Федеральное государственное автономное образовательное учреждение высшего образования

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет гуманитарных наук

КУРСОВАЯ РАБОТА

Оцифровка и анализ русскоязычных телеграмм

по направлению подготовки Фундаментальная и прикладная лингвистика

образовательная программа «Цифровые методы в гуманитарных науках»

Выполнили:

Студенты группы МЦМГН211

Настас Марина

Маслова Мария

Лошкарев Антон

Руководитель:

к.ф.н., доц. Скоринкин Д. А.

Москва 2022

ОЦИФРОВКА И АНАЛИЗ РУССКОЯЗЫЧНЫХ ТЕЛЕГРАММ

Авторы: Марина Настас, Мария Маслова, Антон Лошкарев

Аннотация

Данная работа посвящена оцифровке и анализу русскоязычных телеграмм. Был собран датасет телеграмм из портала открытых данных Министерства культуры РФ, личных архивов, музеев и др. С помощью методов компьютерной лингвистики и анализа данных было изучено тематическое распределение телеграмм по историческим эпохам и динамику варьирования их разнообразия по мере развития альтернативных средств коммуникации, корреляцию между историческими событиями и частотой отправки телеграмм, влияние политики на повседневную коммуникацию посредством телеграмм и интенсивность отправки телеграмм в зависимости от темы.

Ключевые слова: телеграмма, история коммуникации, цифровая гуманитаристика, сетевой анализ, Python

Введение

История мгновенной коммуникации, когда сообщение можно доставить от адресата получателю за очень короткий срок, берёт своё начало не с изобретения интернета, как можно было бы подумать, а с изобретения телеграфа. Телеграммы стали быстрым способом обмениваться сообщениями, и несмотря на то, что почтовая связь уже существовала, телеграф взял на себя несколько иную задачу. Доставка почты занимала слишком долгое время и не подходила набирающему скорость темпу жизни XIX-XX веков. Именно поэтому на смену письмам пришли телеграммы, которые хоть передавали значительно меньше информации, но они могли это делать здесь и сейчас, сообщая сведения, не терпящие задержки.

Цель данной работы — выяснить, каковы характерные особенности телеграмм, исследовать корреляцию исторических событий и тематической составляющей телеграмм, а также географию отправления и получения. Для выполнения этой цели необходимо создать датасет с телеграммами, достаточно большой и репрезентативный, чтобы его можно было изучать. В рамках этой задачи были использованы как личные архивы, так и портал открытых данных Министерства культуры Российской Федерации [1]. Большинство рассматриваемых телеграмм были взяты из последнего источника, что влечёт за собой некоторые особенности выборки: многие из этих телеграмм были адресованы публичным личностям и деятелям

культуры, и, следовательно, мы не вполне можем проецировать полученные результаты на телеграфную коммуникацию в целом.

В ходе работы было поставлено несколько гипотез, часть из которых удалось доказать, однако некоторые так и не были подтверждены. Поставленные гипотезы выглядели следующим образом:

Гипотеза 1: Предполагается, что с развитием других методов коммуникации тематика телеграмм сужалась, причём фокусировалась она на конкретной области, а именно области «важных новостей из жизни»: поздравлений, уведомлений о смерти или болезни. Вследствие этого возникает вопрос: правда ли, что телеграммы, в которых содержались «плохие» новости, выбрасывали?

Гипотеза 2: Когда происходило какое-то историческое событие, которое связано с ситуацией в стране в целом, телеграммы отправлялись чаще, чем обычно. Это нужно рассмотреть на примере конкретного населённого пункта, Москвы, так как при работе с настолько крупным городом будет легче найти достаточно материала для репрезентативной выборки.

Гипотеза 3: Телеграммы и имена собственные: государственные vs. обычные телеграммы. Хочется проверить, насколько часто имена собственные деятелей фигурируют в телеграммах и в каком контексте (поздравления, новости);как большая политика отражается в повседневной коммуникации между людьми. Сама гипотеза состоит в том, что чаще всего это связано со смертью. Гипотеза: несмотря на то, что телеграммы — зачастую личная переписка, в них упоминаются и имена общественных и политических деятелей. Нередко это связано со смертью деятеля.

Гипотеза 4: Интенсивность поздравлений с помощью телеграмм менялась в зависимости от праздников (например, в начале марта будет больше телеграмм с поздравлениями с 8 марта, в конце декабря — с новым годом и др.). Также с помощью телеграмм можно проследить становление новых праздников в жизни общества — в частности, с 1965 по 1985 год начинает повышаться упоминаемость Дня победы.

Помимо четырёх основных гипотез были поставлены и две второстепенные. Как было описано в плане работы, «Одна из них — это то, что если датасет будет достаточно большим, можно будет визуализировать все переписки, имеющиеся у нас, в виде графа, где вершины — это отправители и адресаты, а рёбра — это телеграммы, и этот граф будет полностью связным. Как следствие, можно предположить, что существует некая крупная историческая фигура, с которой абсолютно все вершины графа будут связаны через несколько

«рукопожатий» — в данном случае, телеграмм». Проверить это возможно только в том случае, если в датасете будет обнаружено достаточное количество упоминаний известных личностей: как в качестве адресата или отправителя, так и в тексте самих телеграмм.

Результатом работы стал полностью размеченный датасет и комплексный анализ телеграмм (тематический и исторический) с использованием как ручных, так и программных методов.

Исследований, посвящённых непосредственно телеграммам, на данный момент немного. Они преимущественно опираются на качественные методы анализа, предполагающие анализ и интерпретацию теоретической части вопроса, и только в последнее десятилетие появились работы, совмещающие качественные и количественные исследования телеграмм, в частности, с использованием цифровых методов. В этом исследовании мы продолжаем новую тенденцию анализа телеграмм программными методами при помощи инструментов, которые представляют такие языки, как Python и R.

Обзор существующих подходов

Одной из фундаментальных работ о телеграфной коммуникации стала книга журналиста Тома Стэндейджа «The Victorian Internet» [2], написанная в 1999 году. Как очевидно из названия, автор сравнивает телеграфное сообщение в конце XVIII – начале XIX веков и современный интернет. Он подчёркивает мысль о том, что, хоть оборудование этих двух средств коммуникации и отличается, их влияние на повседневную жизнь пользователей поразительно похоже.

Более поздняя работа Роланда Венцльхюмера [3] ставит своей задачей показать, что изучение телеграмм требует более комплексного подхода, чем проведение параллелей. Исследователь в своей работе спорит со Стэндейджем, доказывая, что телеграфную связь и интернет очень сложно отождествлять, и к таким диахроническим сравнениям нужно относиться осторожно. Венцльхюмер в своей книге вводит тенденцию изучения телеграфной коммуникации и её влияния на глобализацию не только качественными, но и количественными методами. Так, для этой цели автор использует сетевой анализ и цифровую историческую картографию.

Также существует проект «Пишу тебе» [4], который занимается задачей, схожей с тем, чем занимаемся мы, но отличается предмет исследования — открытки вместо телеграмм. Именно поэтому нашим ориентиром стал данный проект. Многое из того, что было сделано в этой работе, было сделано по образу и подобию «Пишу тебе». Несмотря на это, все сложности поиска необходимой информации легли на нас, поскольку пришлось с нуля собирать датасет телеграмм для проверки гипотез.

На сегодняшний день телеграф используется в исключительных ситуациях. Даже в однотысячном датасете современных телеграмм оказалось мало. В приоритете сейчас почтовые сообщения и СМС-уведомления. Таким образом, история телеграммы хоть и не закончилась вовсе, но может считаться именно историей, потому что случаи её нынешнего использования редки. Однако, несмотря на это, даже в настоящее время встречаются случаи, когда один политический деятель поздравляет с чем-то другого с помощью телеграмм.

Точно можно сказать только то, что бытовое использование телеграммы уже себя изжило, потому что на место телеграфным сетям пришла глобальная, которая и дешевле, и быстрее, и позволяет передать большее количество информации.

Данные

Получение и подготовка данных

В качестве малого датасета собрано 1043 телеграммы из различных источников: семейный архив, музеи, форумы коллекционеров, рынок и коллекция музейных экспонатов портала открытых данных Министерства культуры Российской Федерации.

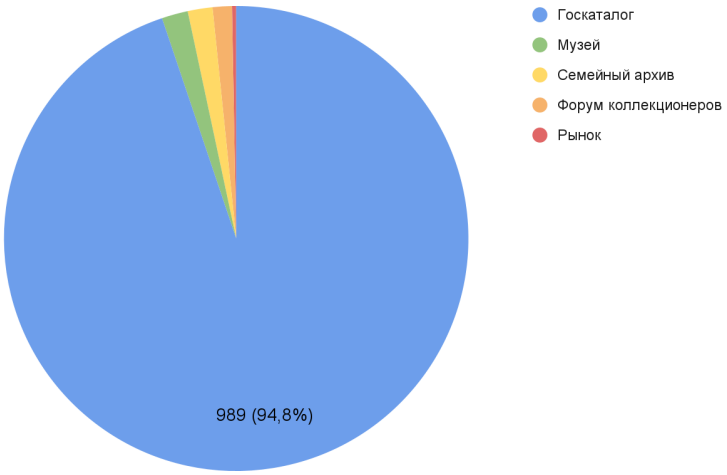

Рис. 1. Распределение телеграмм в нашем корпусе по источникам.

Такой сильный перевес в распределении источников не мог не отразиться и на данных. Главной проблемой стало то, что в госкаталог материалы попадают из музеев, а в музеях в основном представлены телеграммы, принадлежащие известным историческим личностям. Как следствие, большинство имеющихся в датасете телеграмм были написаны или получены известными людьми. Это негативно сказывается на возможности адекватно оценить данные, поскольку в датасете мало телеграмм, принадлежащих «простым» людям.

Процесс разметки телеграмм, собранных из других источников, кроме портала открытых данных, происходил полностью вручную, в то время как обработка данных из портала осуществлялась преимущественно программными методами ввиду большого объёма информации, обработать которую полностью вручную было бы проблематично. В связи с этим данная глава будет посвящена обработке той части датасета, которая была получена из платформы Минкульта РФ.

Для получения данных о телеграммах из портала открытых данных был использован API-ключ[1], который платформа предоставляет для разработчиков, чтобы иметь возможность выгрузить необходимые данные выборочно. В результате такого точечного запроса был получен JSON-файл[2] со следующими данными о телеграммах:

```
some_list = []
for i in data["data"]:
    inner_data = i.get("data")
    new_data = {
        "ID Gov": inner_data.get("id"), # ID экспоната в системе портала
        "Ссылка": [_.get("url") for _ in inner_data.get("images", [])],
        "Название": inner_data.get("name"),
        "Место отправления": inner_data.get("productionPlace"),
        "Место получения": inner_data.get("findPlace"),
        "Описание": inner_data.get("description"),
        "Дата": inner_data.get("periodStr"),
        "Вид бланка": inner_data.get("technologies", []),
    }
    some_list.append(new_data)
```

Рис. 2. Фрагмент кода, который получает объекты из JSON-файла, необходимые для формирования датасета.

Полученные данные подверглись как автоматической, так и ручной обработке. Первым этапом обработки данных было получение изображений телеграмм по ссылкам из базы данных портала с помощью модуля requests[3], чтобы извлечь тексты телеграмм и другие метаданные. Изначально получено ровно 1000 телеграмм из портала Минкульта РФ, однако в процессе ручной проверки изображений выяснилось, что некоторые из них не подходят. Запрос в базе данных портала осуществлялся по поиску в названии экспоната ключевого слова «Телеграмма», и хотя большую часть поисковой выдачи по такому запросу действительно составляли телеграммы, иногда попадались экспонаты, представляющие

[1] API-ключ — индивидуальный код доступа, с помощью которого разработчик может работать с материалами сайта, которые недоступны при обычном использовании ресурсом.

[2] JSON (JavaScript Object Notation) — текстовый формат обмена данными, основанный на JavaScript.

[3] Requests — библиотека для языка Python, которая используется для обработки HTTP-запросов.

собой обложки книг о телеграммах, фотографии и негативы вручения телеграмм и т. д. Подобные результаты выдачи пришлось удалить из датасета.

Далее понадобилось присвоить каждой телеграмме уникальный идентификатор. ID в датасете также формировался особым образом. Номер первого объекта — tg00001-1, где цифры до дефиса обозначают порядковый номер телеграммы в базе данных, а после дефиса — номер изображения в случае, если их приходится несколько на одну телеграмму (например, двусторонняя телеграмма, открытка и др.). Уникальные идентификаторы для телеграмм из портала открытых данных присвоены программными методами.

Затем было необходимо заполнить остальные столбцы базы данных с помощью тех, что были даны в изначальном JSON-файле из портала открытых данных. Так, благодаря колонке «Название» автоматически проставлены отправители и получатели телеграмм с помощью инструментов NER[4].

Содержимое столбца «Название» выглядит приблизительно следующим образом: «Телеграмма Покровскому Н.А. от Простаковых с поздравлением с высокими правительственными наградами». Следовательно, из этого можно определить несколько параметров: именованная сущность в дательном падеже с тегом PER (персона) как получатель телеграммы (Покровскому Н.А.) и именованная сущность в родительном падеже с тегом PER как отправитель телеграммы (Простаковых). Поиск осуществлялся не только по персонам, но и по организациям и локациям, поскольку часто отправителями и получателями телеграмм были не только частные лица, но также коллективы и организации.

Также с помощью колонок «Название» и «Описание» были автоматически проставлены темы и виды бланков некоторых телеграмм. Так, например, если в названии встречается словосочетание «присвоение звания», значит, тема тоже звучит соответствующим образом, что в дальнейшем подтвердилось при ручной проверке автоматически проставленных данных.

Чтобы распознать тексты телеграмм на изображениях, был использован OCR-редактор ABBYY FineReader[5]. Однако такой способ распознавания срабатывал далеко не всегда: например, если ленты телетайпа наклеены на телеграмму криво, или если телеграмма рукописная, программа не справлялась. Большую часть изображений пришлось расшифровывать вручную.

---

[4] NER (Named Entity Recognition) — распознавание именованных сущностей (имена людей, названия организаций, топонимы и др.).

[5] OCR (Optical Character Recognition) — перевод изображений из печатного или рукописного текста в текстовые данные, использующиеся для представления символов в компьютере.

Данные на выходе и их формат

В результате разметки получилась следующая иерархия столбцов в базе данных:

- *Автоматически полученные данные*: данные, полученные из портала открытых данных Минкульта РФ, которые не подвергались дополнительной разметке.

Из этих колонок собиралась информация для разметки других параметров. К этому типу столбцов отнесены ID в рамках базы данных, ID предмета в системе портала, ссылку на изображение, название и описание.

- *Сырые данные*: данные, для которых есть соответствующее поле в тексте телеграммы.

К этой категории относятся следующие столбцы: текст, получатель, место отправления, место получения, дата, время приёма, время передачи и служебный заголовок. Место отправления, получения и дата не принадлежат к предыдущей категории автоматически добытых данных, поскольку они носят не описательный характер, а относятся к содержанию телеграммы, хоть и большая часть данных в рамках этих столбцов действительно была добыта автоматически.

- *Искусственно полученные данные*: данные, содержащие метаинформацию о телеграммах, для которой нет отдельных полей в теле телеграмм.

К ним относятся такие столбцы, как язык, графика, тип телеграммы, вид бланка, тема, тон, персоналии, источник и примечания. Следует отметить, что столбец «Отправитель» относится именно к этой категории, а не к предыдущей, поскольку в теле телеграммы нет отдельного поля для подписи отправителя. Также важно упомянуть, что тональность телеграмм определялась вручную, а не программными методами, поэтому данный параметр относится именно к искусственно полученным данным.

Помимо этого, была разработана специальная система разметки, чтобы фиксировать различные явления в телеграммах. С помощью этой системы отмечена дореволюционная орфография, орфографические ошибки, опечатки, аббревиатуры, сокращения и несоответствие языка и графики телеграммы, которые в дальнейшем использовались для анализа телеграмм.

Поскольку проект предполагает дальнейший сбор телеграмм, внедрение этого датасета в базу проекта «Пишу тебе» и, соответственно, краудсорсинговый формат, где любой желающий может поделиться своей телеграммой или помочь с расшифровкой, была написана подробная инструкция по разметке телеграмм, которая включает подробное описание системы разметки

и всех параметров телеграмм, которые были зафиксированы. Ознакомиться с инструкцией можно в нашем репозитории (см. приложение 1).

Внутр. ID	Внеш. ID	Ссылка	Текст	Название	Отправитель	Получатель	Место отправления	Место получения	Описание
tg00055	10001801	http://goskatak	СОЮЗМИНЛЕГПРОМ СЕРДЕЧНО ПОЗДРА	Телеграмма правительс	Тарасов, организация	Журавлева Тамара М	РСФСР, г. Москва	РСФСР, Алтайский	На бумаге машинопи
tg00056	10001803	http://goskatak	УВАЖАЕМАЯ ТАМАРА МАРКОВНА ГОРЯ	Телеграмма Журавлево	Парамонов А.М., Ром	Журавлева Тамара М	РСФСР, г. Москва	РСФСР, Алтайский	На бумаге набран ма
tg00057	10001804	http://goskatak	КРАЙИСПОЛКОМ ГОРЯЧО И СЕРДЕЧНО	Телеграмма правительс	Раевский В.Н.	Журавлева Тамара М	РСФСР, Алтайский кра	РСФСР, Алтайский	Стандартный бланк
tg00058	10002334	http://goskatak	ВСЕГО СЕРДЦА ПОЗДРАВЛЯЕМ НАГРАЖ	Телеграмма от Кирюшки	Кирюшкин	Островский Николай /	РСФСР, г. Москва	РСФСР, Азово-Чер	NA
tg00059	10002383	http://goskatak	ГОРЯЧО И СЕРДЕЧНО ПОЗДРАВЛЯЕМ В	Телеграмма Журвлевой	Мельникова А.И.	Журавлева Тамара М	РСФСР, Алтайский кра	РСФСР, Алтайский	На типографском бл
tg00060	10002388	http://goskatak	ДОРОГАЯ ТАМАРА МАРКОВНА, ЗА ВЫДА	Телеграмма Журавлево	Сальников, Бологов,	Журавлева Тамара М	РСФСР, Алтайский кра	РСФСР, Алтайский	На бумаге набран ма
tg00061	10002390	http://goskatak	УВАЖАЕМАЯ ТАМАРА МАРКОВНА КОЛЛЕ	Телеграмма. Журавлевс	Белышев, Дикова, ор	Журавлева Тамара М	РСФСР, г. Москва	РСФСР, Алтайский	На бумаге машинопи
tg00062	10002392	http://goskatak	ДОРОГАЯ ТАМАРА МАРКОВНА БАРНАУЛ	Телеграмма правительс	Сидоров М.М.	Журавлева Тамара М	РСФСР, Алтайский кра	РСФСР, Алтайский	На стандартном бла
tg00063	10002821	http://goskatak	ПОЗДНАВЛЯЮ <ПОЗДРАВЛЯЮ> ПОЛУЧЕ	Телеграмма от М. Кольц	Кольцов Михаил Ефи	Островский Николай /	РСФСР, г. Москва	РСФСР, Азово-Чер	NA
tg00064	10005659	http://goskatak	ВОЛХОВЧАНЕ ГОРЯЧО ОДОБРЯЮТ И ПО	телеграмма, правительс	Свердлов А.Г.	Кузьмин А.П., Воробь	РСФСР, Ленинградская	РСФСР, г. Москва	Телеграмма в Моск
tg00065	1000758	NA	NA	Юделевич Лев. Телегра	Юделевич Лев	Покровский Николай /	РСФСР, г. Москва	NA	NA
tg00066	1000759	NA	NA	Алибек Заирбек. Телегр	Заирбек Алибек	Покровский Николай /	УССР, г. Харьков	NA	NA
tg00067	1000761	NA	NA	Дорохины. Телеграмма Г	коллективный (семья	Покровский Николай /	РСФСР, г. Москва	РСФСР, г. Сталингр	NA
tg00068	1000762	NA	NA	Ардаров Г.П. Телеграмм	Ардаров Г.П.	Покровский Николай /	РСФСР, Татарская АС	NA	NA
tg00069	1000763	NA	NA	Сарычева Е. Телеграмм	Сарычева Е.	Покровский Николай /	РСФСР, г. Москва	NA	NA
tg00070	1000765	NA	NA	Белокопытов А. Телегра	Белокопытов А.	Покровский Николай /	РСФСР, г. Москва	NA	NA
tg00071	1000767	NA	NA	Чижова /А.С./, Медовщи	Чижова А.С., Медовщ	Покровский Николай /	РСФСР, г. Ярославль	NA	NA
tg00072	1000769	NA	NA	Чкалова. Телеграмма Пс	Чкалова	Покровский Николай /	РСФСР, Нижегородска	NA	NA
tg00073	1000770	NA	NA	Светлова. Телеграмма Г	Светлова	Покровский Николай /	РСФСР, Московская об	NA	NA
tg00074	1000771	NA	NA	Гилоди Е., Ковалевская	Гилоди Е., Ковалевск	Покровский Николай /	NA	NA	NA
tg00075	1000773	NA	NA	Алексеев. Телеграмма Г	Алексеев	Покровский Николай /	РСФСР, Воронежская с	NA	NA
tg00076	1000775	NA	NA	Марра, Людмила, Всево	NA (Марра), NA (Люд	Покровский Николай /	РСФСР, Воронежская с	NA	NA
tg00077	1000779	NA	NA	Покровский (член Правл	Покровский	Покровский Николай /	РСФСР, г. Москва	РСФСР, г. Сталингр	NA
tg00078	1000780	NA	NA	Толстая Л.И. Телеграмм	Толстая Л.И.	Покровский Николай /	РСФСР, г. Москва	РСФСР, г. Горький	NA
tg00079	1000784	NA	NA	Струков. Телеграмма Пс	Струков	Покровский Николай /	РСФСР, г. Горький	NA	NA
tg00080	1000785	NA	NA	Кац. Телеграмма Покров	Кац	Покровский Николай /	РСФСР, г. Москва	NA	NA
tg00081	1000786	NA	NA	Дорохины. Телеграмма Г	коллективный (семья	Покровский Николай /	РСФСР, г. Москва	NA	NA
tg00082	1000787	NA	NA	Чуколенко, Неклюдовы.	коллективный (семья	Покровский Николай /	РСФСР, Владимирская	NA	NA

Дата	Вр. приёма	Вр. перед.	Служебное	Язык	Графика	Вид бланка	Вид телеграммы	Тема	Тон	Персоналии	Источник	Примечание
xx.xx.1981	NA	NA	233192 ХЛОПОК 111	русский	кириллица	печатный	правительственная	присвоение звания	позитивный	Журавлева Тамара	госкаталог	
xx.xx.1981	NA	NA	МОСКВЫ МТП 4 445	русский	кириллица	печатный	правительственная	присвоение звания	позитивный	Журавлева Тамара	госкаталог	
xx.xx.1981	11:20	11:22	БАРНАУЛ 35/1001 3	русский	кириллица	печатный	правительственная	присвоение звания	позитивный	Журавлева Тамара	госкаталог	
xx.xx.1935	13:20	NA	СОЧИ ОРЕХОВАЯ 4	русский	кириллица	печатный	обычная	получение награды	позитивный	Островский Никол	госкаталог	
xx.xx.1981	13:45	19:20	БАРНАУЛ 99/280 54	русский	кириллица	печатный	обычная	присвоение звания	позитивный	Журавлева Тамара	госкаталог	
xx.xx.1981	NA	NA	ИЗ БАРНАУЛА 19/3-	русский	кириллица	печатный	обычная	присвоение звания	позитивный	Журавлева Тамара	госкаталог	
xx.xx.1980	NA	NA	7(233192 ХЛОПОК.	русский	кириллица	печатный	обычная	присвоение звания	позитивный	Журавлева Тамара	госкаталог	
xx.xx.1981	8:10	8:40	ПРАВИТЕЛЬСТВЕН	русский	кириллица	печатный	правительственная	присвоение звания	позитивный	Журавлева Тамара	госкаталог	
xx.xx.xxxx	16:30	16:45	СОЧИ <СОЧИ> ПИС/	русский	кириллица	печатный	обычная	одобрение	позитивный	Островский Никол	госкаталог	
xx.xx.xxxx	18:30	18:50	СЪЕЗД ГОРОД МО(	русский	кириллица	печатный	бланк съезда КПСС	присвоение звания	позитивный	Островский Никол	госкаталог	
24.07.1944	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
xx.xx.1949	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
31.05.1958	NA	NA	NA	русский	NA	печатный	NA	день рождения	позитивный	Покровский Никол	госкаталог	
25.01.1949	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
25.07.1944	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
xx.xx.1949	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
xx.xx.1944	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
xx.xx.1944	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
xx.xx.1944	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
24.07.1944	NA	NA	NA	русский	NA	печатный	открытка	присвоение звания	позитивный	Покровский Никол	госкаталог	
xx.xx.1944	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
xx.xx.1944	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
22.03.1958	NA	NA	NA	русский	NA	печатный	NA	приглашение	позитивный	Покровский Никол	госкаталог	
xx.xx.194x	NA	NA	NA	русский	NA	печатный	NA	поздравление	позитивный	Покровский Никол	госкаталог	
26.07.1944	NA	NA	NA	русский	NA	рукописный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
xx.xx.1944	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
xx.xx.1944	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	
24.07.1944	NA	NA	NA	русский	NA	печатный	NA	присвоение звания	позитивный	Покровский Никол	госкаталог	

Рис. 3-4. Фрагмент малого датасета. Зелёным выделены автоматически полученные данные, красным — сырые, синим — искусственно полученные данные.

Важно отметить, что место отправления и получения телеграмм фиксировались в соответствии с исторической принадлежностью территории в тот временной период, когда была отправлена телеграмма. Так, если телеграмму отправили из Ленинграда в 1967 году, место отправления было указано как «РСФСР, г. Ленинград», а не «Россия, г. Санкт-Петербург».

Инструментарий

Подготовка датасета

По итогам сбора данных, имелись два датасета в виде таблиц: первый, который далее будет назван «большим» датасетом, состоящий из приблизительно девяноста тысяч телеграмм, которые были получены описанными выше методами и в котором отсутствует текстовая и тематическая разметка, и второй, «малый» датасет, состоящий из 1043 телеграммы, которые были полностью размечены вручную. Несмотря на высокое качество разметки в источнике, данные нельзя было назвать полностью готовыми для машинного анализа, и, как следствие, потребовалось провести предварительную обработку.

Первая задача предобработки данных возникла из-за неоднородной разметки в поле «Дата отправки». Данные в поле могли выглядеть абсолютно по-разному: где-то был указан только год, где-то – полная дата вплоть до дня и месяца, где-то и вовсе только столетие (например, нередко попадалось значение «Вторая половина 19 века»). Основной задачей было привести все даты к единому формату записи для облегчения последующего машинного анализа. Единым форматом был выбран формат записи вида «xx.xx.xxxx», где день, месяц и год были записаны именно в таком порядке двумя, двумя и четырьмя цифрами соответственно. Однозначные числа месяца и дня в обязательном порядке дополнялись нулём слева, а недостающую информацию заменяли латинской буквой «икс» («x»). Данный формат уже успешно применяется в проекте «Пишу тебе», откуда он и позаимствован.

Затем, поскольку из 90000 телеграмм было полностью расшифровано только 1043, требовалось доказать, что эти 1043 телеграмм достаточно репрезентативно представляют общую выборку по интересующим параметрам. Также в процессе исследования введено ещё несколько параметров, в основном представляющих собой обобщения уже существующих. Обе задачи были также решены программными методами.

Следующим шагом было само исследование уже обработанных данных. Инструмент анализа варьировался в зависимости от типа задачи и визуализации: если большинство гипотез было достаточно проверить простыми статистическими подсчётами, которые проводились с помощью языка программирования Python 3, то некоторые из них требовалось проверять с помощью методов сетевого анализа и визуализации карт. Большинство используемых в работе программ находятся в открытом доступе и позволяют совместную параллельную работу. Про каждую из них будет рассказано более детально.

Описание выбранных методов и алгоритмов

Большинство поставленных перед задач можно было выполнить с использованием скриптов, написанных на Python 3. Предобработка и сортировка данных, общий анализ текстов и проверка двух первых гипотез проводилась именно с их помощью. Сами данные хранились в виде таблиц формата .xlsx, также известного как «экселевский». Преимущество такого формата заключается в том, что данные удобно редактировать и просматривать как разметчику-волонтёру, так и программными методами. Код был написан в интерпретаторе Google Colab[6], что позволило делиться результатами работы и оперативно исправлять обнаруженные неточности. Каждая таблица хранилась на Гугл-диске, доступ к которому был у всех участников команды, и подключалась к скрипту с помощью команды для подключения файлов к интерпретатору, взятой из библиотеки google.colab[7]. Для анализа были использованы методы библиотеки pandas. Данная библиотека преобразует таблицу в формат DataFrame[8] («датафрейм»), который считается одним из самых удобных форматов для работы с большими данными и проведения квантитативного анализа.

Самым важным этапом предобработки была унификация формата записи даты. Для выполнения этой задачи применены регулярные выражения, в частности, библиотека re[9]. Сначала было написано регулярное выражение, с помощью которого отсекали все данные, которые уже подходят под выбранный шаблон. Затем с помощью другого регулярного выражения удалось найти группы из четырёх цифр, стоящих подряд, и таким образом большинство телеграмм получило датировку если не годом, то хотя бы десятилетием. Наконец, было применено регулярное выражение для поиска месяцев и дней, причём для месяцев был создан и специальный словарь[10], заменяющий текстовое значение месяца на числовое (например, и слово «февраль», и слово «фев» заменялось на «02»). С помощью метода регулярных выражений удалось отформатировать даты большинства телеграмм. Для облегчения задачи любое обозначение даты, по которому нельзя было определить дату отправки как минимум с точностью до десятилетия, считалось недостаточно информативным и ему автоматически присваивали значение «xx.xx.xxxx». Таким образом, все значения вида «вторая половина 19 века» автоматически не расшифровывались.

---

[6] Google Colab — онлайн-интерпретатор для языка Python, созданный Google.

[7] google.colab — специальная библиотека для языка Python, созданная Google для работы в их онлайн-интерпретаторе.

[8] DataFrame — табличный формат хранения данных, адаптированный под машинную обработку.

[9] re — библиотека для языка Python, предназначенная для работы с регулярными выражениями.

[10] Словарь в языке Python — неупорядоченная коллекция произвольных объектов, записанная в формате «ключ-значение», где ключ—- уникальный индикатор для значения.

Следующей после унификации дат в обоих датасетах стала задача доказательства репрезентативности выборки. Исходя из того, что в большинстве гипотез основополагающим параметром сравнения является дата отправки телеграммы, было принято решение проводить верификацию репрезентативности выборки именно по этому параметру. Поскольку в ходе исследования рассматривается телеграф и телеграммы с исторической точки зрения, наиболее репрезентативным и эргономичным было разбить все телеграммы на группы, исходя из их принадлежности к определённому периоду в истории нашего государства, и таким образом сравнить распределение двух выборок между собой. Был составлен словарь принадлежности, где ключом[11] являлся кортеж[12] из двух числовых значений (начало и конец определённой эпохи в истории России), а значением[13] — название этой эпохи; также была написана функция, которая сопоставляет данное значение с имеющимся в словаре и исходя из года отправки присваивает телеграмме определённую эпоху. После применения данной функции к обоим датасетам и последующей сортировки полученных данных методами из библиотеки pandas[14] репрезентативность выборки удалось доказать. Более подробно итог выполнения этого скрипта описан в разделе «Результаты».

В ходе исследования пришлось ввести ещё несколько категорий, которые не были предусмотрены инструкцией разметчика и, как следствие, не были указаны ни в одном из датасетов. В основном эти категории являлись, как уже было сказано, обобщением какого-либо уже существующего параметра, и добавление новых критериев производилось с помощью уже доказавшей свою эффективность комбинации из словаря соответствия и функции, которую требовалось применить на существующем критерии для получения нового. В таких словарях соответствия ключом всегда является значение данного критерия, а значением — предполагаемое значение будущего критерия. С помощью функции .apply(), входящей в инструментарий библиотеки pandas, можно применять любую функцию, в том числе и самостоятельно написанную, на столбец или несколько столбцов датафрейма сразу. Также потребовалось выполнить несколько математических операций; в частности, для одного из критериев нужно было рассчитать меру энтропии Шеннона[15]. Эта метрика используется для определения разнообразия данных и её числовое значение прямо пропорционально количеству разных значений в выборке. Мера энтропии Шеннона примечательна тем, что она нечувствительна к размеру выборки, поэтому используется в

[11] Ключ — уникальный индикатор данных в словаре.
[12] Кортеж — упорядоченный набор данных.
[13] Значение — содержимое словаря.
[14] Pandas — библиотека для языка Python для работы с датафреймами.
[15] Мера энтропии Шеннона — метрика, которая используется для определения разнообразия данных и её числовое значение прямо пропорционально количеству разных значений в выборке.

качестве средства нормализации датасета. Для подсчётов применялись библиотеки numpy[16] и scipy.stats[17], которые являются одними из самых популярных в решении математических задач программными методами. Для работы с текстовыми данными использовались заготовленные заранее скрипты для лемматизации и удаления стоп-слов: они были сделаны с помощью библиотек pymorphy. Также впоследствии была добавлена библиотека wordcloud для составления облака слов.

Наконец, для большей ясности представления результатов использовались инструменты визуализации. Существует множество библиотек на языке Python 3, подходящих для этой цели, но было решено взять matplotlib[18] по причине её популярности, совместимости с форматом данных вида «датафрейм» и простоты работы с ней. Данная библиотека позволяет рисовать круговые диаграммы, гистограммы и обычные графики на координатной плоскости. С её помощью проведена большую часть работы по визуализации данных.

Помимо уже указанных методов, в решении задач с социальными связями также применялись методы сетевого анализа; задачи же географические решались с помощью интерактивных карт. Благодаря специальному модулю для языка программирования R под названием «Leaflet»[19] координаты для крупных городов были проставлены автоматически; координаты же для более мелких.населённых пунктов были проставлены вручную.

Результаты

В данной работе стояли две цели: практическая — создание базы данных, которую можно было бы пополнять усилиями волонтёров и публиковать в открытом доступе, и академическая — проведение анализа полученных данных и проверка сформулированных гипотез с помощью методов компьютерной лингвистики и Digital Humanities. Обе задачи можно считать выполненными.

Создание базы данных

Для того, чтобы все телеграммы были собраны в удобном формате, было решено создать базу данных, которая находится в формате Excel-таблицы. Для таблицы были тщательно продуманы поля, добавлялись новые, убирались ненужные. В итоге, получилось создать такую базу, в которой можно было как посмотреть текст самой телеграммы, так и узнать дату

[16] Numpy — библиотека для языка Python, добавляющая поддержку больших многомерных массивов и матриц, а также математических вычислений, начиная с базовых функций и заканчивая линейной алгеброй.

[17] Scipy.stats — библиотека для языка Python, позволящая производить статистический анализ данных.

[18] Matplotlib — библиотека для языка Python, позволящая визуализировать данные.

[19] Leaflet — библиотека для языка R, которая позволяет работать с географическими данными и визуализировать их.

её написания / отправления, куда она была доставлена и откуда отправлена. На этом база не ограничивается, иначе анализа бы не вышло. Так, была разработана система, которая отображает тон телеграммы, её тематику — с какой целью отправлялась телеграмма. Такие поля позволяют сравнить, в чём была цель того или иного послания. Существуют также и другие поля, которые позволяют проследить некоторые особенности письма или помогают обработать данные.

Так, были созданы поля с географическими координатами получателя и отправителя, для того, чтобы понимать всю географию телеграмм. Благодаря этому удалось посмотреть на разделение по тону: откуда отправлялись положительные, а откуда отрицательные сообщения при помощи телеграфа. В итоге получились две карты (см. приложение 2): первая из них отражает географию тона для отправителя телеграммы, вторая — для получателя.

Дополнительные пункты в базе данных позволяют определить язык телеграммы, её графику (кириллица или латиница) и персоналии, которые упоминаются непосредственно в тексте телеграммы. Существуют также разные виды телеграмм: обычные, рукописные и фототелеграммы. Подобные элементы таблицы помогают понять, как нужно работать с конкретной телеграммой (например, если для рукописной телеграммы орфографические ошибки или опечатки появляются по вине автора, то для печатной телеграммы сложнее установить виновника).

Предпоследним пунктом по порядку, но одним из важнейших в базе является источник, то есть, откуда телеграмма была получена. Это даёт понимание, как с ней дальше работать. То есть, если телеграмма получена из семейного архива, это одна ситуация. В ней скорее всего не будут упомянуты важные исторические и культурные персоны, она будет адресована близкому семейному кругу и так далее. В случае с музейными экспонатами, существует ровно противоположная ситуация, она вполне может быть адресована крупному начальнику и содержать упоминания множества персоналий. Такие телеграммы требуют другого подхода. Таким образом, таблица призвана охватить весь спектр телеграмм, которые существуют вообще, потому что включает самые разнообразные пункты.

Последним же параметром, который необходим для анализа телеграммы являются примечания: в них попадают как опечатки, так и фактические ошибки автора телеграммы, как например, неверное написание имени и фамилии человека, если речь идёт, например, о телеграмме из семейного архива, датировка телеграммы, которая разнится со штампом на самой телеграмме, устаревшие названия городов, которые сегодня имеют другие официальные наименования и так далее.

Анализ полученных результатов

Общий анализ данных

Для начала стоит уделить внимание общему виду полученных данных. С целью корректировки гипотез проведён поверхностный общий анализ обоих датасетов и получены примерно следующие результаты:

Поскольку всего критериев больше двадцати, будут рассмотрены только самые важные для работы параметры, а именно даты и тематику телеграммы, так как большинство гипотез завязано на них. Самая ранняя телеграмма, которая у нас имеется, датирована 1877 годом. Самая поздняя — 2014 годом. Из общего количества телеграмм 91.33% датированы как минимум годом (с точностью до десятилетия), 31.56% — хотя бы месяцем. В малом датасете результаты примерно такие же: 92.33% датированы как минимум годом, 39% — хотя бы месяцем. Сильные различия в процентах объясняются тем, что при ручной разметке можно получить информацию, которую невозможно извлечь машинными методами, и поэтому ручная разметка более полна. Также получилось вычленить 43 различные темы телеграмм и распознать несколько случаев, когда несколько тем присваивалось одной телеграмме.

Проверка репрезентативности выборки

Поскольку из 90000 телеграмм размечена только тысяча, требовалось доказать, что малый датасет является репрезентативной выборкой из большого. Как уже говорилось, в качестве критерия распределения была выбрана принадлежность телеграмм к той или иной эпохе. Это удалось сделать с помощью уже описанных выше методов. Результаты были следующие:

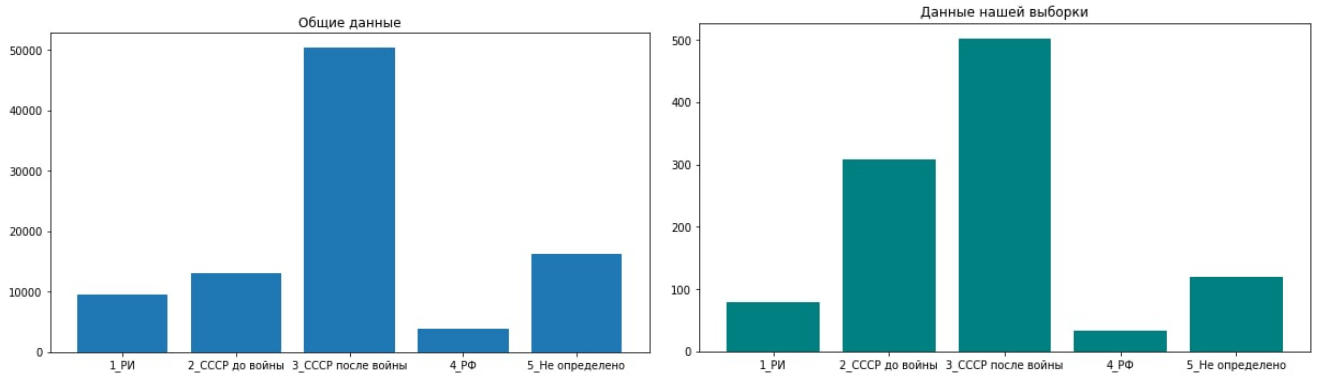

Рис. 5-6. Сравнение временного распределения большого датасета (слева) и малого датасета (справа).

Распределения и правда похожи. Если обратить внимание на второй столбец, «СССР до войны», можно заметить, что в малом и в большом датасетах процентное соотношение к общему числу телеграмм сильно различается (довоенные телеграммы времён СССР

составляют 14% от телеграмм большого датасета, но в малом они занимают 29%). Предполагается, что причиной этому стали обнаруженные и впоследствии исправленные при ручной разметке неточности в датировке. В остальных же столбцах данные примерно совпадают (см. таблицу).

	Большой датасет (% от общего числа телеграмм)	Малый датасет (% от общего числа телеграмм)
1887-1917 годы	10.250416	7.574305
1917-1945 годы	14.026489	29.626079
1945-1991 годы	54.056518	48.130393
1991-2014 годы	4.178240	3.259827
Дата не определена	17.488337	11.409396

Рис. 7. Таблица временного распределения большого и малого датасета.

Гипотезы

Самой важной частью академической части работы является, несомненно, доказательство гипотез. Изначально их было шесть: четыре основных гипотезы и две второстепенные, доказательство которых было бы возможно только при достижении определённых условий. Однако с течением времени гипотезы немного изменились, когда было получено более ясное представление о данных, с которыми предполагается работать.

Было принято решение полностью отказаться от четвёртой гипотезы; она состояла из двух частей, и обе части, к сожалению, доказать не удалось. Поскольку только 39% телеграмм (около 400 при минимальном рекомендованном объёме датасета в 1000 штук) датированы с точностью до месяца, невозможно сделать никаких чётких выводов касательно того, учащались ли поздравления в пору праздников или нет. Также при анализе оказалось, что тема «девятое мая» была упомянута только дважды: один раз в шестидесятые и один раз в восьмидесятые, что тоже не является хорошей почвой для каких-либо выводов касательно упоминаемости Дня победы.

На данный момент гипотезы и их доказательства выглядят так:

Гипотеза 1

«С развитием других методов коммуникации, помимо телеграфа, пул тем телеграмм постепенно сужался, и в последние годы популярности телеграфа среди тем превалировали поздравления и приятные новости.»

В целом можно сказать, что это скорее так: количество разных тем, приходящихся на одно десятилетие, действительно росло, достигнув пика в сороковые годы, а затем медленно спадало с течением времени.

Рис. 8. Разнообразие тем, сгруппированное по десятилетиям. Последний столбец (xxx) — дата неизвестна.

Затем было решено провести похожий анализ, но с помощью других метрик: была принята попытка нормализации данных. Темы объединены в макро-группы, исходя из их примерной принадлежности: «личные поздравления» — это поздравления с персональными памятными датами (дни рождения, свадьбы, рождения ребёнка, присвоение награды), «государственные праздники» (8 марта, Новый год), «устаревшие праздники» (праздники, которые больше не отмечаются, чаще всего — локальные советские), «военно-политические» (манифесты, военные телеграммы), «будничное» — простые сообщения по типу просьбы или сообщения новостей. На графике действительно можно увидеть, что со временем категория «личные достижения» начинает превалировать над остальными.

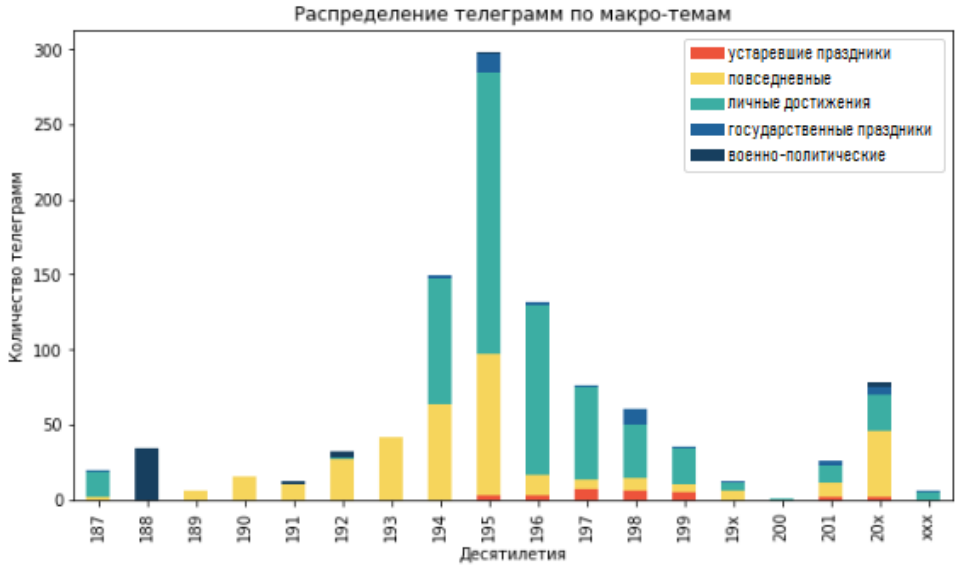

Рис. 9. Распределение телеграмм по тематическим кластерам.

Также заметно, что количество телеграмм на повседневные темы сократилось: скорее всего, это связано с развитием других типов коммуникации, которые позволяли передавать информацию быстрее.

Затем было решено посчитать разнообразие тем с помощью другой метрики — меры энтропии Шеннона. Как уже было сказано выше, данная метрика позволяет измерить «разнообразие» данных в выборке, не обращая внимания на её размер; поэтому релевантно применить её здесь, чтобы посмотреть, работает ли гипотеза, если не брать в расчёт тот факт, что в разные десятилетия было отправлено разное количество телеграмм.

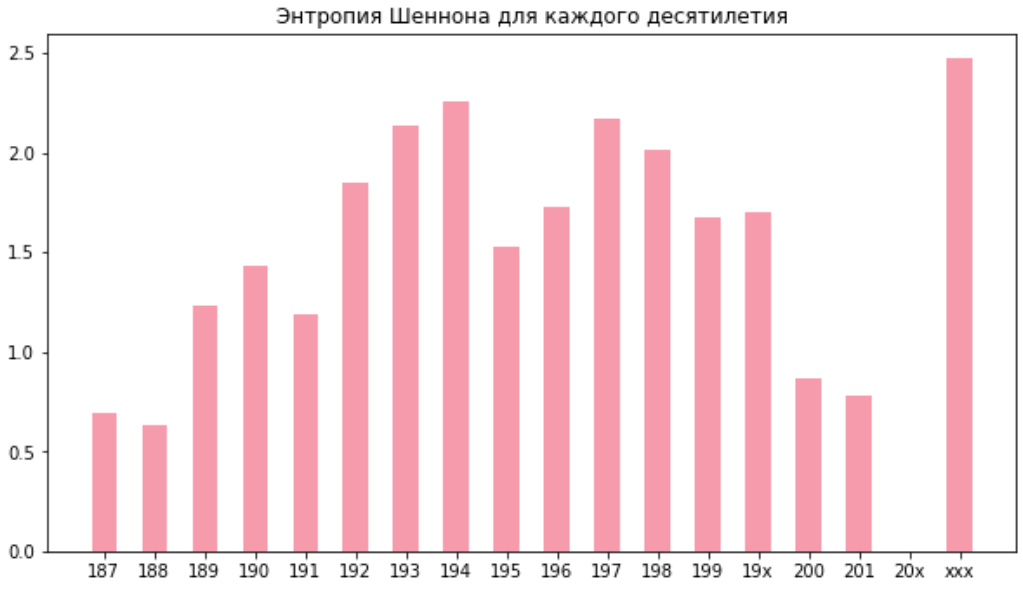

Рис. 10. Разнообразие тем с учётом меры энтропии Шеннона.

Как можно заметить, общее сужение тематического разнообразия во второй половине 20 века можно заметить даже здесь, хотя оно и не так заметно.

Таким образом можно сказать, что гипотеза действительно подтвердилась: пул тем действительно начал сокращаться с появлением альтернативных способов коммуникации, и с течением времени превалировать стали телеграммы с позитивной коннотацией, а именно поздравления различного типа.

Гипотеза 2

«Если объединить и отсортировать телеграммы по году отправки, то, вероятно, окажется, что в годы, которые в истории России обозначены важными событиями, телеграммы отправлялись чаще.»

Здесь следует отметить, что, поскольку на разные периоды приходится разное количество телеграмм, любой рост или падение частотности отправления будет также отображаться по-разному. Однако точное количество отправленных телеграмм для каждого года известно, следовательно, заметить изменение будет возможно даже для временных периодов с небольшим количеством телеграмм.

Для начала требовалось определить, какие именно годы будут считаться *«годами, которые в истории России обозначены важными событиями»*. Было решено взять следующие общеизвестные события предыдущего столетия: Октябрьская революция (1917), Великая Отечественная война (1941-1945), полёт Юрия Гагарина в космос (1961) и распад Советского Союза (1991). Логично, что если реакции на события государственного масштаба в виде роста количества отправленных телеграмм обнаружено не будет, то, следовательно, на события более мелкого масштаба такой реакции тоже не стоит ожидать.

Рис. 11. Частотность отправления телеграмм по годам.

Как можно заметить, тенденции резкого роста частотности отправления телеграмм в годы, когда происходили ключевые исторические события, обнаружено не было. Максимальное значение — пик частотности — приходится на 1950 год; эта дата была присвоена около 2500 телеграммам. Существует предположение, что это ошибка датировки: возможно, если датирующий не мог точно определить, когда именно была отправлена телеграмма, он присваивал ей примерное значение, а 1950 год, который является серединой столетия, отлично подходит на роль фиктивного значения. Учитывая, что в большом датасете попадались значения порядка «Вторая половина 20 века», не исключена такая возможность.

Гипотеза 3

«*Имена общественных и политических деятелей могут упоминаться как в официальной, так и в личной переписке. Нередко это связано со смертью деятеля*».

После применения метода именованных сущностей к размеченным телеграммам оказалось, что это не так — люди, не имеющие непосредственного отношения к политическому аппарату, в личной переписке не обсуждали общественных и политических деятелей, даже после их смерти. Однако, работая с именами политических деятелей, удалось обнаружить информацию, оказавшуюся полезной для подтверждения двух второстепенных гипотез, о чём более подробно будет рассказано в следующем разделе.

Сетевой анализ

С помощью методов сетевого анализа было задумано сопоставить данные о взаимодействиях известных политических деятелей посредством телеграмм в рамках малого датасета с историческими реалиями. Так, например, решено посмотреть круг взаимодействий Иосифа Виссарионовича Сталина с течением времени, по мере того, как нарастал градус политических репрессий и увеличивался список «врагов народа».

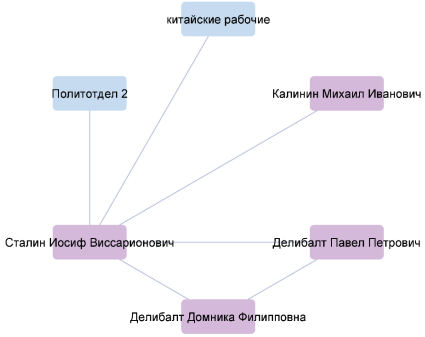

Рис. 12. Взаимодействия И. В. Сталина в период с 1919 по 1930 год.

Данный граф представляет собой взаимодействия Сталина в период с 1919 по 1930 год. Сиреневым цветом выделены отдельные, частные лица, а голубым — коллективы и организации. Как видно, в этот период связи Сталина в рамках корпуса представлены довольно скудно.

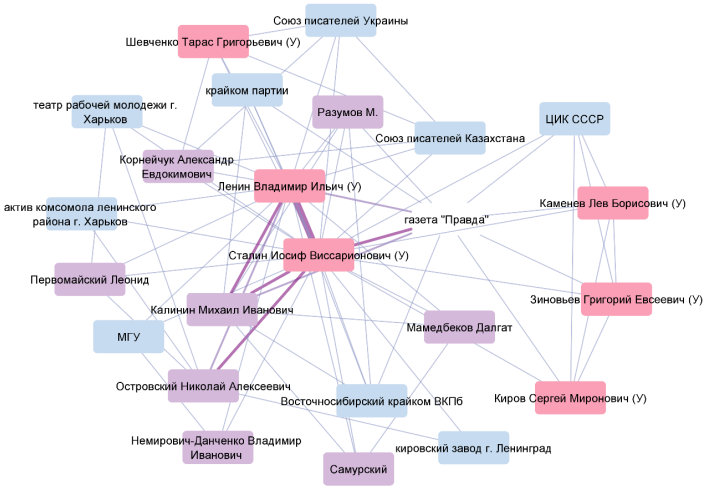

Рис. 13. Взаимодействия И. В. Сталина в период с 1931 по 1940 год.

Второй граф — взаимодействия Сталина в период с 1931 по 1940 год. Легенда остаётся прежней, но добавляется ещё один параметр: розовым цветом выделены упоминания: под упоминанием подразумевается то, что лицо не фигурирует в качестве отправителя или получателя телеграммы, вместо этого просто упоминается в тексте. Ширина рёбер распределена по весу. Здесь видно, что самая сильная по весу связь между двумя узлами — Сталиным и Лениным. Это объясняется тем, что имена этих политиков стали чем-то чуть ли не нарицательным в текстах телеграмм, и их часто упоминали вместе как символов всей советской власти: например, поздравляя известную личность с получением награды, отправитель выказывал уважение получателю, потому что последний поддерживает «дело великих Ленина и Сталина». Также среди упоминаемых политиков можно заметить Каменева и Зиновьева, которые в рассматриваемый промежуток времени были репрессированы и расстреляны. В корпусе каждый из них упоминался лишь однажды — в гневной телеграмме в адрес ЦИК СССР и газеты «Правда» с грубой критикой оппозиционной деятельности репрессированных и поддержкой деятельности Сталина. Также в этой телеграмме упоминался и Киров — политик, убийство которого послужило началом массовых репрессий в СССР.

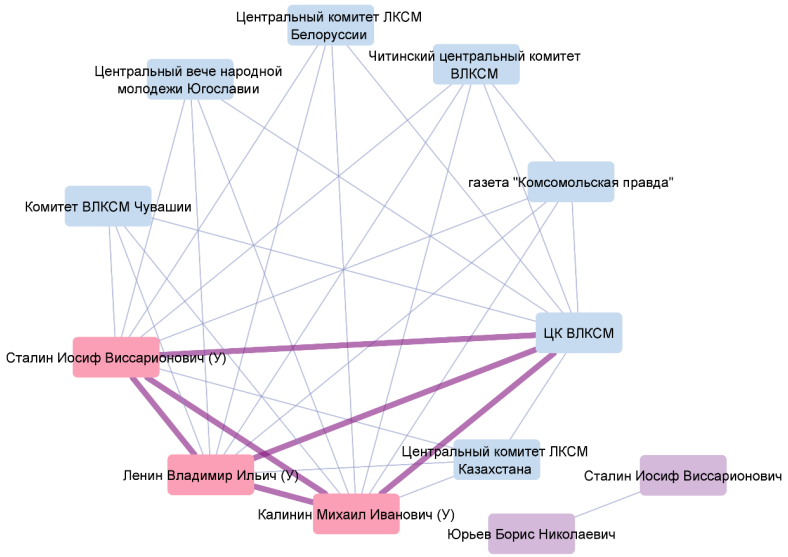

Рис. 14. Взаимодействия И. В. Сталина в период с 1941 по 1950 год.

Третий граф представляет собой взаимодействия Сталина в 1941-1950 гг. Здесь Сталин фигурирует как в упоминаниях, так и как получатель телеграммы. Большая часть телеграмм за этот период была отправлена в 1946 году в связи со смертью Михаила Ивановича Калинина. Функция связи между Сталиным и Лениным остаётся той же, что и в предыдущем графе: имена этих политиков синонимичны самой советской власти; отправители выражают соболезнование по поводу кончины Калинина, соратника «великих Ленина и Сталина».

Помимо этого, в этом графе можно заметить интересную деталь: среди телеграмм, отправленных в 1946 году, есть и телеграмма от Центрального вече народной молодёжи Югославии, опять же, с выражением соболезнований по поводу смерти Калинина и приветствием его деятельности во благо социализма. Как известно из истории, отношения между Советским Союзом и Югославией сильно ухудшились в 1948 году из-за разногласий Сталина и лидера коммунистической партии Югославии Иосипа Броз Тито, поэтому представить подобную телеграмму всего на пару лет позже было бы трудно.

Так, с помощью методов сетевого анализа была рассмотрена корреляция между реальными историческими прецедентами и коммуникацией политических деятелей посредством телеграмм в течение различных промежутков времени.

Дополнительные заметки и наблюдения

Помимо проверки четырёх основных и двух второстепенных гипотез, также был проведён небольшой локальный хакатон[20], целью которого было найти интересные закономерности в данных. Вот что получилось.

Как уже упоминалось ранее, из-за недостатка размеченных данных никакой существенной информации касательно Дня Победы обнаружить не удалось. Было решено посмотреть статистику по двум другим популярным праздникам советского происхождения — Первомаем (отмечается 1 мая) и годовщиной Октябрьской революции (отмечается 7 ноября).

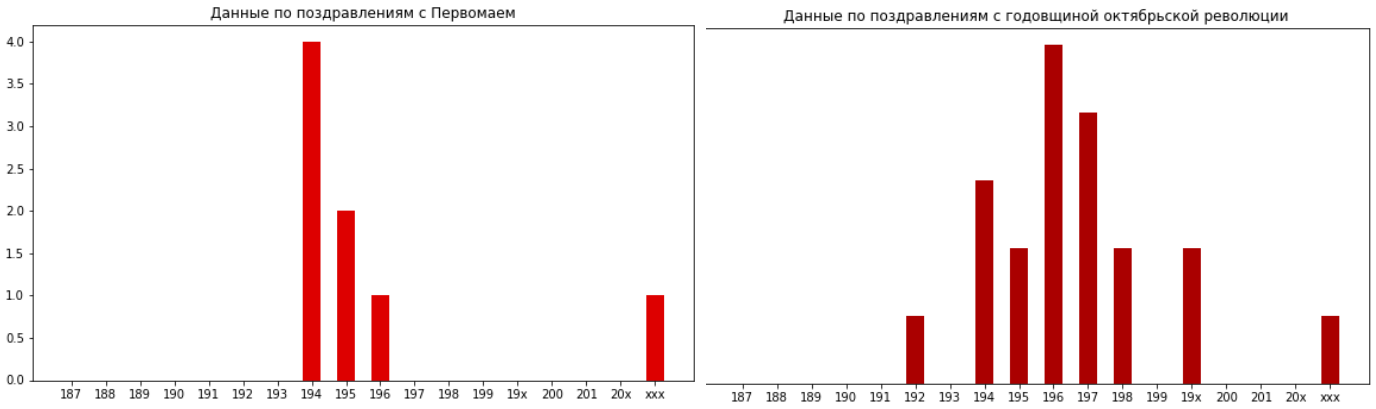

Рис. 15-16. Временное распределение поздравлений с Первым мая и годовщиной Октябрьской революции.

Оба праздника действительно упоминаются в основном только во времена Советского Союза и со временем исчезают из пула тем.

Самая популярная тема в датасете — «новости»; она обсуждалась в 215 из 1043 телеграмм. Однако никакой закономерности распределения обнаружено не было.

[20] Хакатон — мероприятие для разработчиков, на котором они сообща решают какую-либо задачу программными методами.

Рис. 17. Временное распределение темы «новости».

Был проведён анализ и текстового поля. Например, выяснилось, что опечатки в телеграммах встречаются чаще, чем орфографические ошибки. Они же являются самым частым видом нестандартной разметки: орфографические ошибки стоят на втором месте, следом идёт дореволюционная орфография, затем аббревиатура и, наконец, несовпадение графики.

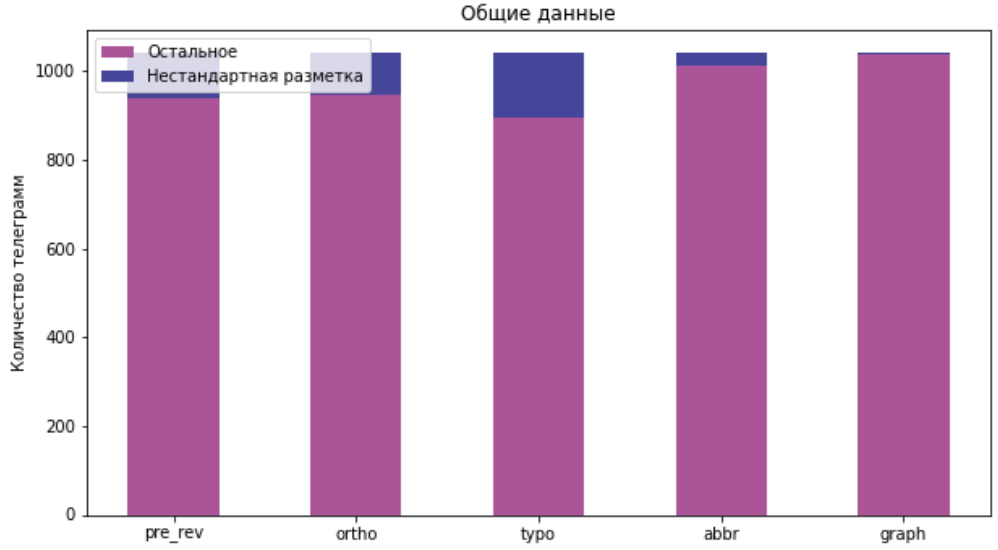

Рис. 18. Отношение нестандартной разметки к нормативной.

Дореволюционная орфография была официально упразднена в ходе орфографической реформы 1917-1918 годов. Однако поскольку телеграммы — это личная переписка, высказано предположение о том, что в них люди общались так, как им привычнее, следовательно, телеграммы с дореволюционной орфографией появятся и после 1917 года. Так и получилось: последние телеграммы, датированные с точностью до года, в которых была обнаружена дореволюционная орфография, датируются 1933 годом. В основном это телеграммы, отправленные от лица или на имя государственных организаций. На графике есть две колонки ещё правее, однако на них обращать внимания не следует, так как это статистика по телеграммам, датированным двадцатым веком и телеграммам, не датированным вовсе.

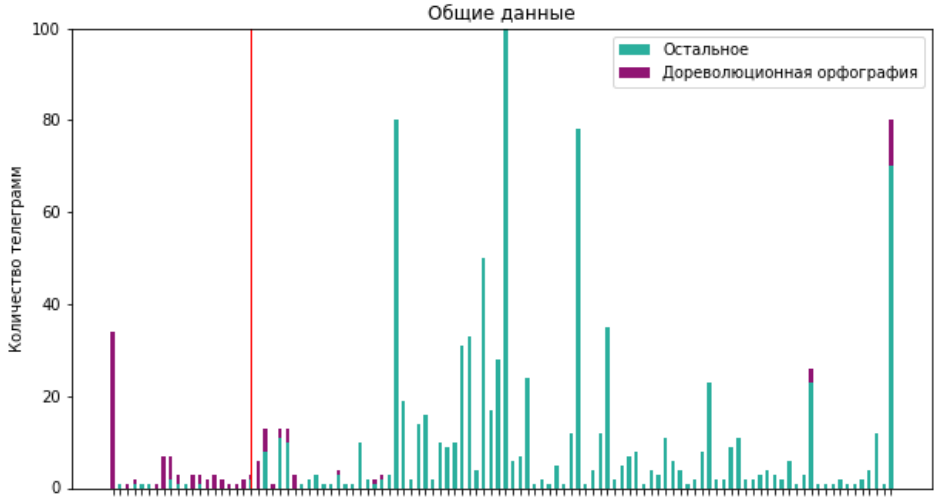

Рис. 19. Временное распределение телеграмм с дореформенной и послереформенной орфографией. Красным обозначен 1917 год.

Была также мысль, что в советское время аббревиатуры встречались сильно чаще, чем в дореволюционное время, и, возможно, чаще, чем после распада СССР, поскольку тенденция сокращать названия была сильно популяризирована именно во времена Советского Союза и частично перешла и в современную Россию. Однако на деле оказалось, что это не так — скорее всего, потому, что не все сокращения являются аббревиатурами (например, *комсомол* и *партбилет*).

Рис. 20. Временное распределение аббревиатур в текстах телеграмм. Красным выделен период существования Советского союза.

Рассмотрена и встречаемость орфографических ошибок в телеграммах, чтобы проверить, повлияли ли советские образовательные реформы на грамотность населения, но тоже не нашли никаких закономерностей.

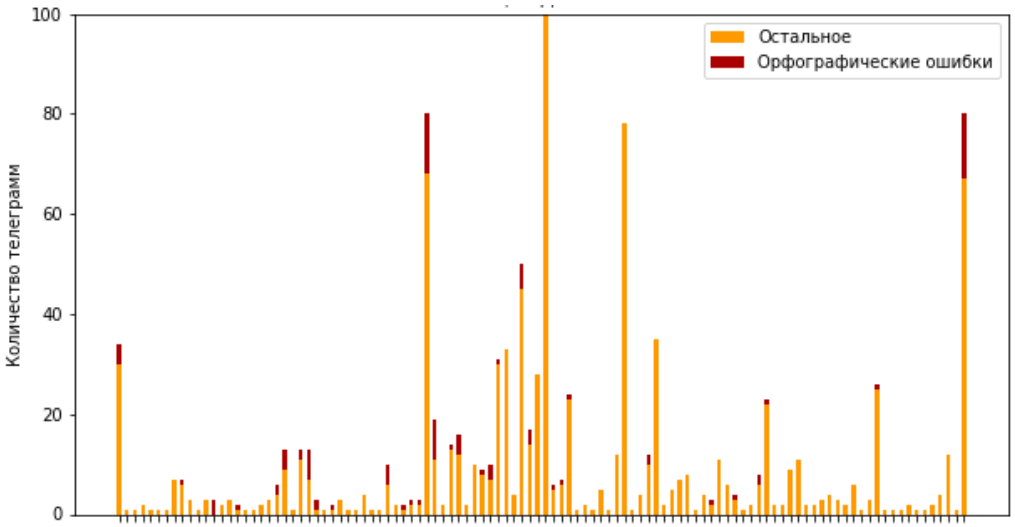

Рис. 21. Встречаемость орфографических ошибок в телеграммах.

Была и идея составить облака слов для каждой из тематических макро-групп, выделенных в процессе проверки первой гипотезы.

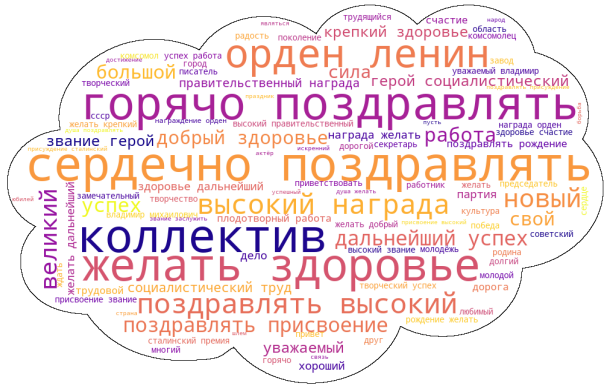

Рис. 22-26. Облака слов с самыми частотными словами по тематическим кластерам. Слева направо и сверху вниз: будничные телеграммы, устаревшие праздники, личные поздравления, военно-политическое и государственные праздники.

В телеграммах, темы которых посвящены каким-либо праздникам, самые частотные слова практически одни и те же — «поздравлять», «желать», «здоровье», «успех» с небольшими различиями в зависимости от типа праздника. Например, в устаревших праздниках также частотны такие слова, как «съезд», «октябрь», «великий»: это связано с тем, что самая распространённая тема в рамках этого кластера — годовщина октябрьской революции. В личных поздравлениях среди самых частотных слов — «высокий», «награда», поскольку чаще всего в этом кластере встречалась тема «получение награды», что, опять же, связано с особенностями выборки.

В будничных телеграммах самые распространённые слова — это «просить», «работа», «советский», «привет», «товарищ», «совет», «партия» и т. д.. Примечательно, что некоторые из этих слов представляют собой атрибуты советской эпохи, что весьма закономерно, потому что большинство телеграмм в корпусе принадлежит именно этому временному периоду.

В телеграммах, посвящённых военно-политической тематике, чаще всего встречаются слова соответствующей направленности — «потеря», «генерал», «войско», «отряд», «атаковать» и др. Самое частотное слово в этой категории — «турок», и это не случайно: ряд телеграмм в рамках этого кластера представляет собой новостные сводки времён Русско-турецкой войны 1877-1878 годов. Также одним из самых частотных слов в этой категории является «октябрь», как и в устаревших праздниках, что объясняется политической аффилиацией данного слова в контексте Октябрьской революции 1917 года.

Будничные телеграммы		Устаревшие праздники		Государственные праздники		Личные праздники		Военное и политика	
советский	42	поздравлять	23	поздравлять	32	поздравлять	269	турок	35
привет	33	желать	20	желать	30	желать	218	потеря	22
товарищ	31	праздник	14	новый	20	здоровье	148	генерал	19
москва	30	успех	11	здоровье	16	успех	129	войско	18
армия	30	здоровье	10	счастие	16	высокий	105	взять	17
совет	30	работа	10	успех	16	награда	89	турецкий	15
телеграмма	30	съезд	10	город	11	работа	81	октябрь	14
рабочий	26	октябрь	9	творческий	10	горячо	79	отряд	14
молодёжь	25	великий	9	дорогой	9	сердечно	75	позиция	14
принять	24	сердечно	8	трудящийся	7	дальнейший	72	офицер	14

Рис. 27. 10 самых частотных слов во всех тематических кластерах.

Заключение

В результате исследования удалось подтвердить гипотезу о том, что диапазон тем стал сокращаться по мере развития альтернативных средств коммуникации. Другая гипотеза была опровергнута: не было выявлено связи между историческими событиями и частотностью отправляемых телеграмм. Две оставшиеся гипотезы не удалось проверить не удалось. Тем не менее, получилось проверить одно из второстепенных предположений — мы проследили взаимосвязь между реальными политическими событиями в СССР и правительственной перепиской в телеграммах, что позволяет по-новому взглянуть на историю.

Также важно отметить ту работу по созданию базы данных с телеграммами, которая была нами проделана. Несмотря на то, что особенности нашей выборки накладывают определённые ограничения, нам всё же удалось осуществить поставленные цели. Мы стремимся и дальше дополнять наш корпус и продолжать наше исследование. В частности, хотелось бы исследовать корреляцию времени приёма и передачи с видами телеграмм (например, проверить, какие виды телеграмм доставлялись быстрее — срочные, правительственные или местные). Желательно бы также создать сайт нашего проекта и попробовать TEI-разметку[21] для хранения данных.

[21] TEI (Text Encoding Initiative) — стандарт электронного представления текста, распространённый в среде цифровой гуманитаристики.

Помимо этого, мы стремимся создать отдельный корпус телеграмм, который не привязан к порталу открытых данных Минкульта РФ, чтобы иметь возможность публиковать данные. В настоящий момент мы имеем право на публикацию лишь текстовых полученных телеграмм, но не самих изображений, и, соответственно, мы не можем их разместить на своём потенциальном сайте.

Подводя итог, необходимо отметить, что телеграммы — большой и неизведанный пласт информации, который всё ещё открыт к исследованиям и интерпретации, несмотря на то, что телеграфное сообщение устарело как вид коммуникации. Телеграммы позволяют посмотреть под другим углом на жизнь и события предыдущих эпох.

Источники

1. Портал открытых данных Министерства культуры Российской Федерации [Электронный ресурс] // URL: https://opendata.mkrf.ru/opendata (дата обращения: 08.06.2022)
2. Standage, T. The Victorian Internet: The Remarkable Story of the Telegraph and the Nineteenth Century’s on-Line Pioneers. New York, N.Y.: Berkley Books, 1999.
3. Wenzlhuemer, R. Connecting the Nineteenth-Century World: The Telegraph and Globalization. Cambridge: Cambridge University Press, 2012. doi:10.1017/CBO9781139177986
4. «Пишу тебе» [Электронный ресурс] // URL: https://sysblok.ru/postcards (дата обращения: 10.06.2022)

Приложения

1. Репозиторий проекта — https://github.com/marinanastas/telegrams
2. География телеграмм — https://rpubs.com/purrmorning/900880