ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Факультет гуманитарных наук

Есаян Диана Айковна, Герасименко Дарья Александровна. Гольцман Михаил Андреевич, Гончаренко Георгий Иванович, Семашко Ирина Кирилловна

Создание карты культурной (литературной) топонимики Москвы

МАГИСТЕРСКИЙ ПРОЕКТ
по направлению подготовки 46.04.01
образовательная программа «Цифровые методы в гуманитарных науках»

Руководитель
канд. филол. наук, доцент
Орехов Борис Валерьевич

Москва 2022

Содержание

Аннотация

Работа посвящена анализу частотности употребления московских топонимов в русской литературе XIX–XXI веков и сопоставлению ее с социологической оптикой. Описаны методы извлечения именованных сущностей из текстов, предшествующие исследования, проведено сравнение работы существующих NER-библиотек и описаны дальнейшие планы исследования.

Ключевые слова: распознавание именованных сущностей, Natasha, Spacy, социология города, топонимы, русская литература

Abstract

The work is devoted to the statistical analysis of the frequency of the use of Moscow toponyms in the Russian literature of the 19th–21st centuries and its comparison with sociological optics. Methods for extracting named entities from texts and previous studies are described, a comparison of the operation of existing NER libraries is made, and future research plans are described.

Key words: named entities recognition, Natasha, Spacy, urban sociology, toponyms, Russian literature

Введение

В русской литературе XIX–XXI веков закономерно встречается довольно много топонимов, среди них есть и московские. Примеры тому можно найти и в классической литературе, и в современной фантастике, будь то Патриаршие пруды, где в том числе разворачивается сюжет «Мастера и Маргариты», или гостиница Космос из романов Сергея Лукьяненко. Не используя количественные методы, можно заметить, что некоторые топонимы встречаются чаще — например, локации центральных районов города.

За последние несколько десятилетий количество методов цифрового анализа литературы многократно увеличилось. В том числе сформировалось направление в программировании — Natural Language Processing[1] (NLP). Современные NLTK библиотеки[2] позволяют находить в тексте имена собственные, строить деревья зависимостей между словами, определять автора и даже процент заимствований. Это открывает широкий спектр возможностей перед исследователями: при достаточном объеме данных, можно работать, с литературоведческими или другими гуманитарными задачами, получая результаты, которые без количественных методов были бы недоступны.

С одной стороны, актуальность работы подтверждает недостаток исследований в области пересечения литературоведения, социологии и культуромики, с другой — популяризация анализа литературы с помощью цифровых методов. Наше исследование подразумевает анализ как литературного, так и социологического пространства с помощью цифровых инструментов.

Целью нашего проекта является анализ частотности локаций, которые посещают герои русской литературы XIX–XXI веков. Эти семиотические данные мы планируем наложить на существующие социологические представления о популярности тех или иных районов Москвы. Частотность

[1] Обработка естественного языка.

[2] Библиотека NLTK, или NLTK, — пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования Python. Содержит графические представления и примеры данных.

извлеченных топонимов интерпретируется нами в культуромической оптике, то есть как уровень значимости. В социальной проекции это можно было бы назвать степенью видимости тех или иных точек пространства. В этом смысле мы представляем пространство как нечто непрерывное, что культура дробит на отдельные единицы. В результате некоторые элементы этого пространства отражаются в культуре, а некоторые нет.

Важным дополнением к теме является вопрос об увеличении Москвы и включению в область нашего рассмотрения локаций окраин города. В нашей исследовательской оптике — окраина города, или, например, локации, считавшиеся в XIX веке окраиной, считаются московскими. К примеру, Чертаново всегда было в орбите Москвы, задолго до того, как вошло в нее официально. С социальной точки зрения, которую мы взяли как ориентир, Московская область — это тоже Москва. Все мегаполисы «высасывают» человеческий капитал из окружающих регионов, а те служат их придатком. Усадьбы и деревни, не входившие в XIX веке в понятие «город Москва», противопоставлены Москве по линии «город-деревня», но нами это противопоставление не учитывается.

Поставленная цель определила задачи исследования:

- анализ литературы посвященной проблеме понимания и определения социологического пространства города;
- сбор корпуса русской литературы для анализа;
- выбор и сравнение инструментов цифрового анализа;
- написание кода на языке Python, который позволил бы извлечь из текстов корпуса московские топонимы;
- создание метрик для оценки захвата и точности кода;
- анализ полученных результатов.

Теоретический контекст и предшествующие исследования

Основные цели данного проекта включают в себя определение соотношения упоминаемости московских улиц в литературных текстах XIX–XXI веков с социологическим пространством города, а также составление интерактивной карты московских топонимов. В связи с чем возникает вопрос о понятии социологического пространства. Что это такое и какое значение имеет для создания подобной карты? В данной статье мы опираемся на определение, основанное на статье Пьера Бурдье[3]. Социальное пространство — это абстрактное пространство, сконструированное из таких областей как экономика, политика, культура, которые обязаны своей структурой неравному распределению отдельных видов капитала. Другими словами, социальное пространство, существующее наряду с физическим, определяется как пространство социальных процессов, отношений и практик, связанных между собой. Бурдье отмечал, что наличие капитала позволяет присвоить как физическое, так и социальное пространство, и чем оно дефицитней, тем более ценно. Это и является целью социальной борьбы. Отсутствие капитала, наоборот, приковывает к месту.

Таким образом, в городе выделяются районы территориально более ценные, в том числе из-за проживающего там контингента. По О. Е. Трущенко, Москве престижем в первую очередь обладают центральный и западные районы[4]. Это обусловлено также делением данного пространства и проживанием в нем социально господствующих групп: партийных представителей советской власти. Если рассматривать данные от 1989 года именно кварталы к западу и северо-западу наиболее густо населены привилегированными слоями населения, быстрее заполняются общественными, культурными и образовательными учреждениями, что связано с приходом советской власти. При этом рабочий класс (как и до революции) обитает в восточных и юго-восточных районах Москвы.

---

[3] Бурдье П. Социология социального пространства / Пер. с франз. Н.А. Шматко. СПб.: Алетейя, 2007.

[4] Трущенко О. Е. Престиж Центра: Городская социальная сегрегация в Москве. М.: Socio-Logos, 1995.

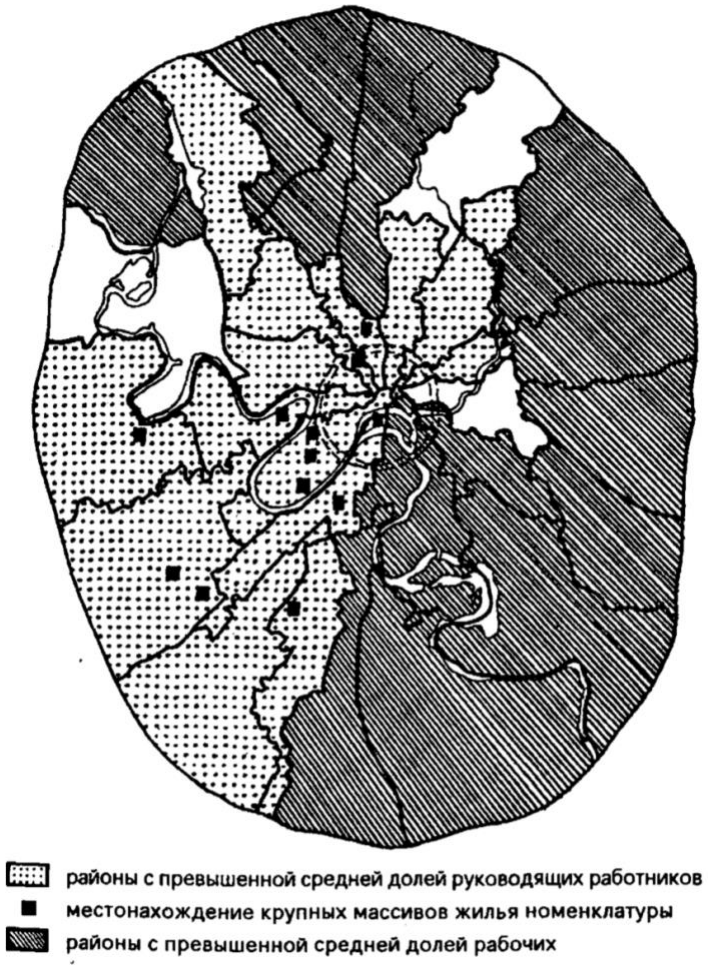

Схема расположения районов с превышенными средними показателями доли руководящих работников и доли рабочих в 1989 году (в пределах МКАД).[5]

Важно также отметить, что в статье позиция современного наблюдателя и современника равны: по Трущенко, все матрицы иерархизации пространства сложились очень давно и сейчас работают инерционно.

Как уже было сказано, основными пунктами для сравнения в нашем исследовании являются семиотическое (культуромическое) и социологические пространства. Культуромика как подход была описана в статье «Quantitative Analysis of Culture Using Millions of Digitized Books» (2011)[6] создателей поискового сервиса «Google Books Ngram Viewer» Жана-Батиста Мишеля, Эреза Либермана Эйдена и др. Описанный ими подход заключается в том, чтобы при

[5] *Рисунок из статьи О. Е. Трущенко*. Там же.
[6] Michel J. B. et al. Quantitative analysis of culture using millions of digitized books //science. T. 331. №. 6014. 2011. С. 176-182.

помощи статистических данных (например, частотности слов), полученных на основе больших корпусов текстов, изучить процессы, протекающие в культуре.

Другой важный текст в контексте нашего исследования — это статья «The Emotions of London»[7] Райна Хойзера, Франко Моретти и Эрика Штайнера. Авторы «Эмоций Лондона» использовали тематическое моделирование для извлечения географической информации из романов XIX века. Вопрос состоял в том, могут ли лондонские топонимы быть важной составляющей анализа эмоциональной географии города. Авторы исследования сначала идентифицировали имена собственные в корпусе с помощью программ распознавания именованных сущностей, затем удалили из списка те термины, которые никаким образом не относятся к Лондону (например, иностранные топонимы и/или имена персонажей). Похожим образом в нашем исследовании мы извлекали именованные сущности для создания черного списка, который включал в себя неподходящие для нашей работы топонимы.

Соответственно, в данной статье, опираясь на теорию Бурдье и культуромический подход, мы сопоставим социальное пространство Москвы с семиотическим, выявленным в ходе анализа литературы XIX–XXI веков, взяв за визуальную основу исследование «The Emotions of London».

Гипотеза

Гипотеза исследования заключается в том, что социологическое представление о престижности районов Москвы, описанное в статье О. Е. Трущенко, напрямую соотносится с культуромическим, то есть полученным при подсчете частотности употребления тех или иных топонимов в русской литературе XIX–XXI веков. Здесь мы исходим из предположения, что распределение слов отражает существующие в культуре представления. К примеру, предположим, что топоним «Патриаршие пруды» очень часто упоминается в русской литературе. Это может быть связано и с тем, что район, в котором находится эта локация, с социологической точки зрения считался

[7] Heuser R., Algee-Hewitt M., Lockhart A. Mapping the emotions of London in fiction, 1700–1900: A crowdsourcing experiment //Literary mapping in the digital age. Routledge, 2016. С. 43-64.

престижным и имел свой ореол значений. Таким образом, частое его появление в литературе обусловлено именно этим фактором.

Методы и инструменты

Для тестирования описанной гипотезы нами были выбраны две библиотеки для задач обработки естественного языка (NLP) — Natasha и Spacy. С их помощью мы привели слова к нормализованной форме и извлекли именованные сущности, в частности локации из произведений русской литературы. Одной из сложностей работы с этими библиотеками было то, что именно они понимают под именованной сущностью, какие категории именованных сущностей выделяют и как. Типы именованных сущностей в библиотеках отличаются, однако некоторое пересечение тем не менее существует и включает в себя, как правило, PERSON, то есть некоторое одушевленное существо с именем, ORGANIZATION, то есть некоторую организацию, и LOCATION, то есть некоторую географическую локацию или область. Как видно из определения LOCATION, эта категория именованных сущностей включает в себя названия стран, городов, рек, озёр и т.д. В рамках исследования было необходимо извлечь из категории LOCATION только те именованные сущности, которые относятся к городским топонимам.

Spacy является одной из наиболее популярных библиотек для обработки естественного языка. Алгоритм ее работы выглядит следующим образом:

- токенизация;
- лемматизация;
- определение частей речи;
- извлечение именованных сущностей.

Токенизация является первым этапом NER-обработки, на котором текст исследуемого корпуса разбивается на отдельные элементы — токены. Далее для корректного анализа и подсчета именованных сущностей необходимо привести полученные токены к нормальной форме — лемматизировать. Таким образом, мы получаем лемматизированный корпус токенов. Это является стандартным

алгоритмом при обработке естественного языка. Дальнейшие этапы зависят от цели конкретного исследования или проекта. В нашем случае — это извлечение именованных сущностей, представленных топонимами.

Natasha работает аналогичным образом. Этот инструмент широко используется в работе с русскоязычными корпусами и демонстрирует достаточно высокую точность. Она также решает основные NLP задачи, такие как токенизация, сегментация, морфологическая разметка, нормализация, разбор синтаксиса и извлечение именованных сущностей. Преимущество библиотеки Natasha заключается в том, что она включает в себя библиотеки, которые заточены на отдельные NLP задачи, что и делает её работу более точной. Например, библиотека Razdel делит русскоязычный текст на токены и предложения. Она построена на правилах русского языка и работает только на текстах, оформленных правильно орфографически и пунктуационно. Следовательно, библиотека с более высокой точностью работает, например, с художественными текстами, а с постами из социальных сетей будет работать с более низким качеством. Библиотека SlovNet используется для моделирования NLP на основе глубокого обучения для русского языка. Slovnet предоставляет высококачественные практические модели для NER русского языка, его морфологии и синтаксиса.

Использование двух библиотек для обработки исследуемого корпуса позволит в дальнейшем сравнить их согласованность и сделать выводы о точности результатов.

Перед тем, как протестировать две вышеописанные библиотеки, нами были вручную размечены несколько глав из произведения Булгакова «Мастер и Маргарита». Скачав текстовый файл с отдельной главой, перед каждой локацией мы добавляли знак решетки (#).

«Однажды весною, в час небывало жаркого заката, в Москве, на #Патриарших прудах, появились два гражданина»[8].

[8] Отрывок из романа «Мастер и Маргарита» Михаила Булгакова.

Пример выделения начальной позиции локации вручную.

Это позволило нам извлекать топонимы из данного текста, ссылаясь на знак решетки, и определять их начальное положение в тексте.

Вручную разметив три главы из романа «Мастер и Маргарита», мы написали код, который извлекает локации и их положение в тексте с помощью библиотек Natasha и Spacy. Обращаясь к положению локации в тексте как к первичному ключу, мы объединили результаты ручной разметки и выводы библиотек Natasha и Spacy и составили CSV-файл с помощью оператора outer join для дальнейшего сравнения точности и согласованности работы библиотек. CSV-файл состоял из четырех столбцов: start (положение в тексте), markup (локация, извлеченная вручную), natasha (локация, извлеченная библиотекой Natasha) и spacy (локация, извлеченная библиотекой Spacy).

	start	markup	natasha	spacy
0	115	Патриаршие пруды	Патриаршие пруды	
1	1096	улица Малая Бронная	Малая Бронная улица	малый бронной улица
2	1868	улица Малая Бронная	Бронная	
3	2185	Патриаршие пруды	Патриарших	
4	22772	Патриаршие пруды	Патриаршие пруды	
5	24581	Патриаршие пруды	Патриаршие пруды	
6	30506	Патриаршие пруды	Патриарших	патриарших
7	41387	улица Малая Бронная		
8	41411	Скатертный переулок	Бронная	
9	41821	Патриаршие пруды	Патриарших	
10	47702	Арбат	Арбат	
11	48282	Арбат	Арбат	арбат
12	48683	Арбат	Арбат	арбат

Структура таблицы с результатами ручной разметки и выводы библиотек Natasha и Spacy

Итак, тестовый прогон трех глав показал, что помимо локаций обе библиотеки также ошибочно извлекают имена персонажей, а также другие слова и словосочетания. К сожалению, Named Entity Recognition — это система далекая от идеала: в данный момент количество ошибочных и

ложноположительных срабатываний все еще высокое. Это связано с тем, что определить критерии отличия организации от географической локации зачастую невозможно. Пока нет понимания контекста, даже человек может неправильно интерпретировать фразу «Мы в Праге»: это может быть как ресторан или город в Чехии, так и торт. То же касается имен персонажей, улиц и людей.

Согласованность результатов двух библиотек Natasha и Spacy мы измерили с помощью коэффициента Каппа Коэна. Данная мера определяется и задается следующей функцией:

$$k = frac p_o - p_e 1 - p_e = 1 - frac 1 - p_o 1 - p_e,$$

где p_o = относительное наблюдаемое согласие,
а p_e = гипотетическая вероятность случайного соглашения

После подсчетов был получен коэффициент равный 0,64. По существующим шкалам интерпретации результатов полученный нами уровень согласия работы моделей можно считать существенным.

Так как в исследовании внимание уделяется только московским топонимам, перед нами стояла задача отсеять московские локации от всех остальных. С этой целью был создан черный список, в который вошли названия всех стран мира, основных гор, морей, океанов, рек и озер, небесных тел, городов России и других стран, регионов России и частей света.[9] Впоследствии данный список пополнялся итеративно после каждого прогона произведения из корпуса для первых прогонов. Дополненный черный список был добавлен в пайплайн кода для вывода локаций с помощью Spacy и Natasha. Следовательно, после извлечения топонимов проводилась фильтрация по черному списку.

Затем мы объединили таблицу с отфильтрованными локациями, извлеченными с помощью библиотеки Natasha, с таблицей с отфильтрованными локациями, которые были извлечены библиотекой Spacy. Соединение производилось с помощью оператора inner join, который

[9] Ссылка на черный список: https://bit.ly/3Oq6yoO

объединяет записи из двух таблиц, если в связующих полях этих таблиц содержатся одинаковые значения. В нашем случае связующим полем служил столбец с позицией начала сущности. Таким образом, у нас получилась таблица с московскими топонимами, извлеченными и библиотекой Natasha, и библиотекой Spacy.

Корпус

Для отработки инструментов мы использовали размеченные вручную романы «Доктор Живаго» Бориса Пастернака и «Мастер и Маргарита» Михаила Булгакова. Для «Доктора Живаго» были размечены все главы, для «Мастера и Маргариты» — три[10] выбранные. В случае с романом Булгакова была также найдена позиция каждого топонима в тексте.

Помимо перечисленных размеченных романов, мы также создали первичный корпус для итеративного создания и отладки инструментов поиска московских топонимов. Он состоит их тех произведений русской литературы, в которых точно содержатся московские топонимы. Среди них, к примеру, «Москва и москвичи» Михаила Загоскина, выбранные главы «Войны и мира» Льва Толстого, «Счастливая Москва» Андрея Платонова и другие. Этот корпус был необходим для первых прогонов выбранных нами NER-библиотек и составления черного списка, о котором подробнее рассказано в пункте «Методы и инструменты».

В качестве корпуса для создания финального набора данных и статистики мы планируем использовать электронную библиотеку librus.ec. Финальный прогон не предусматривает отбора, как в случае с первичным корпусом — на этом этапе и с помощью этого корпуса обрабатываются все доступные тексты русской литературы XIX–XXI веков.

Промежуточные выводы

Для анализа первичных результатов было проведено сравнение трех вариантов разметки текста, с использованием библиотеки Natasha, Spacy и ручной разметки. Каждый вариант разметки был вынесен в CSV-файл, в

[10] Были размечены 1, 5 и 22 главы.

котором есть колонка для именованной сущности и ее начального символа. Далее все три таблички были соединены по принципу outer join, то есть именованные сущности из всех таблиц были вынесены в одну. Если сущность извлекалась и с помощью библиотеки Natasha, и с помощью библиотеки Spacy, и с помощью ручной разметки, она появлялась во всех трех колонках. В противном случае сущность фигурировала только в одной или двух.

Сложность интерпретации результатов дополняется тем, что ручная разметка включала в себя только названия московских топонимов, в то время как разметка с помощью Python позволяла выделять все сущности типа LOC, а субкатегории этого типа выбрать не позволяла, что привело к тому, что с помощью кода были найдены такие локации как Москва или Кисловодск. Согласованность работы библиотек была признана удовлетворительной, основываясь на метрике Каппа Коэна.

Первичные результаты также указали на то, как избежать ошибок и трудностей при работе с большим объемом данных в будущем. Невозможность выбирать подкатегории из выделяемых с помощью Python именованных сущностей определила необходимость их автоматической категоризации. Создание черного списка с городами, реками и немосковскими топонимами позволило отфильтровать то, что используемые библиотеки понимают под локацией, в формат необходимый нам в рамках данного исследования. Дальнейшая фильтрация топонимов будет осуществляться с помощью инструментов визуализации: на карте будут отображаться топонимы только в рамках интересующего нас города, то есть Москвы.

Планы

Нашей следующей задачей является обработка корпуса всей русской литературы XIX–XXI веков через отработанные инструменты — библиотеки Natasha и Spacy. Далее мы планируем получить финальный CSV-файл, проинтерпретировав который, мы сможем либо опровергнуть, либо подтвердить описанную в работе гипотезу.

Помимо этого, мы планируем создать визуализацию полученных результатов — тепловую карту. Для этого был выбран инструмент API Яндекс

Карт — это набор сервисов, которые позволяют использовать картографические данные и технологии Яндекса в личных проектах. Чтобы создать подобного рода карту мы планируем выполнить следующие шаги:

1. с помощью Геокодера API, который используется для перевода географических координат в адрес и наоборот, получить координаты для всех извлеченных локаций;

2. с помощью JavaScript API создать свою карту.

Визуализация данных позволит упростить процесс интерпретации результатов, поскольку среди социологических работ о престижности районов Москвы существует много аналогичных представлений информации. Следовательно, нашу карту можно будет сравнить с существующими социологическими.

Источники

1. Michel J. B. et al. Quantitative analysis of culture using millions of digitized books //science. Т. 331. №. 6014. 2011. С. 176-182.
2. Heuser R., Algee-Hewitt M., LOCKHART A. Mapping the emotions of London in fiction, 1700–1900: A crowdsourcing experiment //Literary mapping in the digital age. Routledge, 2016. С. 43-64.
3. Бурдье П. Социология социального пространства / Пер. с франз. Н.А. Шматко. СПб.: Алетейя, 2007.
4. Трущенко О. Е. Престиж Центра: Городская социальная сегрегация в Москве. М.: Socio-Logos, 1995.

**Хранилище данных проекта: https://github.com/diana-esaian/MOSCOW-TEXT*