слово считается значимым для заданного текста и получает большой вес TF-IDF. Косинусная мера рассчитывает угол между векторами значимых слов для двух текстов, что позволяет оценить близость двух текстов. В свой работе для построения матрицы схожести текстов мы использовали данный метод.

Для каждого текста было построено векторное представление по 70% объема слов корпуса с самым высоким значением TF-IDF. Далее была рассчитана матрица попарной косинусной близости полученных векторов. По данным матрицы были построены графы близости текстов разных рубрик и граф взаимосвязи рубрик.

В графах близости текстов узлы представляют собой тексты, а ребра соединяют пары текстов, для которых значение косинусной меры превышает пороговое значение. В качестве порога эмпирическим путем было выбрано значение косинусной меры 0.5. Вес ребра соответствует косинусной мере близости текстов.

В графе взаимосвязи рубрик ребра соединяют между собой рубрики, для которых доля близких по косинусной мере текстов по отношению к общему числу пар текстов превышает пороговое значение:

$$S(r_1, r_2) = \frac{N(r_1 r_2)}{N(r_1) * N(r_2)} * 10000,$$

где $N(r_1 r_2)$ -- число пар текстов под рубриками r_1 и r_2, для которых значение косинусной меры больше 0.5, $N(r_1)$ -- число текстов под рубрикой r_1, $N(r_2)$ -- число текстов под рубрикой r_2.

Значения предложенного коэффициента близости рубрик (S) зависят от объема рубрик, однако такой коэффициент позволяет отразить взаимосвязь рубрик между собой. В качестве порога для коэффициента близости рубрик было выбрано значение 0.4 -- среднее значение S для всех пар рубрик, которые связывают между собой близкие по косинусной мере тексты.

Кластеризация

Другой подход к анализу близости текстов разных тематик представляют собой методы кластеризации. Цель кластеризации состоит в разбиении множества текстов на группы таким образом, что тексты внутри группы близки между собой, а тексты

разных групп находятся далеко друг от друга в терминах некоторой метрики. Таким образом, по результатам кластеризации корпуса текстов можно проанализировать, из каких областей тексты попадают в один кластер, или оказываются в разных кластерах.

В настоящем исследовании мы рассматривали результаты автоматической кластеризации текстов с использованием двух методов:

1. Метод k-средних (*k-means*).
2. Иерархическая агломеративная кластеризация (Manning at al. 2008).

Кластеризация множества векторов методом *k-средних* основана на итерационном процессе. В алгоритме изначально задается число кластеров k, и случайным образом выбирается k центральных векторов кластеров (центроидов). Все вектора входного множества распределяются по кластерам в зависимости от того, до какого центроида расстояние от входного вектора минимально. После распределения векторов координаты центроида для каждого кластера пересчитываются как среднее векторов, отнесенных к данному кластеру на первой итерации. Далее аналогичным образом распределение векторов по новым кластерам и перерасчет центроидов повторяется до тех пор, пока координаты центроидов и распределение векторов не зафиксируются.

Иерархические методы кластеризации предполагают построение древовидной структуры вложенных кластеров на множестве входных векторов. Агломеративная кластеризация строит иерархическое дерево кластеров “снизу-вверх”: изначально каждый вектор относится к своему кластеру, затем постепенно ближайшие друг к другу отдельные кластеры объединяются в более крупные кластеры следующего уровня иерархии. Процесс останавливается, когда структура кластеров сходится к одному кластеру в основании дерева.

В своей работе мы провели кластеризацию TF-IDF векторов научно-популярных текстов на 18 кластеров (по числу рубрик, представленных в корпусе) с использованием описанных методов. Полученные результаты были проанализированы с точки зрения того, тексты каких рубрик объединялись в один кластер после кластеризации различными методами.

Снижение размерности

Векторное представлении текстов, где в качестве признаков векторов используются слова, дает высокую размерность признакового пространства векторов. Для анализа структуры корпуса текстов с целью выделения групп, соответствующих различным тематикам, могут быть использованы методы снижения размерности векторного пространства.

Метод *t-SNE* (t-distributed stochastic neighbour embedding) преобразует вектора высокой размерности в двумерные или трехмерные вектора, так чтобы в пространстве меньшей размерности сохранялась структура расположения векторов в исходном пространстве. Метод основан на переходе от евклидова расстояния между многомерными векторами в к условным вероятностям, характеризующим близость

этих векторов в исходном пространстве (вероятность того, что одна точка близка к другой точке). Далее по вероятностному распределению для исходного пространства можно найти эквивалентное двумерное или трехмерное распределение, наиболее близкое к исходному. Полученные вектора меньшей размерности могут быть отображены на плоскости или в трехмерном пространстве, что дает возможность для визуального анализа выделения групп среди множества векторов (Maaten, Hinton 2008).

В своей работе мы построили визуализацию двумерных векторов текстов различных рубрик, полученных в результате применения *t-SNE*. В качестве предварительной обработки текстов было построено векторное TF-IDF представление и проведено первоначальное снижение размерности до 100 компонент с помощью сингулярного разложения.

Тематическое моделирование

Задача тематического моделирования состоит в том, чтобы определить тематику каждого документа в коллекции и распределение ключевых слов по тематикам. При этом каждая тема представляется вероятностным распределением слов, а каждый документ определяется вероятностным распределением тем. Слова и документы являются наблюдаемыми переменными, а темы представляют собой скрытые (латентные) переменные.

LDA (Latent Dirichlet Allocation) – байесовский метод тематического моделирования, предназначенный для определения тематической структуры в коллекции документов (Воронцов 2017). Документ представляет собой пересечение нескольких тем, каждая из которых описываются дискретным распределением слов. Решение задачи тематического моделирования основано на итерационном процессе нахождения параметров распределения слов в теме и распределения тем в документе. Семантическая близость тематических слов определяет интерпретируемость темы.

На корпусе научно-популярных текстов была построена тематическая модель LDA для извлечения 18 тем.

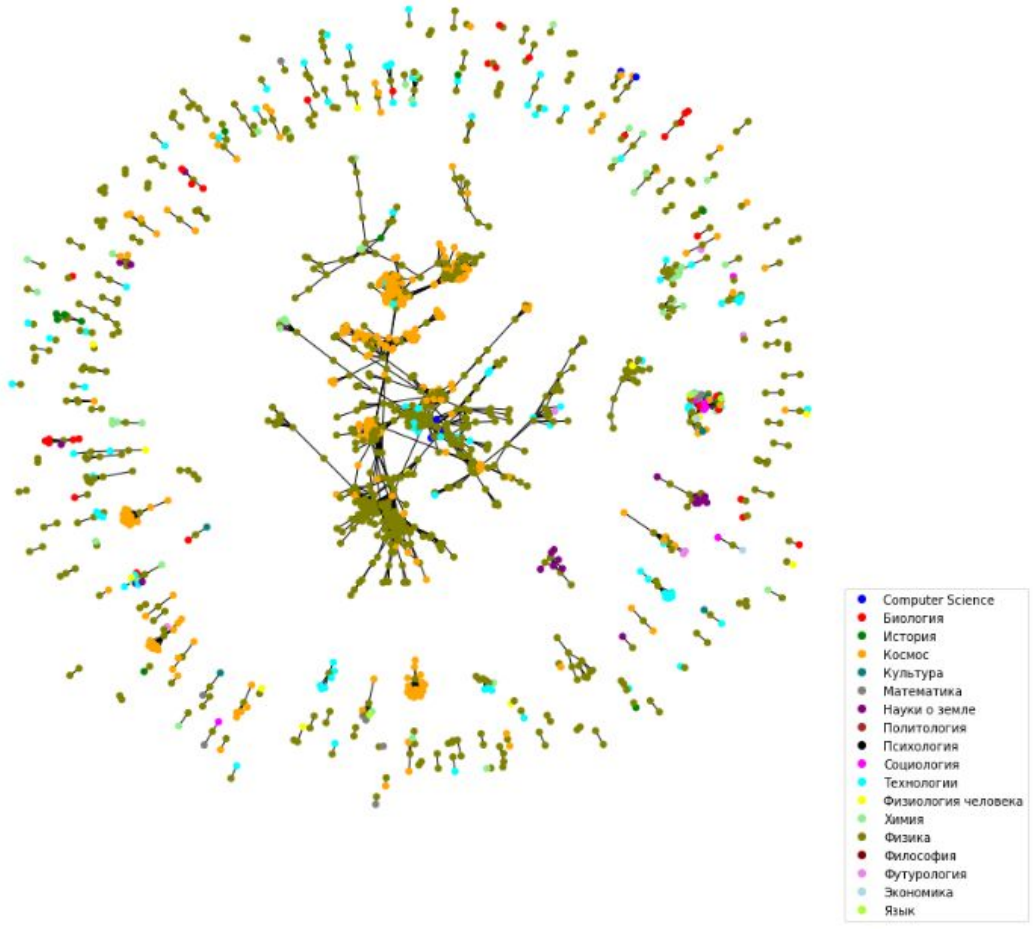
4.4. Результаты

Text similarity

Для анализа близости текстов из разных областей был построен граф, в котором ребра соединяют между собой пары текстов со значением косинусной близости выше порогового значения (0.5). В качестве веса ребра задано значение косинусной меры для пары текстов. Цвет каждой вершины графа отражает принадлежность текста к определенной рубрике. Такое представление позволяет проанализировать схожесть текстов в соответствии с их тематиками. Построенный граф содержит 9734 вершин. Таким образом, около 40% текстов корпуса оказываются схожими между собой (9734 из 24611 текстов). Из 24387 ребер графа 4972 ребра соединяют тексты из разных рубрик, что говорит о достаточно высокой степени междисциплинарных связей в корпусе. Для более наглядного представления взаимосвязей между текстами из разных

областей науки были построены отдельные графы для текстов каждой рубрики и их связей (Приложение 4).

На примере графа для текстов рубрики *Физика* можно проанализировать близость статей по *Физике* и статей из других областей. Граф содержит 4120 ребер, соединяющих 1437 текстов. Среди текстов по *Физике* выделяется группа, связанная с текстами рубрики *Космос* (1181 связей), значительное количество текстов связано с текстами, относящимися к рубрике *Технологии* (305 связей). Есть и отдельные случаи близости с текстами других рубрик (*Биология*, *Computer Science, Футурология, Химия*), но они не так многочисленны (50-80 связей).



В результате анализа графов для разных рубрик, представленных в корпусе, мы вывели наиболее выраженные взаимосвязи:

Рубрика	**Связанные рубрики**
Физика	Космос, Технологии
Космос	Физика, Технологии

Технологии	Физика, Космос, Биология
Биология	Физиология человека, Технологии История, Науки о земле
История	Язык, Биология
Computer Science	Физика, Технологии Биология, Физиология человека
Язык	История, Культура
Науки о Земле	Биология
Футурология	Космос, Технологии
Психология	Физиология человека

Интересно, что структура взаимосвязей текстов внутри разных рубрик оказывается различной: тексты некоторых рубрик образуют больше связей друг с другом (*Язык*, *Технологии*, *Космос*), в то время как тексты других рубрик более разобщены (*История*, *Computer Science*).

Какие тексты связывают различные науки между собой?

Рассмотрим отдельные случаи взаимосвязи различных областей науки на примере рубрик: *Космос* и *Биология*.

Значительное количество статей рубрики *Космос* вполне ожидаемо оказались близкими к статьям из области *Физики*, т.к. в них освещаются физические свойства космических объектов и физические явления в космосе (черные дыры, ядерные реакции в звездах[12]). Другая большая группа текстов рубрики *Космос* связана с текстами о *Технологиях*: в них представлены различные разработки и изобретения для космических полетов и исследований (летательные аппараты, устройства и приборы[13]). Более неожиданной оказывается взаимосвязь рубрик *Космос* и *Биология* на основе статей, посвященных исследованиям поведения растений и организмов в космических условиях (например, эксперимент по выращиванию картофеля в условиях,

[12] Здесь и далее приведены названия и ссылки на тексты, связывающие разные рубрики. Космос, Физика: “FAQ: Гравитационные волны и черные дыры. 7 фактов об исследованиях, необходимых для доказательства существования черных дыр” (https://postnauka.ru/faq/7342), “Действительно что-то узнать о черной дыре можно, только прыгнув в нее. Интервью с астрофизиком Сергеем Поповым о геометрии пространства и времени, горизонте черных дыр и проблеме доказательства их существования” (https://postnauka.ru/talks/26212).

[13] Космос, Технологии: ”Три жизни «Любопытства». Марсоход «Кьюриосити» — пять лет на Марсе” (https://chrdk.ru/tech/tri-zhizni-lyubopytstva), “Первый на Марсе. К 20-летию посадки марсохода «Соджорнер»” (https://chrdk.ru/tech/sojourner).

приближенных к условиям на Марсе[14]). Пересечение тематик *Космос* и *История* можно видеть на примере текстов об астрономических исследованиях древних народов[15].

Для *Биологии* самой близкой рубрикой стала *Физиология человека*, т.к. физиология входит в биологию. Помимо этого, значительная часть текстов по *Биологии* связана с текстами из рубрики *Технологии*. Предметом таких статей являются современные разработки в медицине (например, бионические протезы[16]), или других областях биологии (например, технологии искусственного производства мяса[17]). Также интересна связь между Биологией и Историей, возникающая на основе текстов о генетических особенностях древних людей и происхождении народов[18].

Граф взаимосвязи рубрик

Для анализа связей между рубриками был построен граф, в котором между двумя рубриками проводится ребро, если доля пар близких по косинусной мере текстов по отношению к общему числу пар текстов в рубриках (коэффициент близости рубрик S) превышает пороговое значение. В качестве весов для ребер графа задаются значения коэффициента S для пары рубрик.

Граф взаимосвязи рубрик отражает наиболее сильные связи между рубриками. Кроме того, положение вершин графа определяется с учетом весов ребер, соединяющих различные узлы. На графе выделяется группа рубрик, связанных с техническими науками: *Космос*, *Технологии*, *Физика*, *Computer Science* и *Футурология. Физика* также связана с *Химией*. Группа взаимосвязанных естественных наук включает *Биологию, Физиологию* и *Науки о Земле.* Помимо этого, выделяется группа областей *История, Политология, Социология, Культура*, связанных с рубрикой *Язык. Политология* и *Культура* также связаны с *Философией*. *Психология* соединяет между собой группы гуманитарных и естественных наук (т.к. она связана с *Социологией* и *Физиологией*). Гуманитарные и технические области связываются через *Математику* (*Математика* связана с *Социологией* и рубриками *Computer Science, Физика*).

---

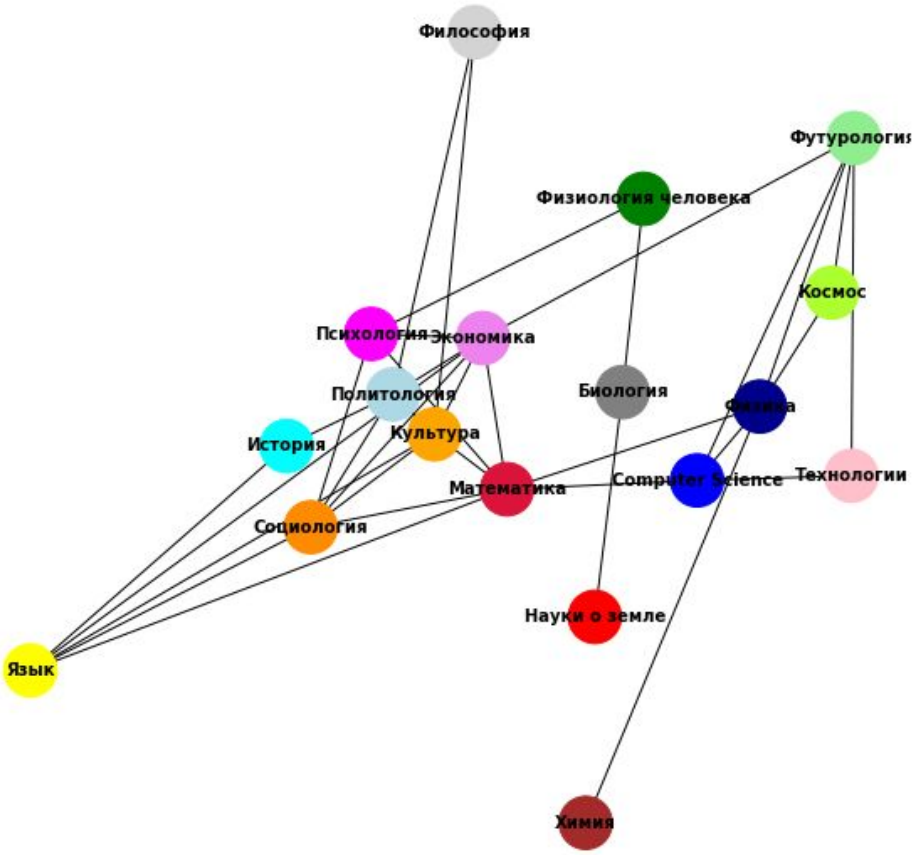
[14] Космос, Биология: “Куст картофеля вырос почти на Марсе. В Перу вырастили картофельный куст в условиях, близких к марсианским, и растение даже смогло образовать клубни” (https://chrdk.ru/news/kartofan-na-marse), “Ученые подтвердили возможность выращивания на Марсе картофеля” (https://habr.com/post/373243/).

[15] Космос, История: “Майя знали о нерегулярности синодического периода Венеры” (https://habr.com/post/396819/), “Майя рассчитали период вращения Венеры вокруг Солнца более тысячи лет назад” (https://chrdk.ru/news/maiya_rasschitali_period_vrashheniya_venery_vokrug_solntsa_bolee_tysyachi_let_nazad).

[16] Биология, Технологии: “Бионический протез Ossur управляется мыслью” (https://habr.com/post/379741/), “Новый бионический протез ноги без проблем справляется со спуском по лестнице” (https://habr.com/post/365645/).

[17] Биология, Технологии: “Бифштекс из искусственной говядины существенно подешевел” (https://habr.com/post/374877/), “Почти мясо. Как развиваются технологии создания мяса «в пробирке»” (https://chrdk.ru/tech/almost_meat).

[18] Биология, История: “Здесь инки не приставали. Молекулярные биологи ставят точку в споре о происхождении аборигенов острова Пасхи” (https://chrdk.ru/sci/heyerdahl-was-wrong), “Остров Пасхи, Америка и генетика” (http://polit.ru/article/2017/10/14/ps_rapanui/).



Таким образом, в результате анализа графов близости текстов и рубрик, построенных с помощью методов *text similarity*, мы выявили некоторые взаимосвязи между областями. В корпусе научно-популярных текстов выделяются группы взаимосвязанных технических, гуманитарных и естественно-научных рубрик, которые соединяются между собой за счет текстов на стыке нескольких областей.

Кластеризация

По результатам кластеризации методом k-средних и иерархической кластеризации на 18 кластеров можно заметить общие закономерности:

1. Выделяется большой блок, сочетающий в себе тексты рубрик *Технология* (~3000 из 6000 текстов рубрики Технология), *Космос*, *Физика*.
2. Тексты рубрики *Космос* разбиваются на несколько кластеров (два больших кластера и несколько небольших).
3. Выделяется блок гуманитарных наук (*История*, *Социология*, *Культура*).
4. Выделяется блок: *История* и *Биология*.
5. Выделяется блок: *Биология* и *Физиология*.

6. Выделяется блок: *Язык* (300 из 500 текстов).

При этом метод k-средних объединяет в один кластер тексты рубрик: *Технологии* (3000 текстов), *Биология* (2000 текстов), *Физика*, *Физиология*, *Космос*. Иерархическая кластеризация выделяет два больших кластера, включающих статьи из областей: *Технологии, Космос, Физика*.

Таким образом, аналогично результатам анализа графов на основе методов *text similarity,* при кластеризации хорошо выделяются группы технических и естественно-научных рубрик. Также явно заметна группа гуманитарных наук, несмотря на то, что эти области представлены в корпусе не таким большим числом текстов по сравнению с техническими областями. *Язык* оказывается более обособленной областью, как это было видно и по графу взаимосвязи рубрик, и значительная часть текстов этой рубрики не связана с другими гуманитарными областями.

Методы кластеризации явно выделяют связь между рубриками *История* и *Биология*, которая возможно возникает за счет текстов о происхождении народов и эволюции различных видов. Такая связь не была отражена в графе взаимосвязи рубрик, т.к. количество соответствующих текстов оказывается небольшим в сравнении с общим количеством текстов, относящихся к *Биологии*. Однако на графе близости текстов для рубрики *История* можно заметить отдельные связи с текстами по *Биологии*.

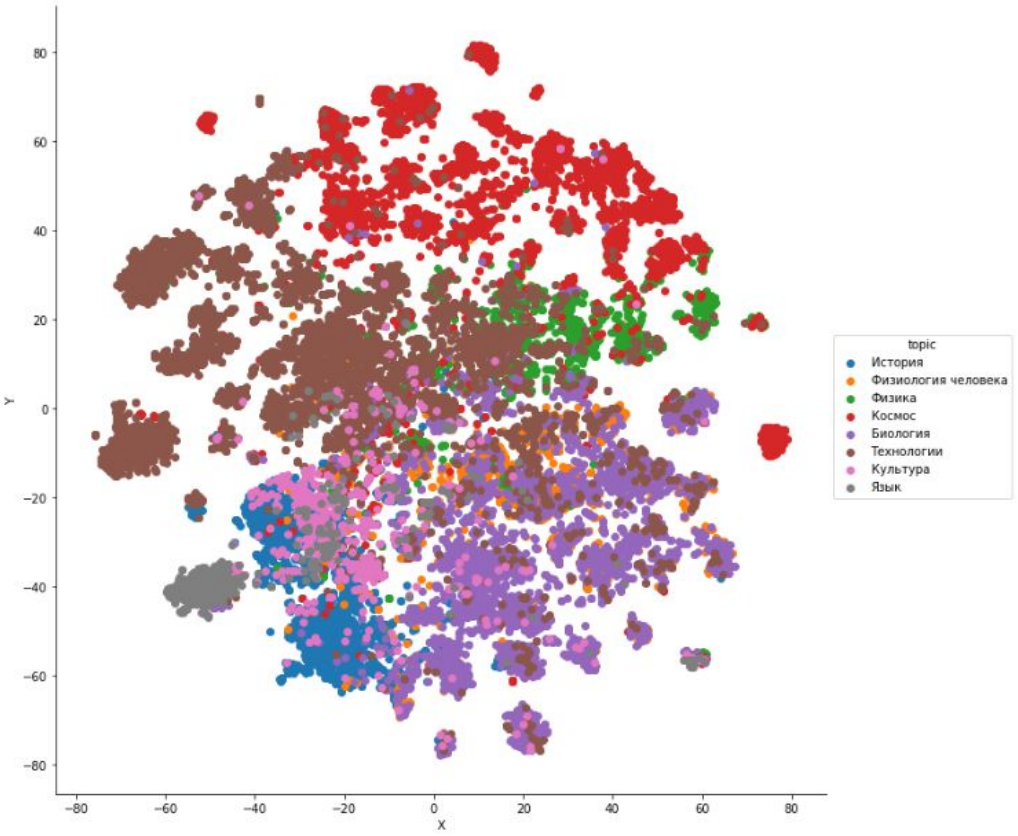
Разбиение рубрики Космос на несколько кластеров может быть вызвано тем, что эта область представлена в корпусе достаточно широко и включает в себя различные направления (связанные с физическими процессами в космосе, технологиями для освоения космоса и другими аспектами).

Снижение размерности

Для применения метода t-SNE были отобраны 8 наиболее крупных рубрик, количество текстов по которым превосходит 500: *История*, *Физиология человека*, *Физика*, *Космос*, *Биология*, *Технологии*, *Культура*, *Язык*.

В построенном двумерном пространстве хорошо выделяется кластер, соответствующий рубрике *Космос*, и кластер *Технологий*. Тексты по *Физике* пересекаются с текстами рубрик *Технологии* и *Космос*, рубрика *Физиология человека* смешивается с *Биологией*. Можно заметить, что среди текстов рубрики *Язык* часть выделяется в обособленный кластер, а другая часть смешивается с текстами из рубрик *История* и *Культура* (например, в смешанный кластер входят статьи, посвященные расшифровке письменности древних народов[19]).

[19] “Шифровка с острова Пасхи. Почему ученые не могут расшифровать письменность с острова Пасхи” (https://chrdk.ru/sci/pochemu_uchenye_ne_mogut_rasshifrovat_pismennost_ostrova_pashi).



Тематическое моделирование

Модель LDA, обученная на научно-популярном корпусе, выделила тематические слова для 18 тем. С помощью полученной модели мы определили темы для текстов корпуса, чтобы увидеть, какие рубрики относятся к каждой теме.

В таблице приведены списки из 15 наиболее значимых тематических слов для 18 тем, выявленных с помощью модели LDA, вариант интерпретации тем по словам и рубрики текстов, отнесенных к каждой из тем по результатам применения модели. Для каждой рубрики приведено количество текстов, попавших в данную тему.

Тема	**Тематические слова**	**Интерпретация**	**Рубрики**
0	частица, теория, квантовый, физика, энергия, нейтрино, взаимодействие, эксперимент, модель, электрон, состояние, физик, масса, атом, материя	Физика	Физика (630)
1	век, город, война, король, надпись, храм, музей, стать, год, робот, текст, слово, римский, войско, монета	История	История (598) Культура (118) Технологии (113)

2	вид, птица, учёный, животное, самка, самец, насекомое, тело, исследование, университет, человек, исследователь, яйцо, женщина, случай	Биология	Биология (1047) Технологии (115) Физиология человека (70)
3	учёный, материал, структура, метод, молекула, использовать, атом, свет, свойство, температура, помощь, поверхность, исследование, получить, волна	Физика, Химия	Физика (768) Технологии (543) Химия (315)
4	компания, устройство, технология, система, доллар, проект, производство, автомобиль, сша, использовать, новый, цена, использование, разработка, помощь	Технологии	Технологии (1832)
5	учёный, найти, обнаружить, остров, исследование, вид, находка, век, земля, современный, археолог, северный, древний, останки, место	История, Археология	История (981) Биология (645) Науки о земле (347)
6	число, пространство, точка, мочь, теория, математика, большой, вопрос, иметь, знать, вселенная, уравнение, существовать, гипотеза, модель	Математика	Математика (61)
7	звезда, галактика, планета, земля, солнце, объект, система, телескоп, масса, астроном, чёрный, волна, вселенная, излучение, дыра	Космос	Космос (1888) Физика (144)
8	должный, вопрос, советский, проект, стать, история, система, фильм, человек, говорить, корабль, компания, возможность, сделать, союз	Технологии	Космос (476) Технологии (139)
9	растение, неандерталец, учёный, реакция, энергия, соединение, химический, получить, водород, ядерный, вид, вещество, кислород, использовать, реактор	Химия, Биология, Физика	Биология (111) Физика (50) Технологии (44) Химия (33)
10	аппарат, марс, космический, поверхность, километр, луна, полёт, комета, земля, спутник, проект, получить, посадка, орбита, метр	Космос, Технологии	Космос (1149) Технологии (562)
11	язык, слово, человек, разный,	Язык	Язык (294)

	говорить, образ, русский, группа, языковой, некоторый, мочь, мозг, рука, друг, речь		Физиология человека (109)
12	человек, ребёнок, система, задача, социальный, игра, информация, исследование, слово, результат, разный, сеть, использовать, метод, исследователь	-	Технологии (637) Computer Science (191) Биология (191) Физиология человека (165)
13	наука, страна, россия, институт, научный, российский, вопрос, университет, лекция, государственный, общество, проблема, человек, стать, развитие	Наука (общая)	Технологии (234) История (220) Социология (171) Экономика (124) Физиология человека (105)
14	самолёт, двигатель, система, скорость, метр, испытание, проект, километр, полёт, новый, аппарат, получить, установка, составлять, крыло	Технологии	Технологии (1165) Космос (343)
15	клетка, ген, вирус, организм, днк, бактерия, болезнь, геном, учёный, человек, белка, исследование, белок, заболевание, мутация	Биология, Физиология	Биология (1292) Физиология человека (537)
16	книга, дело, город, история, свой, век, власть, вопрос, самый, иметь, мочь, культура, знать, стать, говорить	История, Культура	История (301) Культура (218)
17	человек, исследование, мозг, пациент, группа, уровень, результат, болезнь, ребёнок, поведение, состояние, случай, сон, фактор, активность	Психология, Физиология	Физиология человека (547) Биология (358) Психология (146)

По тематическим словам можно интерпретировать большую часть тем.

В построенной тематической модели хорошо выделяются темы, соответствующие техническим и естественным наукам. Среди гуманитарных областей выделяются темы, связанные с Историей и Языком, при этом некоторые темы объединяют Историю и Культуру.

Интересно, что тематическая модель выделяет различные аспекты внутри одной тематики. Например, выделено несколько тем, связанных с Космосом: астрономия и физические явления в космосе (7), космические технологии (10). Тексты рубрики *Технологии* попадают в темы, относящиеся к разным наукам, т.к. в корпусе представлено большое число текстов, посвященных применению технологий в различных областях (Биология, Космос, Физика, Культура).

Выводы

Применение различных методов компьютерной лингвистики позволило проанализировать взаимосвязи между различными науками в корпусе научно-популярных текстов. Было выявлено, что автоматические методы связывают между собой как тексты из близких областей (Физика, Технологии, Космос), так и тексты, относящиеся к не связанным на первый взгляд областям (Биология, История).

Такой результат может быть вызван тем, что в научно-популярном жанре часто освещаются явления на стыке наук. Однако полученные выводы о взаимосвязях тем могут послужить основой для разработки уточненной классификации научно-популярных текстов с учетом пересечения различных областей науки и выделения отдельных направлений внутри некоторых областей.

5. Оценка readability

5.1 Что такое readability? Постановка задачи

Readability — это удобочитаемость, то есть величина, показывающая, насколько понятным и легким для прочтения является текст. Чаще всего оценки readability опираются на различные статистические характеристики текста: среднюю длину предложения, среднюю длину слов, количество “сложных” слов (то есть с большим количеством слогов), и так далее (Crossley 2017).
. В компьютерной лингвистике readability — традиционная область: первые метрики появились еще в середине XX века. Способы оценки readability во многом создавались как инструмент для подбора текстов в образовательных целях: для школьной программы и обучения иностранному языку. Большинство популярных метрик ранжирует тексты как раз согласно уровням школьного образования: 1 — понятный первокласснику, 5 — пятикласснику, 10 — ученику выпускных классов, больше 12 — студенту университета и старше. Для большинства популярных метрики эти уровни ассоциированы с классами американской школы, а при ранжировании текстов для обучения иностранному языку они соотносятся с уровнями владения языком (традиционные A1, A2, B1 и так далее) (Gallagher et al. 2017, Solorio et al. 2017). Современные исследования в этой области направлены на поиск новых характеристик, помогающих точнее определять уровень сложности текста.

Оценка readability научно-популярных текстов кажется нам важной задачей по нескольким причинам. При помощи оценки readability мы можем понять, насколько однородными по сложности являются научно-популярные публикации. Как правило,

их уровень удобочитаемости может быть очень неодинаковым даже в рамках одного сайта: в них может быть разная концентрация терминов, удачная или не очень редактура, в конце концов, затронутая тема может быть известна читателям со школы — а может быть понятна только узким специалистам. Некоторые ресурсы учитывают это и вводят собственную оценку сложности материалов (например, сайт «N+1»), правда, основанную на субъективном мнении редактора: “Наша сложность - это некоторая коллективная оценка редакции интеллектуальных усилий, которые понадобились редактору для написания заметки” (Readability). В перспективе маркирование научно-популярного текста уровнем сложности могло бы быть полезным для читателя: он может еще до ознакомления с материалом понять, имеет ли смысл ему читать эту статью, или он скорее всего она кажется для него слишком сложной. Иногда заголовок и даже первые несколько вводных абзацев могут создавать иллюзию понятности текста, тогда как по мере чтения он становится все сложнее и непонятнее, и напоминать скорее сугубо научную статью, чем научно-популярное изложение. Кроме того, можно предположить, что тексты, посвященные разным областям науки (физика, биология, литературоведение, история и т.д.) могут иметь разный уровень сложности, либо их “сложность” будет определяться по-разному.

Специфика оценки readability научно-популярных текстов прежде всего состоит в том, что традиционные метрики readability предлагают ранжирование, как говорилось выше, в соответствии со школьными классами (или уровнями владения языком), а научно-популярные ресурсы обычно позиционируют себя как ориентированные на студентов и людей с высшим образованием. Соответственно, стоит ожидать, что значения стандартных метрик будут выходить за пределы традиционной шкалы (1-12). Поэтому представляется затруднительным оценить сложность текстов, которые *уже* сложные в соответствии со стандартными измерениями. Кроме того, в российских научно-популярных медиа далеко не всегда авторами текстов являются профессиональные журналисты: часто статьи пишут ученые, у которых, как правило, неодинаковый опыт в популяризации. На некоторых сайтах (postnauka.ru, polit.ru) представлены минимально отредактированные расшифровки лекций — с одной стороны, они не задумывались как текст, с другой — их все равно читают и оценивать их “понятность” тоже имеет смысл.

Для решения этих проблем мы предлагаем два важных изменения. Во-первых, разделять тексты на **три уровня** сложности:

1) базовый: понятный ученикам старшей школы и людям без высшего образования, эрудированным, но не специалистам. Такие тексты написаны на общие, неспециальные темы, понятны с первого прочтения и не требуют специфических знаний.
2) продвинутый: тоже могут быть понятны людям без специальных знаний, но требуют более вдумчивого чтения и погружения в тему, возможно, ознакомления с некоторыми терминами

3) сложный: тексты на узкоспециализированные темы, с обилием терминов и сложной информации, полностью понятные специалистам либо тем, кто специально увлекается предметом

Три уровня сложности показались нам оптимальным решением, так как бинарная классификация была бы слишком простой для нашей задачи и не отражала всей специфики наших текстов, а более дробная напротив, могла оказаться слишком путаной и создавать трудности при аннотировании данных. Кроме того, тернарная классификация соотносится с основными уровнями образования: школьник, студент, специалист. Во-вторых, помимо стандартных метрик, о которых будет рассказано ниже, мы ввели несколько новых характеристик текста: количество разных частей речи и количество разговорных слов. Такой набор признаков поможет изучить особенности научно-популярных текстов, а также может служить хорошей базой для машинного обучения. Мы предполагаем, что такая тернарная классификация (то есть разбиение на три группы) будет оптимальной и позволит максимально адекватно ранжировать научно-популярные тексты по сложности. Иначе говоря, мы рассчитываем, что научно-популярные тексты действительно можно разделить на три уровня сложности, основываясь на выведенных нами количественных признаках. В идеале хотелось бы получить на основе наших признаков, классификатор, который может предсказывать сложность научно-популярных текстов, опираясь на вышеуказанные признаки.

5.2 Способы измерения: основные метрики

Традиционные способы оценки сложности текста оптимизированы под статистические расчеты на основании различных характеристик текста. Наиболее полный список измеряемых параметров включает в себя следующие пункты:

- длина текста в символах
- длина текста в словах
- количество предложений в тексте
- среднее количество слов в предложении (и обратная пропорция — «количество предложений на слово»)
- среднее количество символов в словах
- среднее количество символов в предложениях
- количество слогов в тексте
- среднее количество слогов в предложениях
- среднее количество слогов в словах
- количество «сложных» слов (в которых более 3 слогов для английского языка и более 4 для русского)
- процент «сложных» слов (Saggion 2017).

Начиная с 40-х годов прошлого века учеными было разработано несколько метрик для оценки readability. Самая популярная из них — индекс удобочитаемости Флеша (Flesch reading Ease – FRE), которая применяется во многих сервисах для оценки сложности текстов, в частности, ее использует Word от Microsoft Office.

Индекс удобочитаемости Флеша

Индекс удобочитаемости Флеша рассчитывается следующим образом:

$$FRE = 206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

где ASL — это средняя длина предложений, а ASW — средняя длина слова в слогах.
Эта метрика была разработана в 1975 году. Предполагается, что она должна возвращать значение в пределах от 100 до 0. Чем меньше полученная цифра, тем сложнее текст. Например, значение 60 характеризует текст как довольно простой, а 20 — достаточно сложный (Readability).
Существует еще один вариант той же формулы — «уровень Флеша-Кинкайда» (Flesch-Kincaid grade level). Он выглядит вот так:

FKG = 0.39*ASL + 11.8*ASW.

и возвращает предполагаемый уровень школьного образования в американской системе, необходимый для понимания текста (Readability).

Общей проблемой для этих и нижеперечисленных метрик и является то, что они разработаны только для английского языка. Это особенно заметно на индексе удобочитаемости Флеша: при указанных коэффициентах получаемый результат во многих случаях выбивается из принятой шкалы значений и показывает то отрицательное число, то более 100. В русском языке слова обычно длиннее, а предложения в среднем короче, чем в английском, так как служебных слов в среднем используется меньше (Reynolds 2016). Поэтому для индекса Флеша, как и для других метрик, существуют адаптации под русский язык. Для FRE наиболее распространенная версия была разработана И. В. Оборневой (Reynolds 2016) и предлагает следующие коэффициенты:

$$FRE = 206.835 - (1.3 \times ASL) - (60.1 \times ASW)$$

С такими значениями русскоязычные тексты значительно реже выбиваются за пределы значений шкалы.

Индекс Дэйл-Челл

Один из самых ранних индексов readability — Dale-Chall readability formula, был разработан еще в 1948 году. Рассчитывается по формуле:

$$0.1579\left(\frac{\text{difficult words}}{\text{words}} \times 100\right) + 0.0496\left(\frac{\text{words}}{\text{sentences}}\right)$$

Для русского языка здесь изменяется принцип подсчета «сложных слов»: к таковым относятся те, у которых 4 и более слога, а не 3 и более, как для английского. Отличается и шкала: она показывает значение от 4 до 10, где наиболее высокое число соответствует уровню студента колледжа (Saggion 2017).

Индекс Gunning Fog

Индекс был разработан Робертом Ганнингом в 1952 году. Рассчитывается он по следующей формуле:

$$0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right) + 100\left(\frac{\text{complex words}}{\text{words}}\right)\right]$$

то есть учитывается среднее количество слов в предложении и отношение сложных слов к их количеству в тексте. Возвращает число от 17 до 6, чем выше значение, тем сложнее текст (также в соответствии со школьной программой) (Saggion 2017).

Индекс Колман-Лиау

Рассчитывается следующим образом:

$$CLI = 0.0588L - 0.296S - 15.8$$

L – среднее количество букв на 100 слов, а S — среднее количество предложений на 100 слов. Похож на Автоматический индекс удобочитаемости (automated readability index (ARI)):

$$ARI = 4.71 \times \frac{C}{W} + 0.5 \times \frac{W}{S} - 21.43$$

ARI использует среднее количество символов в слове (C/W) и среднее количество слов в предложении (W/S). В нашем исследовании мы не использовали ARI, потому что его результаты оказывались практически идентичны CLI (Saggion 2017).

Индекс SMOG (Simple Measure of Gobbledygook)

Эта метрика была разработана в 1969 году как альтернатива индексу Gunning fog. Некоторыми исследователями оценивается как более точная и удачная, чем индекс удобочитаемости Флеша.
Рассчитывается следующим образом:

$$\text{grade} = 1.0430\sqrt{\text{number of polysyllables} \times \frac{30}{\text{number of sentences}}} + 3.1291$$

Для русского языка адаптивность опять же достигается за счет изменения критерия отнесения слова к «сложному»: если для английского языка это 3 слога, то для русского традиционно берут 4. Как FKG и все вышеперечисленные метрики, возвращает необходимый для понимания текста уровень школьного образования (Saggion 2017).

Как уже говорилось выше, метрики имеют свои ограничения. Так как их шкалы рассчитаны на уровни школьного образования, а научно-популярные тексты из русскоязычных медиа, как правило, не предназначены для школьников младших и средних классов, мы предполагаем, что по всем этим метрикам наши тексты будут оценены как предназначенные для старшеклассников либо студентов, и даже выйдут за пределы шкал. Таким образом, в нашем исследовании мы не будем пытаться присвоить научно-популярным текстам уровень сложности в соответствии со школьной программой, но изучим, как изменяются эти значения в соответствии с нашей классификацией текстов по сложности.

5.3. Актуальные исследования

Исследования по нашей теме можно разделить на две основные группы. К первой относятся, ориентированные на анализ текстов на английском языке. Авторы сочетают в своих работах различные подходы — методы машинного обучения, ручные правила, применение нейронных сетей. Другая часть работ относится к исследованиям русского языка и чаще всего ориентирована на преподавание русского как иностранного (Batinic et al. 2016, Reynolds 2016). Кроме того, при помощи описанных выше метрик активно анализируется сложность юридических, медицинских и других специальных текстов (Vargas et al. 2017, Smith et al. 2017, Loughran, McDonald 2011) . Но такие работы не содержат в себе исследовательской составляющей: они не стремятся анализировать сами метрики, дополнить их новыми, а только констатируют высокую сложность специальных текстов — и так, в принципе, очевидную. Поэтому такие работы, несмотря на кажущуюся близость исследовательского материала к нашему, не представляют для нас большого интереса.

Одной из самых содержательных работ оказалась статья с EACL-2017 «A Multi-task Approach to Predict Likability of Books» (Solorio et al. 2017). Авторы поставили перед собой задачу бинарной классификации: при помощи традиционных алгоритмов машинного обучения и нейронных сетей предсказать успех литературного произведения (то есть отнести текст к классу «успешных» или «неуспешных»). На

успех книги среди читателей влияет множество факторов, начиная от сюжета и заканчивая личностью автора— насколько он известен и авторитетен. Чтобы поделить тексты на успешные и неуспешные, авторы взяли за основу оценки книг на сайте Goodreads (авторы одной из предшествующих работ на схожую тему для разделения книг на «успешные» и «неуспешные» брали количество скачиваний с Project Guttenberg). Для повышения качества классификации авторы, с одной стороны, выделили несколько признаков, которые помогли бы обучить классификатор, с другой, прибегли к методам глубокого обучения посредством рекуррентной нейронной сети. Для своего датасета авторы взяли коллекцию из 800 текстов из Guttenberg project и связали их с оценками с сервиса Goodreads. Посчитав средний рейтинг книг на сайте и количество отзывов, авторы установили, что «успешная» книга должна иметь оценку не менее 3, 5 баллов (из 5). Книги с меньшей оценкой относились к классу неуспешных. Среди признаков для обучения был стандартный набор количественных характеристик текста (средняя длина слов и предложений в тексте, количество «сложных» слов,, и т.д.), метрики readability для английского языка, тональность текста, и так называемый «писательский стиль». При оценке книг авторы разбивали текст на н-граммы, измеряли TF-IDF, а также оценивали тональность текста на основе данных SentiWordNet: текстам присваивалась одна из категорий «нейтральный», «позитивный», «негативный». Кроме того, авторы применили инструмент анализа тональности SenticNet, разбили слова на н-граммы и назвали получившуюся модель Bag of concepts. Свою нейронную сеть авторы обучали при помощи модели Doc2Vec и назвали полученные результат Book2Vec. Так как, по утверждению авторов, рекуррентные нейронные сети плохо справляются с длинными последовательностями, авторы применили многозадачный подход: сеть рассматривала каждый документ как последовательность из последовательностей. На этапе эксперимента авторы смешали вместе книги разных жанров, и разделили датасет на обучающую и тестовую выборки. Всего они натренировали 25 моделей со случайными гиперпараметрами и весами. Нейронная сеть показала результаты чуть лучше, чем ручные признаки (и те, и те колебались в пределах точности 0,7). Но наилучший результат (0,73) дал алгоритм, который совмещал в себе «ручные» признаки и глубокое обучение. Авторы показывают, что в своей наиболее успешной модели самые значительные веса они давали признакам, связанным с тональностью и н-граммам, свидетельствующим о наличии диалогов в тексте (обращения, «он сказал», и т.д.). Книги, наполненные диалогами зачастую оказывались успешными, что позволяло предположить, что читателям нравится читать чужие беседы. Также популярными оказались авторы с наибольшей «плотностью» письма (эта характеристика включала в себя длину слов и предложений, количество восклицательных и вопросительных знаков, повторяемость лексики). Кроме того, при помощи ряда экспериментов, авторы пришли к выводу, что их алгоритму достаточно обработать порядка 200 предложений, чтобы дать свою оценку тексту. Статья вполне отражает текущее состояние в обработке текстов на естественном языке и оценки его «читабельности»: наилучший результат дает совместное использование «ручных» методов и машинного обучения.

Задача, которую ставят перед собой авторы на первый взгляд мало перекликается с нашей. Но те аспекты, на которые обратили внимание авторы при подборе дополнительных признаков, являются для нас очень значимыми: действительно, читатели научно-популярных текстов, возможно, лучше воспринимают информацию, если автор (журналист или ученый) использует разговорные выражения, эмоциональные слова, при помощи которых проще воспринимать сложную информацию.

Из других заметных работ стоит отметить статью “Creating an extensible, levelled study corpus of Russian” (Batinic et al. 2016) авторства Долорес Батиник, Сандры Бирцер и Хейке Цинзмейстер. Их целью было разработать классификатор текстов на русском языке. Это исследование также служит задаче подбора текстов для обучения русскому языку как иностранному. Авторы поделили тексты на две группы сложности и соотнесли их с уровнями владения языком:

Class	TRKI	CEFR	Sem	#Texts
I	elementary	A1	1st	43
	basis	A2	2nd	43
	1	B1	2nd	50
II	2	B2	3rd	38
	3	C1	4th	30
	4	C2	indep	5

Авторы не только обучили классификатор, но и создали онлайн-сервис Lestcor (на май 2018 года сайт http://lestcor.com был недоступен из-за истечения срока пользования доменом), в котором можно ввести текст и получить информацию об его уровне сложности и разли Исследования авторов были нацелены на оценку readability русских текстов, поэтому для большинства метрик примененены корректировки коэффициентов для русского языка, а также использовались словари “общих” слов и разметка текста по частям речи. В результате этого появился Lestcor -- сайт, на котором можно ввести текст и получить информацию и его метриках readability и предполагаемый уровень сложности (тоже исходя из американской школьной системы).

5.4. Дополнительные метрики

С учетом особенностей наших текстов и предыдущего опыта исследований, мы решили ввести несколько дополнительных признаков, которые могли бы улучшить качество классификации. Все они -- количественные: отражают степень присутствия того или иного вида слов в текстах.

Чтобы посчитать эту информацию корректно, нам было необходимо лемматизировать все тексты -- то есть привести слова в них к начальной форме.

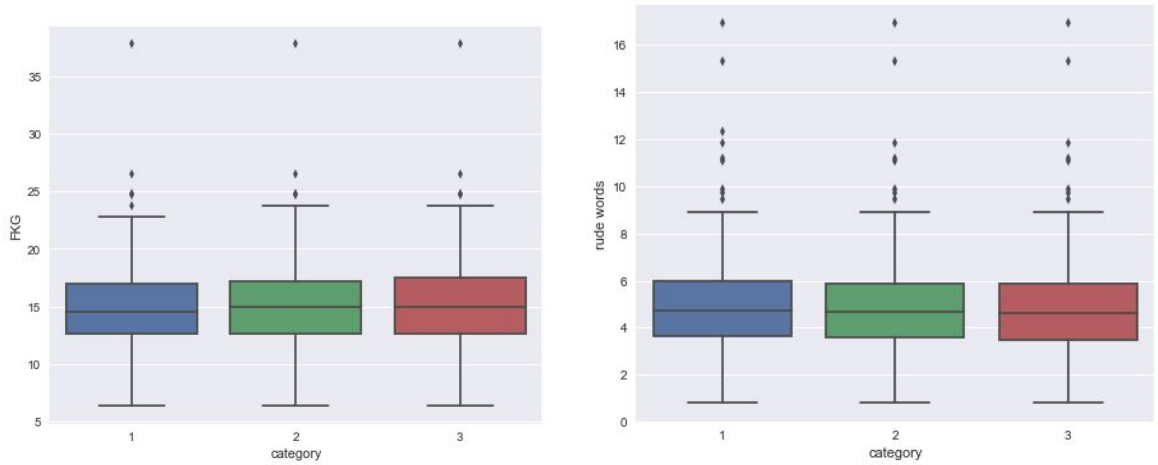
Обработанные таким образом тексты мы сравнили с несколькими специальными списками слов: списком наиболее частотных слов русского языка (предполагается, что такие слова обычно покрывают около 80% среднестатистического текста (Batinic et al. 2016), списком “разговорных” и “грубых” слов, собранных из Викисловарей. Подсчитывалось количество слов из каждого списка в каждом тексте и делилось на общее количество слов в этом тексте.
Кроме того, мы предположили, что в сложных текстах может встречаться большое количество специальных слов, образованных при помощи приставок из системы СИ. Поэтому, взяв список этих приставок, при помощи регулярных выражений мы посчитали, как часто такие слова встречаются в каждом тексте.
Мы выдвинули еще одну гипотезу: для прочтения окажутся более сложными тексты, в которых больше существительных, а тексты с большой концентрацией глаголов и союзов окажутся проще для прочтения. Чтобы проверить ее, при помощи pymorphy2 мы разметили лемматизированные тексты по частям речи и посчитали долю существительных, прилагательных, глаголов, союзных слов и предлогов в каждом тексте.

5.5 Данные

Наши данные -- это корпус из **529** текстов-публикаций с русскоязычных научно-популярных ресурсов (ПостНаука, N+1, Индикатор, Чердак, Geektimes, Полит.ру). С этих сайтов были собраны все подходящие нам по формату и содержанию тексты (например, тесты и короткие анонсы в них не попали), и часть из них на основе экспертной оценки была разбита на три класса, в соответствии с нашей классификацией: простые, средние и сложные. Как было написано выше, для обучение мы использовали только количественные признаки.

5.6 Визуализация и отбор признаков

Перед машинным обучением мы решили визуализировать распределением признаков для разных классов. Мы сделали боксплоты (“ящики с усами”), показывающие интерквартилиальное распределение признака. Признаки рассчитывались при помощи пайплайна, написанного нами на языке программирования Python. Оказалось, что для большинства признаков между классами нет различий. Например, как здесь:

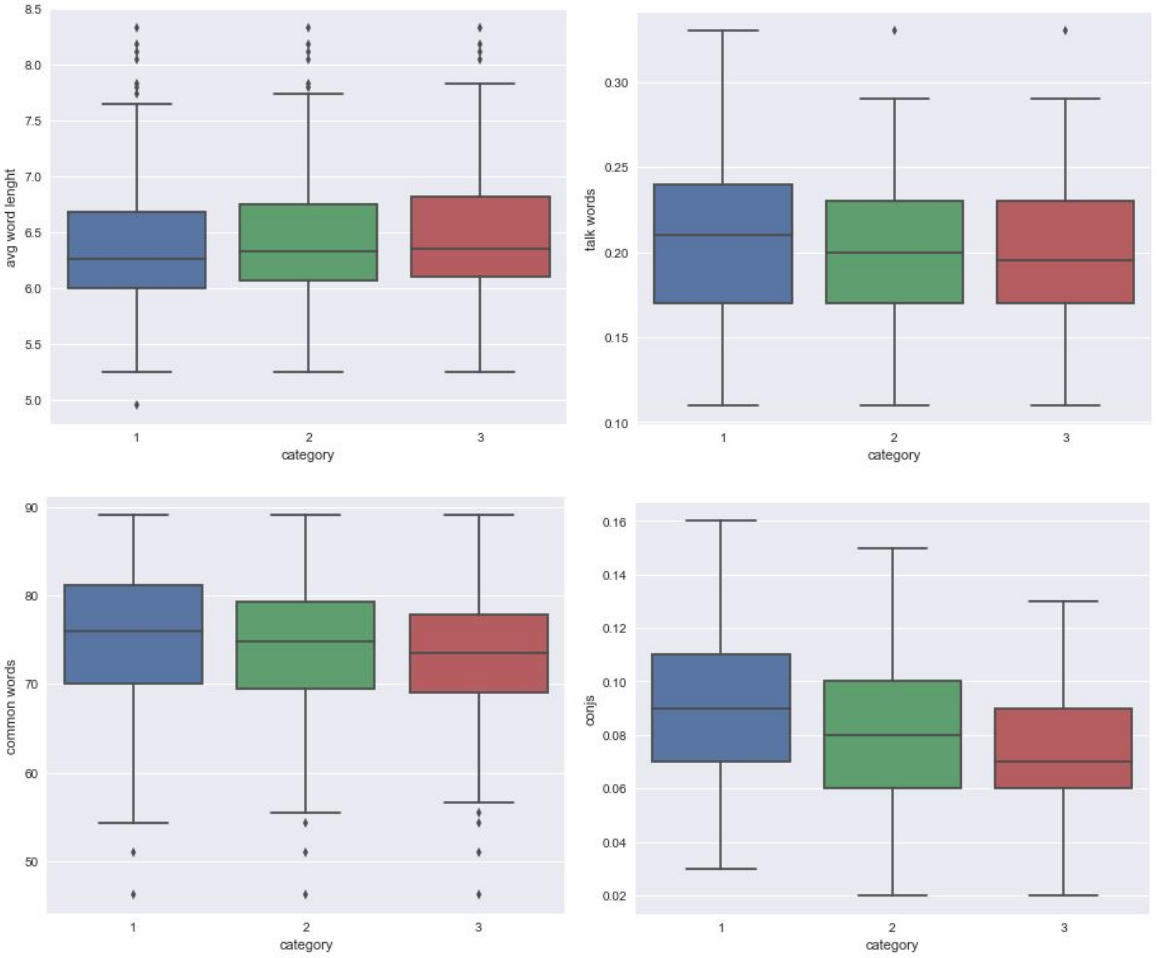


Распределение признака FKG, распределение признака "грубые слова". Здесь и далее: 1 — "простые тексты", 2 — "средние по сложности тексты", 3 — "сложные тексты"

Для некоторых признаков результат все же оказался чуть лучше, но не значительно. Например, для признака "грубые слова" медианы и границы "коробочек" находятся на одном уровне (5, 4 и 6 соответственно).

Визуализация признаков позволила проверить ряд гипотез. Так, для признака "разговорные слова" видно, что тексты из категории 1 (самые простые) имеют более вытянутые "усы", то есть такие тексты содержат больший процент разговорных слов, хотя если обратить внимание на значения на оси у, становится понятно, что речь все равно идет об очень близких значениях -- 0, 24 и 0, 22. "Общие слова" также чаще встречаются в более легких текстах, а в текстах большей сложности слова в целом чуть длиннее. То есть в целом общие и разговорные слова действительно покрывают более значительную часть "простых" текстов, чем все остальных.

Кроме того, как показал анализ, в текстах большей сложности используется меньше союзов но больше существительных, тогда как объем употребления глаголов практически не различается. Важным наблюдением также является, что традиционные метрики -- индекс Флеша, индекс Колиман-Лиау и другие очень слабо отображают различия между текстами. Вопреки ожиданиям, наличие приставок из СИ тоже мало сказывается на сложности текста.



Распределение признаков “средняя длина слова”, “разговорные слова”, “общие слова”, “союзы”

Те различия, которые мы видим между классами в некоторых признаках, кажутся довольно незначительным. Может ли такое распределение признаков стать хорошей основой для машинного обучения? Это кажется сомнительным, но мы попробуем.

5.7 Обучение

5.7.1. Тернарная классификация

Для машинного обучения мы использовали метод ближайших соседей и метод опорных векторов и случайный лес.

Мы натренировали 10 моделей с разными параметрами и разными комбинациями признаков (беря все подряд или исключая наиболее незначительные). Но так или иначе, наилучший результат оказался следующий:

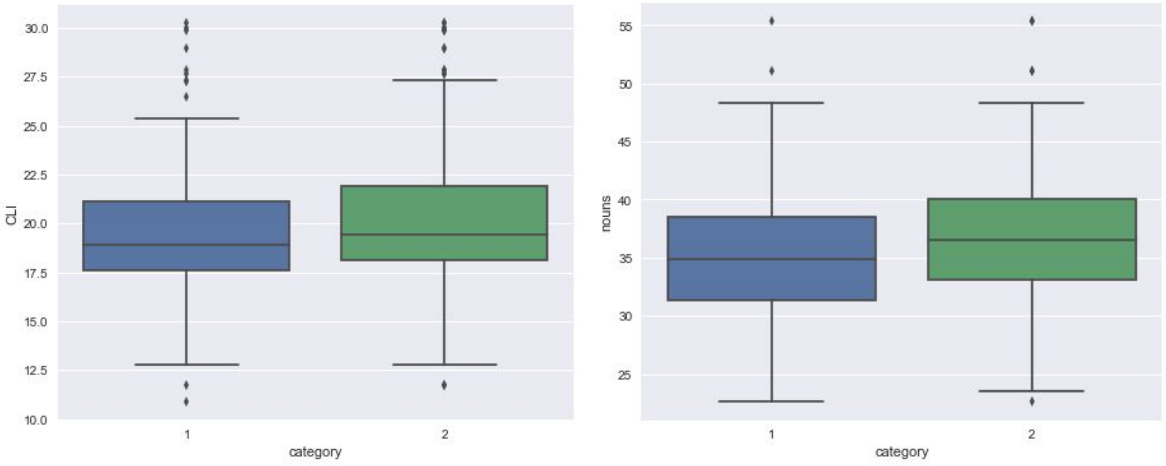
	precision	recall	f1-score	support
1	0,39	0,59	0,47	86

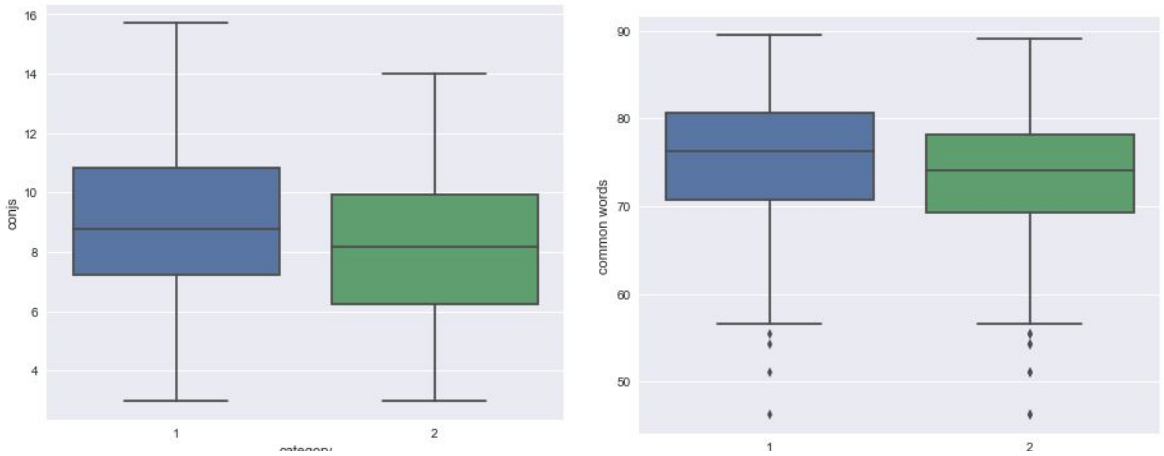
2	0,34	0,14	0,20	71
3	0,27	0,25	0,26	55
avg / total	0,34	0,35	0,33	212

После анализа признаков стало понятно, что отличать друг от друга тексты по уровню сложности алгоритмам машинного обучения будет непросто. Так и оказалось: наилучший результат по f-мере для модели -- 33%, что практически идентично случайному выбору.

5.7.2. Бинарная классификация

При визуализации данных мы обратили внимание на то, что, несмотря на в принципе слабое различие между категориями (по нашим признакам), классы 2 и 3 были более похожи друг на друга. Кроме того, заметно, что 3 класс выделяется несколько хуже остальных двух. Отсюда мы предположили, что, возможно, тернарная классификация текстов на данном этапе невозможна, но более эффективной может оказаться бинарная классификация, то есть разделение текстов на простые и сложные. Мы объединили категории 2 и 3 в одну, и посмотрели, как теперь распределяются признаки.





Распределение признаков CLI, “существительные”, “союзы”, “общие слова”. 1 — “простые тексты”, 2 — “сложные тексты”

Мы использовали те же способы машинного обучения, что и для тернарной классификации и натренировали такое же количество моделей. Наилучший результат оказался следующий:

	precision	recall	f1-score	support
1	0,62	0,29	0,39	63
2	0,65	0,89	0,75	96
avg / total	0,64	0,65	0,61	159

F-мера составила 61%, что немногим лучше случайного, но в целом предполагает более качественный результат, чем обучение с тернарной классификацией. Для других классификаторов и моделей результат был походим и отличался на 1-2%. Видно, что объединенная категория из более сложных текстов классифицируется лучше, чем простые тексты.

5.8. Выводы

На основании экспертной выборки мы попытались разделить тексты из научно-популярных русскоязычных медиа на несколько категорий сложности. Чтобы понять, можно ли воспроизвести эту классификацию при помощи компьютерной лингвистики, мы выделили несколько признаков, на которых попытались построить алгоритм машинного обучения. Предварительные результаты оказались невысокими: наилучший из них — 33%, то есть идентичный случайному. Мы предположили, что проще будет поделить тексты только на две группы, простые и сложные, и попробовали решить задачу бинарной классификации. Результаты оказались немного

лучше, но незначительно. Отсюда следует вывод, что тернарная классификация научно-популярных текстов не является оптимальной (по крайней мере, на используемых алгоритмах и моделях), а бинарная не вполне может стать альтернативой ей. Дальнейшие исследования будут направлены на разработку новых признаков для обучения и проверки моделей, а также внедрение категориальных признаков-тем, так как гуманитарная или естественно-научная направленность текста может служить хорошим маркером его уровня сложности.

Кроме того, анализ получившихся в результате подготовки признаков их средних значений показал, что большинство научно-популярных текстов отличаются достаточно высоким уровнем сложности, то есть наша теория о том, что такие тексты в целом выходят за рамки школьной программы подтверждается показаниями индексов ридабилити и долей наиболее распространенных слов. Метрики ридабилити опираются на разные количественные признаки, но каждый из использованных индексов показывал, что научно-популярные тексты относятся к “сложным”. А значит, для корректного разделения такого материала “более сложный” и “менее сложный” требуется серьезная разработка дополнительных признаков, связанных, в частности, с семантикой и смысловой наполненностью текстов.

Источники

1. http://dfgm.math.msu.su/people/konyaev/
2. https://chrdk.ru/about
3. http://www.media-atlas.ru/contentmanager/?a=view&id=2885
4. https://ru.wikipedia.org/wiki/%D0%A5%D0%B0%D0%B1%D1%80
5. https://istina.msu.ru/profile/nicola/
6. https://nplus1.ru/difficult/

Литература

1. *Ahrenberg, Lars* . "Term extraction: a review", 2013
2. *Batinic, Dolores, Birzer, Sandra*, "Creating an extensible, levelled study corpus of Russian", 13th Conference on Natural Language Processing (KONVENS 2016).
3. *Crossley, Scott A., Stephen Skalicky*, Mihai Dascalu, Danielle S. McNamara & Kristopher Kyle, "Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas", Discourse Processes, Volume 54, 2017.
4. *Gabrilovich E., Markovitch S.* Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, Proceedings of the 20th International Joint Conference on Artificial Intelligence, 6–12, 2007.
5. *Gallagher, Tiffany, Xavier Fazio, Katia Ciampa*, "A Comparison of Readability in Science-Based Texts: Implications for Elementary Teachers", Canadian Journal of Education / Revue canadienne de l'éducation 40:1, 2017.
6. *Gomaa, Wael,* H., Fahmy, Aly A. A Survey of Text Similarity Approaches. International Journal of Computer Applications 68(13):13-18, April 2013.
7. *Lample Guillaume, Ballesteros Miguel, et al*, Neural Architecture for Named Entity Recognition. Proceedings of NAACL, 2016.
8. *Loughran, Tim, Bill McDonald,* "Measuring Readability in Financial Disclosures", Journal of Finance, 2011.
9. *Manning, Christopher D.*, Raghavan, Prabhakar, Schütze, Hinrich, Introduction to Information Retrieval. Cambridge University Press, 2008.
10. *Maaten van der,* L.J.P. , Hinton G.E. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.
11. *Plavén-Sigray P., Matheson GJ., Schiffler BC., Thompson WH.,* "The readability of scientific texts is decreasing over time", 2017.
12. *Solorio, Thamar, Manuel Montes-y-Gómez*, Suraj Maharjan, John Edison Arevalo Ovalle, Fabio A. González, «A Multi-task Approach to Predict Likability of Books», e 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017.

13. *Smith, K., P. Buchanan, P. McDonald*, “How easy is it for a lay audience to read medical journals? A survey of the readability scores of a sample of research papers on diabetes”, 2017.
14. *Shen, Yuru*, “On Improving Text Readability by Creating a Personal Writing Style”, English Language Teaching, Vol 10, No, 2017.
15. *Saggion, Horacio,* “Automatic Text Simplification”, 2017
16. *Reynolds, Robert*, “Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories”, 11th Workshop on Innovative Use of NLP for Building Educational Applications, 2016.
17. *Vargas, Christina R., Joseph A.Ricci*, Michelle Lee, Adam M.Tobias, Daniel A.Medalie, Bernard T.Lee, “The accessibility, readability, and quality of online resources for gender affirming surgery”, Journal of Surgical Research, Volume 217, September 2017.
18. *Воронцов К.В.* Вероятностное тематическое моделирование: обзор моделей и регуляризационный подход. 2017.
19. *Гринёв-Гриневич, С. В.* “Терминоведение”, Учебное пособие. — М.: Академия, 2008, С. 29-30.
20. *Клышинский Э.С., Кочеткова Н.А.* “Метод извлечения технических терминов с использованием меры странности”, 2014
21. *Кругосвет* – Энциклопедия “Кругосвет”, статья “Терминология”.
22. *Ляшевская О.Н., Шаров С.А.* “Новый частотный словарь русской лексики”, 2009.
23. *Митрофанова О.А. Захаров П.В.* “Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике - Automatic analysis of terminology in the Russian text corpus on corpus linguistics”, 2009.
24. *Соколова Е. Г. , Семенова С. Ю.* “Особенности подготовки терминов для русско-английского тезауруса по компьютерной лингвистике”, 2011.
25. *Readability* -- http://www.readabilityformulas.com/

Приложения

Приложение 1

Список глаголов:
'эксгумировать', 'интенсифицироваться', 'экспоненциально', 'отсеквенировать', 'интериоризовать', 'дезинвестировать', 'дифференцировать', 'матрицировать', 'проецироваться', 'имплицировать', 'лигнифицироваться', 'детерминировать', 'апробироваться', 'редуцироваться', 'утилизироваться', 'легировать'

Приложение 2

Список прилагательных
'рефлективный', 'проницаемый', 'гетерогенный', 'переменный', 'гибридный', 'аморфный', 'протокультурный', 'вещественный', 'инклюзивный', 'ксенофобский'

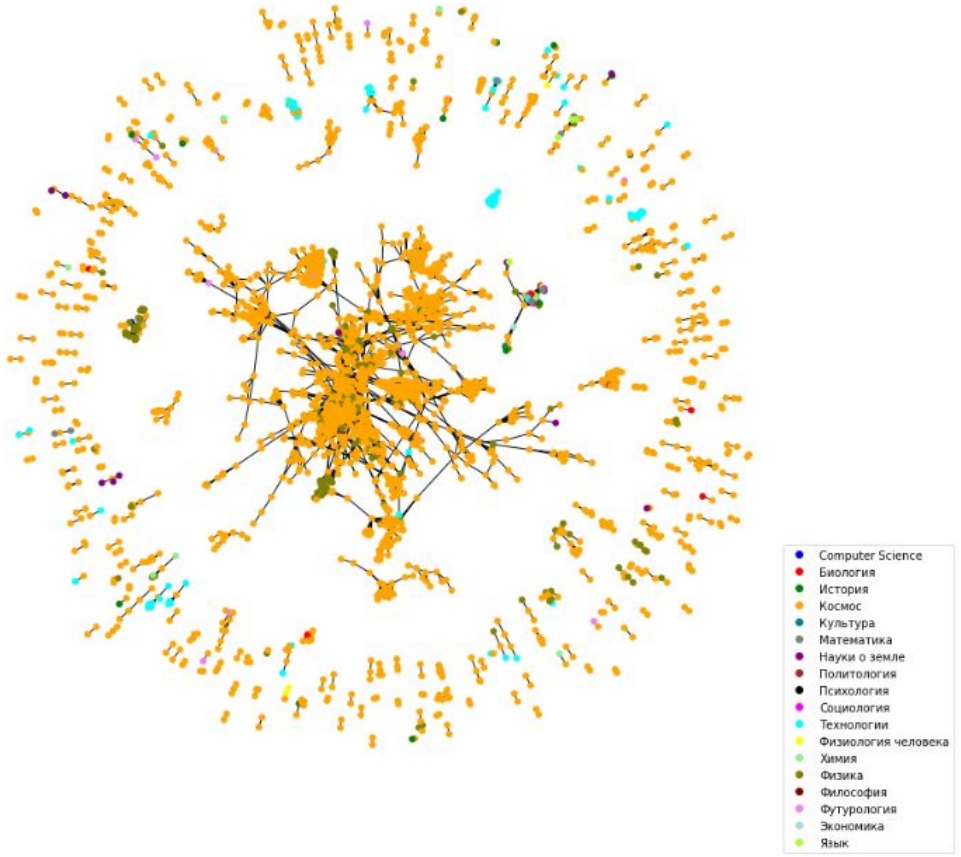
Приложение 3

Контекстные слова
изменение, анализ, количество, концентрация, образец, объем, понятие, амплитуда, действие, использование, использовать, исследование, метод, под названием, называется, называют, например, развитие, разновидность, рост, система, содержание, содержать, стадия, структура, термин, увеличение, уменьшение, уровень, формирование, формировать

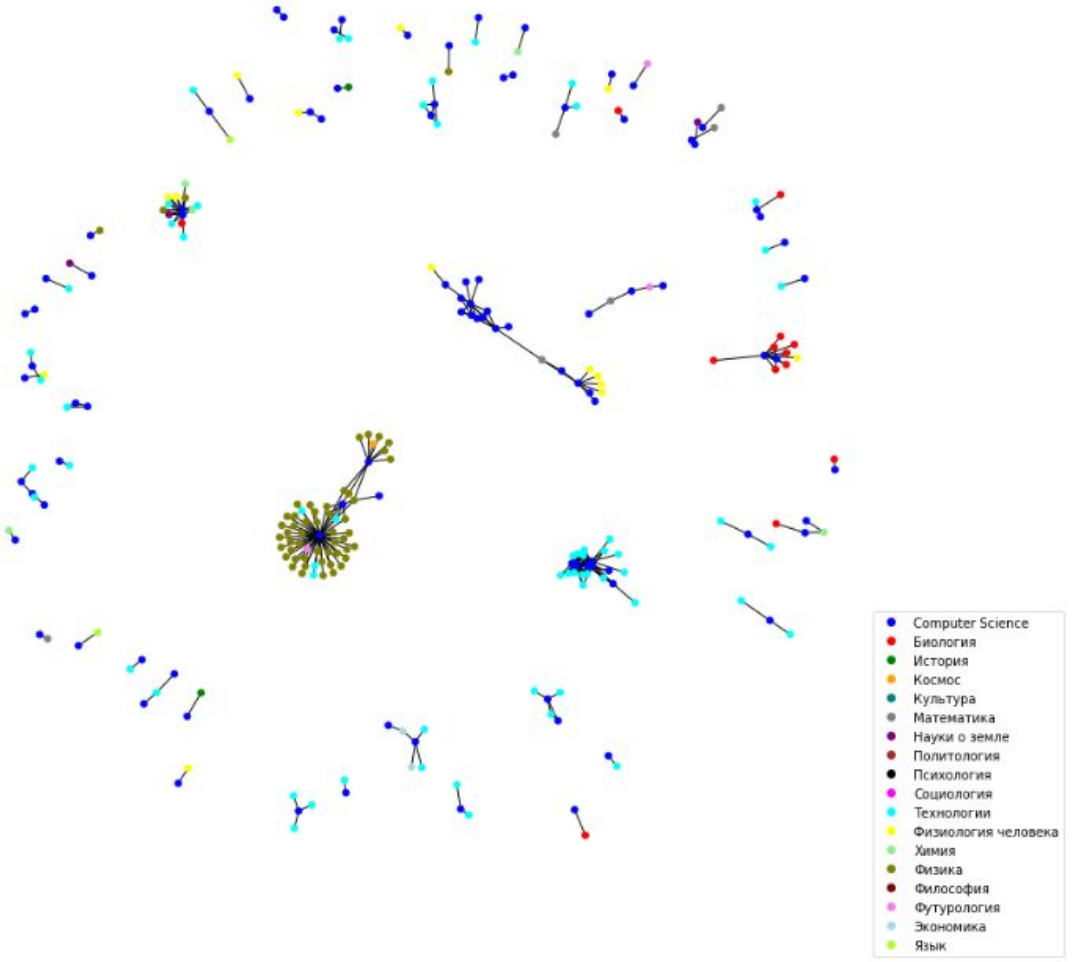
Приложение 4

Графы связей текстов по рубрикам

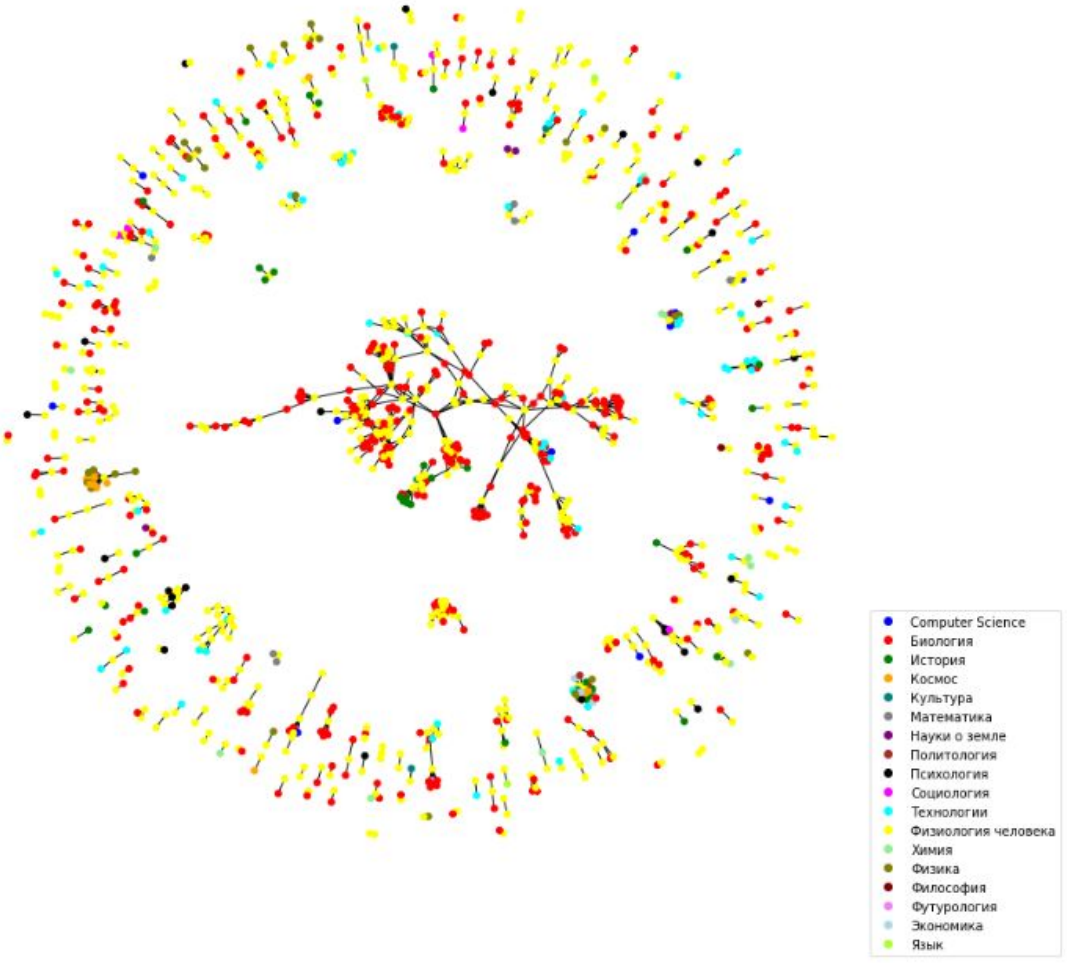
Космос


Computer Science
Биология
История
Космос
Культура
Математика
Науки о земле
Политология
Психология
Социология
Технологии
Физиология человека
Химия
Физика
Философия
Футурология
Экономика
Язык

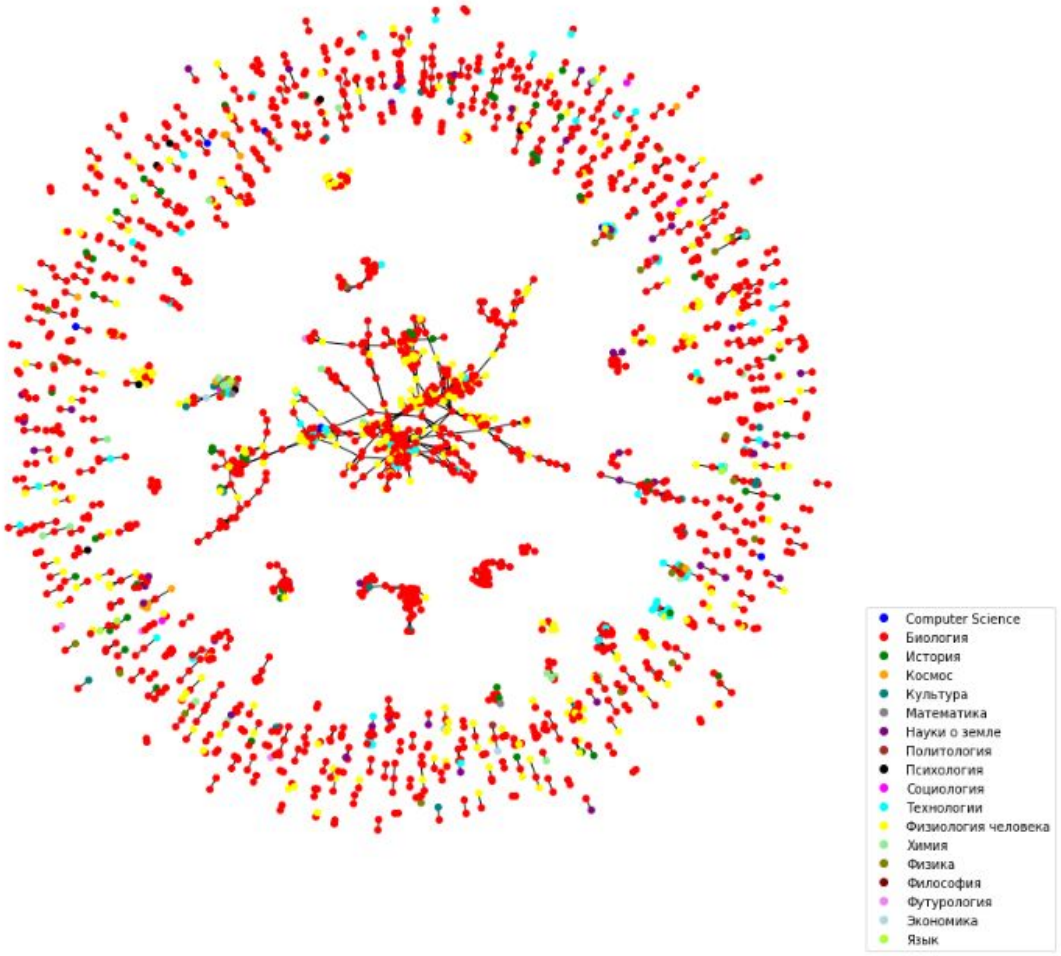
Computer Science



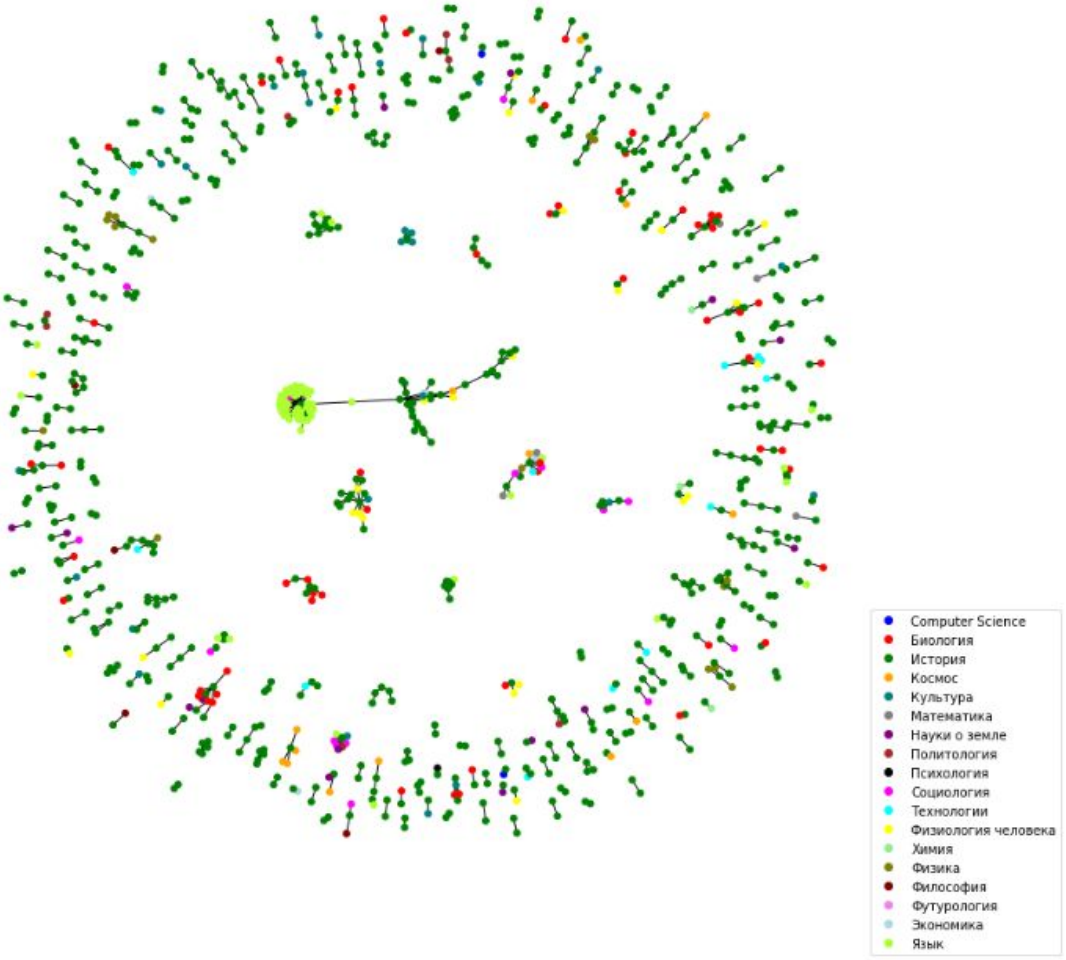
Физиология


Computer Science
Биология
История
Космос
Культура
Математика
Науки о земле
Политология
Психология
Социология
Технологии
Физиология человека
Химия
Физика
Философия
Футурология
Экономика
Язык

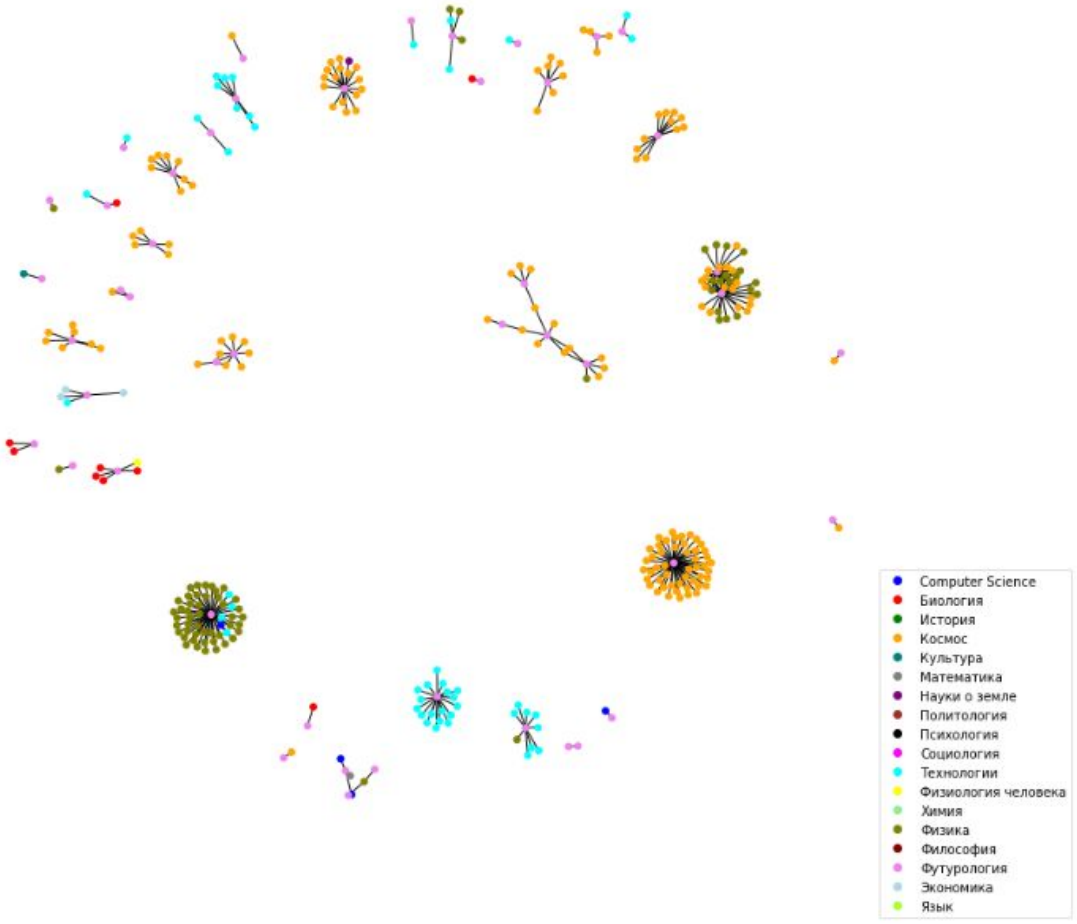
Биология


Computer Science
Биология
История
Космос
Культура
Математика
Науки о земле
Политология
Психология
Социология
Технологии
Физиология человека
Химия
Физика
Философия
Футурология
Экономика
Язык

История


Computer Science
Биология
История
Космос
Культура
Математика
Науки о земле
Политология
Психология
Социология
Технологии
Физиология человека
Химия
Физика
Философия
Футурология
Экономика
Язык

Футурология



Науки о Земле