Муниципальная бюджетная образовательная организация “Средняя общеобразовательная школа №27 им. И. А. Курышева”

Исследовательская работа

“Частотный анализ речи современных политиков на основе микроблога Twitter”

Выполнил: ученик 11Б класса
Якунин Олег Сергеевич

г.Чита

2021

Оглавление

Twitter, политики и профориентация **3**

Глава 1. Объект исследования. Методика исследования. **6**

Обзор литературы 6

Объект исследования 6

Получение доступа 7

Получение и обработка данных 7

Глава 2. Результаты исследования **9**

Основные темы в речи 9

Наблюдение о распределении значимых слов в корпусе 11

Риторические инструменты политиков 12

Владимир Соловьёв 12

Владимир Жириновский 13

Алексей Навальный 14

Михаил Светов 15

Итоги исследования **16**

Список использованной литературы **16**

Приложения **18**

Приложение А. Письма с собеседования 18

Приложение Б. Список стоп-слов 20

Приложение В. Программный репозиторий 21

Приложение Г. Графики 22

Twitter, политики и профориентация

«Твиттер» — это одна из крупнейших социальных сетей мира, где взаимодействие строится на «твитах», которые видны всем и на которые все могут отвечать.

Поскольку социальная сеть изначально планировалась как платформа с мобильным доступом, но создавалась в 2004-2005, когда мобильный интернет был несовершенен и не все телефоны могли вообще им пользоваться, публикация была основана на SMS: пользователь отправляет на номер «Твиттера» сообщение, тот смотрит, к какому профилю номер привязан, и публикует из-под него сообщение на обозрение всем. Эта мера была остроумным решением для тех лет, однако это создало ограничение: поскольку исторически максимальный размер SMS — 160 латинских символов, то и твит не мог быть больше этой величины. Чтобы сравнять тех, кто пишет с компьютера и с телефона, для всех было установлено ограничение в 140 символов (в 2017 расширено до 280). SMS-доступ вскоре был упразднён, а ограничение осталось в качестве изюминки платформы.

Таким образом, твит — это ёмкая, сконцентрированная законченная мысль (в лучших традициях Спарты!). Для того, чтобы вместить в ограничение свою мысль, нужно тщательно выбирать наиболее выразительные и ёмкие слова. Для обычного человека это новый подход к выражению мысли и повод переосмыслить свою речь и количество бессмысленной болтовни в ней. А для политиков это особенное соревнование. Людям, чья работа — много говорить, приходится отказываться от распространённого современного паттерна "заговаривай народ цистерной популизма и демагогии с умным видом, чтоб никто ничего не понял, но проникся твоей серьёзностью" и возвращаться к ещё античной модели — "чем меньше слов, тем лучше (но склонить на свою сторону всё же надо)". Опять-таки, вспоминается Спарта, но это также касается древнеримских политиков, мыслителей, императоров — тех, кто ещё помнил, что такое риторика как ораторское искусство.

Такая чистая, сжатая мысль является очень удобным материалом для анализа — в ней сосредоточена вся суть высказывания.

К вопросу о том, насколько “серьёзны” твиты как источник политической речи. Нельзя недооценивать вес короткого высказывания в Интернете. Как показывает практика, пара предложений — или всего несколько слов — от имени значимого человека в Сети имеют не меньшие последствия для реального мира, чем их живые выступления. Пример: один из известнейших пользователей Твиттера — действующий на момент исследования президент США Дональд Трамп. Через Твиттер лидер делает все обращения, рассказывает о том, какие готовятся законопроекты, делает

заявления, в том числе обращенные лидерам других стран, чем зачастую едва не вызывает военные конфликты[1]. В декабре 2016 он одним твитом обвалил[2] акции военной корпорации Lockheed Martin, уронив ее капитализацию на $3,5 млрд, а 2 апреля 2020 года, рассказав в твите о якобы свершившейся сделке России и Саудовской Аравии, о которой он узнал из телефонного разговора с Путиным, немедленно поднял[3] стоимость нефти Brent на 21%. Фактически через эти короткие высказывания он не только управляет собственной страной и напрямую общается с народом (чтобы СМИ не искажали его позицию, по его собственному высказыванию), но и влияет на мировую экономику. Судьба нации умещается в 280 символов.

Создание микрокорпуса из тысяч твитов политических активистов и применение к нему метода частотного анализа позволяют не только оценить обстановку в обществе, но и сравнить письменную речь выбранных лиц, найти риторические инструменты и приёмы, которыми они пользуются.

Цель моей работы — профориентация. Возможность строчкой кода получить, а затем и обработать десятки тысяч (или миллионов) объектов сообщений подводят нас к таким областям ИКТ, как Data Science и прикладная компьютерная лингвистика, и дают мне возможность попробовать себя интерном в профессии лингвиста либо Data Scientist, то есть "учёный по «большим данным»". Это специальности, которые я рассматриваю как свою потенциальную профессию. Суть этой цели в том, чтобы, получив необходимые навыки и ознакомившись с профессиональным инструментарием, попробовать и решить, действительно ли мне подходят эти специальности, попутно решив прикладную задачу — ведь именно в этом суть профессий.

Задачи моей работы:

- Ознакомиться с похожими исследованиями, их методами анализа и результатами.
- Научиться пользоваться необходимым инструментарием и технологиями для проведения исследований такого рода.
- Определить 20 основных тем, которые волнуют российское общество (на начало марта 2020). Это возможно, поскольку темы, поднимаемые политиками, прямо отражают вопросы, которые волнуют социум. Эта задача относится к социологии.
  Социология — наука, изучающая общество, его структуру и взаимоотношения его членов, то есть людей.

[1] https://news.tut.by/world/643150.html
[2] https://www.rbc.ru/rbcfreenews/584eef239a7947658b57e99a
[3] https://www.kp.ru/daily/27112/4190029/

- Найти риторические инструменты, которыми пользуются выбранные политики. **Риторика** в общем смысле — наука об ораторском искусстве, то есть умении убеждать и побуждать людей с помощью красноречивой, выразительной, заранее подготовленной речи. Под **риторическими инструментами** имеются в виду характерные для политической речи слова, которые используют ораторы: побуждающие глаголы и выразительные прилагательные.
  Ораторские качества являются основополагающими для профессиональной деятельности политика. Анализ речи состоявшегося политика позволяет выделить значимые, сильные слова, которые он использует для убеждения аудитории. Поиск таких слов – вечная задача риторики с древнейших времён, которая никогда не может быть в полной мере решена, и потому всегда актуальна.

- Получить практику технического английского языка, так как взаимодействие с платформой осуществляется на английском языке, и на нём же написана документация к ней и к большинству необходимых программных инструментов.

Актуальность цели исследования обусловлена тем, что выбор профессии для меня чрезвычайно важен в данный момент — необходимо как можно скорее и точнее выбрать специальность и место, где я буду её получать, и это исследование если не поможет сделать выбор, то хотя бы уменьшит круг возможных направлений. Кроме того, профессия, которую я примеряю, очень востребована в сфере информационных технологий и на их стыке с самыми разными дисциплинами, и её практическая значимость лишь растёт вместе с объёмом данных, которые человечество начинает собирать и обрабатывать во всех областях жизнедеятельности. Помимо этого, в процессе проведения этого исследования решаются задачи, актуальность которых перманентна, покуда есть человечество, а равно общество и лидеры масс в нём. Большинство актуальных социальных тем постоянно меняются, и потому имеет смысл даже раз в несколько месяцев или лет проводить такое исследование снова и снова; риторика, будучи неточной, гуманитарной наукой, у одного человека реализуется не так, как у другого, и риторические инструменты меняются от человека к человеку и от десятилетия к десятилетию.

Глава 1. Объект исследования. Методика исследования.

Обзор литературы

Исследований такого рода мне практически не удалось найти, но научная статья “Корпусные исследования политического дискурса в лингвистике” [1] является наиболее близкой, и я принял во внимание методологию и подход, описываемый в ней.

В статье обсуждается возможность использования корпусных методов для исследования современной политической коммуникации. На материале двух микрокорпусов, включающих тексты предвыборных выступлений Хиллари Клинтон и Дональда Трампа, показаны возможности выделения и анализа наиболее частотной лексики и ключевых слов. Целью анализа является выявление риторических способов воздействия на аудиторию. Проще говоря, в этом исследовании была проведена почти аналогичная работа со схожей задачей и тоже с политиками, но для английского языка, который является более простым для обработки и частотного анализа, поскольку слова в нём не изменяются (по родам, падежам и т.д.). Были проведены как частотный анализ для обоих политиков, так и выделение ключевых слов. Недостатком этой работы я считаю выбор лишь двух политиков-соперников, что, впрочем, обусловлено тем, что Клинтон и Трамп были наиболее успешными и притом твёрдо противопоставленными оппонентами президентской гонки. Я же в моей работе предпочёл рассмотреть более широкий спектр разнородных политиков, что даёт посмотреть на разнообразные подходы к риторическим способам воздействия на аудиторию, к тому же на русском языке.

В работе “Основные коммуникативные тактики в публичных выступлениях женщин-политиков” [2] более подробно рассматривается политическая коммуникация и речевой портрет политика.

Объект исследования

В качестве объекта исследования был взят микрокорпус, составленный автоматически из 10 000 твитов за авторством 4 политиков. Для исследования были взяты известные политические деятели: двое из них придерживаются взглядов, поддерживающих действующую власть (В. Соловьёв, В. Жириновский), а двое других – оппозиционеры (М. Светов, А. Навальный). Такой выбор позволяет как сравнить темы, о которых говорят по разную сторону политических баррикад, так и уменьшить влияние речевых особенностей конкретного человека на общий анализ. Также доподлинно известно, что выбранные люди ведут свои микроблоги самостоятельно, и эти высказывания – их прямая речь, а не составленная для них редакторами. Всего выборка содержит около 5-6 тысяч слов на каждого человека.

Получение доступа

Для сбора данных необходимо получить доступ к API[4] Twitter. У этой компании довольно строгая политика предоставления доступа к своим данным и возможностям, что, очевидно, обусловлено заботой о конфиденциальности пользовательских данных во избежание скандалов, участившихся в последнее десятилетие. Для получения доступа к возможностям разработчика необходимо пройти индивидуальное собеседование в несколько этапов на английском языке. На первом этапе, происходящем на сайте в специальной форме, необходимо выбрать тип своей деятельности и расписать целое эссе о своих намерениях по каждому из интересующих компанию пунктов: рассказать, в чём принцип работы приложения-обработчика, объяснить, что у него некоммерческая направленность, рассказать, какие данные будут использоваться и описать методы их анализа, рассказать, в каком виде и где будут показываться пользовательские данные вне “Твиттера”. После того, как отправленная форма будет одобрена сотрудником компании, на email запрашивающего отправляется письмо, где требуется уточнить информацию или ответить на дополнительные вопросы. Если ответ на это письмо устроит проверяющего, запрашивающий получит доступ к панели разработчика в своём аккаунте Twitter. После этого необходимо сгенерировать ключ доступа для конкретного приложения, где повторно вкратце описать, что оно из себя представляет. Этот ключ и обеспечивает доступ к данным.

Мой ответ на второй этап собеседования и письмо, подтверждающее его прохождение, приведены в приложении А.

[4] API, Application Programming Interface — “программный интерфейс приложения”, совокупность протоколов взаимодействия между приложениями, в данном случае — набор документированных HTTP-запросов к веб-приложению, которое возвращает HTTP-ответы, содержащие данные в формате JSON.

Получение и подготовка данных

Я разработал программное решение, которое осуществляет составление запроса к API, получение данных, сохранение их в базу данных и дальнейшую обработку вплоть до получения готового к интерпретации результата. В качестве предварительной обработки было выполнено разбиение текстов на слова, а также очищение исходного материала от “мусора”, специфичного для микроблога: упоминание других пользователей в формате @user, ссылки, хэштеги; кроме того, были исключены все некириллические словоформы, выполнено приведение всех слов в нижний регистр, избавление от знаков препинания и цифр. Из промежуточного микрокорпуса также исключались т.н. стоп-слова, то есть служебные части речи, дейктические единицы и другие слова, не имеющие значения для анализа (исключение было сделано для местоимений “я” и “мы”). Такая обработка искореняет главные недостатки частотного анализа — преобладание в верхней части итогового списка “ненужных” слов. Затем все слова были приведены в исходную, т.е. словарную форму с помощью морфологического анализатора pymorphy2 в составе моего программного решения. После этого материал пригоден для дальнейшего анализа. Основным методом обработки данных в моей работе является частотный анализ.

Частотный анализ, анализ частоты употребления слов — это способ выявления коммуникативно значимых языковых единиц внутри одного корпуса. Для всего корпуса производится подсчет встречаемости каждого слова, после чего составляется новый корпус, своеобразный рейтинг частоты употребления слова.

Также для удобства интерпретации для готового рейтинга было проведено разделение общего списка на списки определённых частей речи. Для исследования были выбраны такие части речи, как существительное (для социологии), прилагательное и глагол (для риторических изысканий). Разделение было выполнено тоже с использованием морфологического анализатора.

Побочным эффектом использования морфологического анализатора для приведения в исходную форму и разбиения по частям речи стало появление в выборке неточных данных: так, некоторые слова были насильно приведены в исходную форму, даже если в том не нуждались или обычно используются во множественном числе (например, “деньги”, приведённые к “деньга”). Кроме того, некоторые слова были отнесены не к тем частям речи: для существительных и глаголов микрокорпус почти идеален, а к прилагательным попали лишние слова, например фамилии. Это обусловлено тем, что анализатор, не найдя слова в словаре, строит предсказательную модель, которая делает предположение, что раз “зеленский” обладает

свойствами прилагательного (окончание -ий, например), то, скорее всего, им и является. Тем не менее, доля ошибок незначительна, а ошибки явны, и потому их можно просто игнорировать на этапе интерпретации. Однако следует отметить, что некоторые “странные” словоформы из итогового корпуса являются жаргонизмами (например, “тви” у В. Соловьёва — это “Твиттер”).

Основным материалом для дальнейшей интерпретации является полученный в результате всех описанных выше действий файл words.xls, представляющий собой таблицу Excel с четырьмя листами, подписанными фамилиями политиков, где на каждом листе представлены подписанные столбцы как всех слов, так и отдельно исследуемых частей речи для этого политика. Получить файл можно в репозитории (приложение В).

Глава 2. Результаты исследования

Основные темы в речи

В первую очередь было обращено внимание на то, какие существительные занимают верхние позиции корпуса, поскольку, как было сказано выше, предполагается, что именно они являются тем, что волнует российское общество на момент исследования. В итоге наиболее встречаемыми существительными, не считая некоторых бытовых, таких как “день”, оказались те, что приведены в таблице 1. Следует отметить, что словам с одинаковой встречаемостью было присвоено одинаковое место в получившимся “рейтинге”, то есть если слово “аниме” встречается у Жириновского лишь раз, то оно занимает 70 место, хоть и находится на 3524 строчке итоговой таблицы, деля это место с ещё 1926 другими словами, тоже встречающимися 1 раз (далее мы увидим, насколько значимость места падает с каждой позицией).

В целом можно заметить, что наибольшее внимание всех сторон привлекают война и мир, закон и Москва; у оппозиционных политиков с большим отрывом на первом месте Россия, и их вовсе не беспокоят Украина и США. Также оппозиционерам свойственно упоминание власти, митингов и денег. У их оппонентов оказалось неожиданно мало общих частых тем, кроме уже упомянутых отметить особо нечего, зато бывают пересечения с оппозиционерами: например, Владимир Соловьёв и Михаил Светов оба очень часто упоминают историю, а остальные — очень редко. Права и свободы волнуют только представителя Либертарианской партии России Светова, коррупция, пенсии и депутаты — антикоррупционера Алексея Навального. Именно эти слова хочет слышать от них аудитория, именно это волнует людей, которые за ними следуют.

Таблица 1. 20 наиболее важных существительных

Слово	Место среди существительных			
	Соловьёв	Жириновский	Навальный	Светов
Россия	2	4	**1**	**1**
Путин	**1**	62	**2**	13
народ	21	26	34	19
Москва	6	11	8	8
Украина	**8**	**10**	34	39

США	**10**	**19**	30	36
депутат	42	20	**7**	42
война	27	29	27	27
митинг	39	21	**15**	**3**
мир	21	16	30	18
страна	16	**8**	18	**4**
власть	30	28	**17**	**10**
право	28	44	29	**5**
свобода	41	62	28	**9**
история	**5**	46	26	**7**
деньги	33	37	**9**	**14**
суд	**13**	62	**17**	23
коррупция	45	55	**18**	46
закон	23	22	29	**16**
пенсия	38	62	**27**	отсутствует

Наблюдение о распределении значимых слов в корпусе

Посмотрев на то, как уменьшается встречаемость каждого следующего места в корпусе, можно заметить, что между первыми строчками чаще всего значимый разрыв по встречаемости, зато по мере движения в сторону уменьшения встречаемости этот разрыв переходит на единицы, а слова, занимающие одно место, начинают исчисляться сотнями и тысячами. Для того, чтобы определить, является ли это общей закономерностью, а также из интереса, какую долю от частотности всего корпуса занимают 10% самых встречаемых слов, я программно построил графики падения встречаемости словоформ для общего корпуса (все слова без разбиения на части речи) каждого политика. График для Михаила Светова представлен на рис. 1, все графики приведены в приложении Г, данные и исходный код программы — в приложении В.

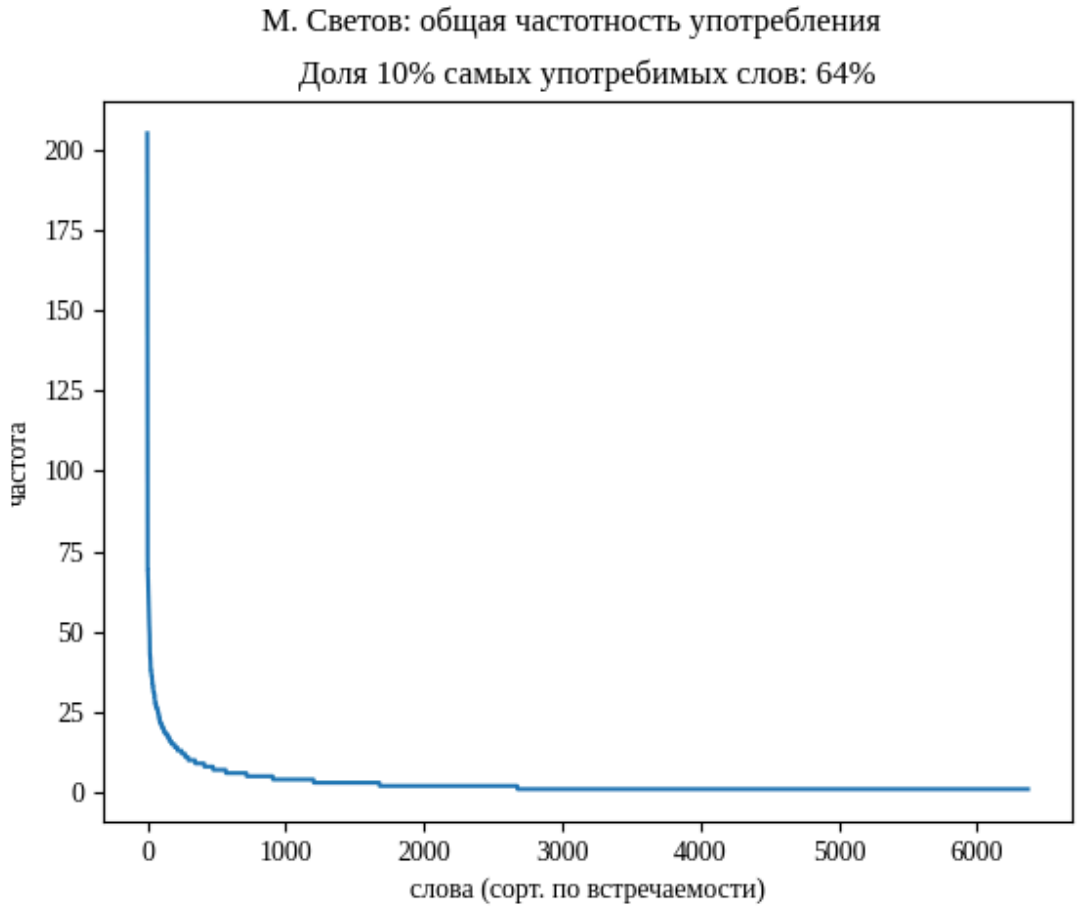


рис. 1

Оказалось, что это действительно закономерность, и она работает для всех выбранных лиц и для любой части речи, строя схожий график. При этом доля встречаемости 10% первых слов сортированного корпуса по отношению к встречаемости всех слов превышает 60%, а у Владимира Жириновского достигает 70%, то есть в первых 10% фактически заключены все значимые, характеризующие речь и человека слова, а остальные 90% —

вспомогательные, эпизодические, не несущие особого смысла в масштабах такого исследования. На практике это означает, что для подобной исследовательской работы реальной интерпретации подлежит лишь относительно небольшая часть слов, а остальные следует отбросить, беря во внимание лишь при сравнениях встречаемости слова у разных людей — у кого оно входит в 10% и является значимым, а у кого нет. Табл. 1 и весь остальной анализ в этой работе были проведены с учётом полученных практических сведений.

Риторические инструменты политиков

Рассмотрев, о чём говорят избранные нами лидеры мнений, можно подойти к вопросу, *как* они это говорят.

Хотя риторика в своей основе — искусство говорить, доступность и удобство текстового представления информации ведёт к тому, что люди начинают больше писать, обращаясь друг к другу или к широкой аудитории в Интернете. Один из основополагающих факторов риторики — голос, его интонация и передаваемое через него усилие воли, вызывающие нужный психологический настрой аудитории — в данном случае не работает, поэтому остаётся полагаться лишь на языковые конструкции, выразительные и цепляющие прилагательные, а также “сильные”, то есть убеждающие и побуждающие к действию глаголы. Именно поэтому я рассматриваю эти части речи у выбранных политиков, причём в данном случае мы не будем сопоставлять их между собой, находя некие “правильные” слова (универсальных решений не существует), но рассмотрим собственный ораторский набор каждого из них.

Владимир Соловьёв

Таблица 2

прилагательные	глаголы
приятный	работать
должный	заявить
российский / русский	очнуться
тупой	лгать
глупый	предложить
свежий	врать
способный	пытаться

интересный	тупить
трусливый	судить
великий	призывать

У Владимира Соловьёва в наблюдается неожиданное соседство слов, очень разных по настроению: например, приятный и тупой (таблица 2). Такая ситуация — уникальная для этой выборки — возникает из-за того, что говорящий, говоря нейтрально или положительно о том, что ему нравится, при столкновении с другим мнением предпочитает не приводить каких-либо качественных аргументов, а сокрушить оппонента морально, унизить, оскорбить. С этим же связан тот факт, что Владимир Рудольфович — единственный в выборке, у кого среди частых существительных встречаются слова *лжец, ложь, чушь, враньё, тварь, дурак, мразь, идиот, ненависть*, в то время, как у всех остальных эти слова не встречаются вообще или лишь один раз (не входя в значимые 10%). Это агрессивный, но рабочий подход, заслуживающий внимания не меньше, чем более консервативные и “спокойные” подходы других ораторов. Кроме того, представляет интерес очень частый призыв *очнуться*, являющийся особенностью речи В. Соловьёва, его регулярным риторическим инструментом.

Владимир Жириновский

Таблица 3

прилагательные	глаголы
русский	поздравлять
новый	обсудить
нужный	поддержать
хороший	предлагать
важный	требовать
большой	выступать
родной	голосовать
искренний	запретить
главный	убрать

великий	отменить

Ораторский набор Владимира Жириновского является наиболее ярким и образцовым (таблица 3). Политику свойственны “положительные” прилагательные, создающие у аудитории приятные, светлые ощущения и ассоциации, а также патриотические чувства: *русский*, *родной, великий*. Среди глаголов видны прямые побуждения к действию, после каждого хочется поставить восклицательный знак. Создаётся образ активного человека, который знает, что и как нужно делать: требует, поддерживает, предлагает что-то, призывает убрать и запретить. Поскольку среди изучаемых политиков В. Жириновский добился наибольших успехов в карьере, является лидером крупной партии и регулярным участником президентской гонки, можно считать, что его ораторский набор и чёткий подход к риторическому воздействию являются наиболее достоверными в качестве образца политической риторики.

Алексей Навальный

Таблица 4

прилагательные	глаголы
новый	требовать
единый	думать
путинский	работать
большой	получить
хороший	купить
важный	устроить
отличный	запретить
прекрасный	врать
сегодняшний	украсть
иностранный	закрыть

Набор Алексея Навального немного похож на набор Владимира Жириновского: в целом используются схожие прилагательные, есть общие глаголы. Отличительной особенностью является частое употребление прилагательного “путинский”, обозначающего явление, которое связано с периодом президентства Владимира Путина и Дмитрия Медведева с 2000 г. по сей день и является негативным результатом их политики, по мнению

говорящего. Чаще всего это прилагательное используется в словосочетании “*путинская* Россия”, несущем негативную окраску. Оно является отличительной особенностью речи оппозиционеров, в нём заключается сама суть их оппозиции — выставление действующей власти в негативном свете и противопоставление ей “*прекрасной* России будущего” — как буквальной России будущего, представляющей собой абстрактный утопический образ, так и одноимённой партии, председателем которой Алексей является.

Михаил Светов

Таблица 5

прилагательные	глаголы
путинский	понимать
государственный	любить
огромный	рассказывать
политический	выйти
важный	защищать
частный	организовать
смешной	запретить
умный	обязать
левый	победить
страшный	спасти

Михаил Светов, тоже будучи оппозиционером, имеет прилагательное “путинский” среди наиболее встречаемых. Особенностью его набора риторических инструментов являются слова *защищать* и *спасти*. Часто встречается прилагательное *частный*, т.е. частная собственность или частная жизнь. То, что он осуждает, он чаще всего называет *смешным*, если относится к этому с иронией, или *страшным*, если считает крайне недопустимым. Часто призывает *выйти* — на митинг, из зоны комфорта и т.д.

Итоги исследования

Я добился поставленной цели — прошёл весь путь типичного задания для инженера IT-направления Data Science, от получения данных и постановки задачи до обработки и анализа, и в целом пришёл к выводу, что такая работа мне интересна и по силам, и я буду выбирать учебное заведение, где смогу овладеть ею на более глубоком уровне.

Были успешно выполнены поставленные задачи:

- Я изучил существующие исследования на похожие темы, изучил и применил методы и подходы, которые в них используются, нашёл и исправил то, что мне показалось недостатком подхода;
- Овладел такими инструментами для работы с данными, как pandas и matplotlib, улучшил свои навыки программирования на Python и Ruby и работы с git;
- Были найдены темы, волнующие российское общество: это внешнеполитические отношения, США, Украина, война и мир, новые законы, коррупция, история, пенсии, и, разумеется, сама Россия и её президент;
- Были исследованы способы риторического воздействия российских политиков, их особенности;
- Я получил бесценный опыт практического применения технического английского языка, усовершенствовал навыки владения им;

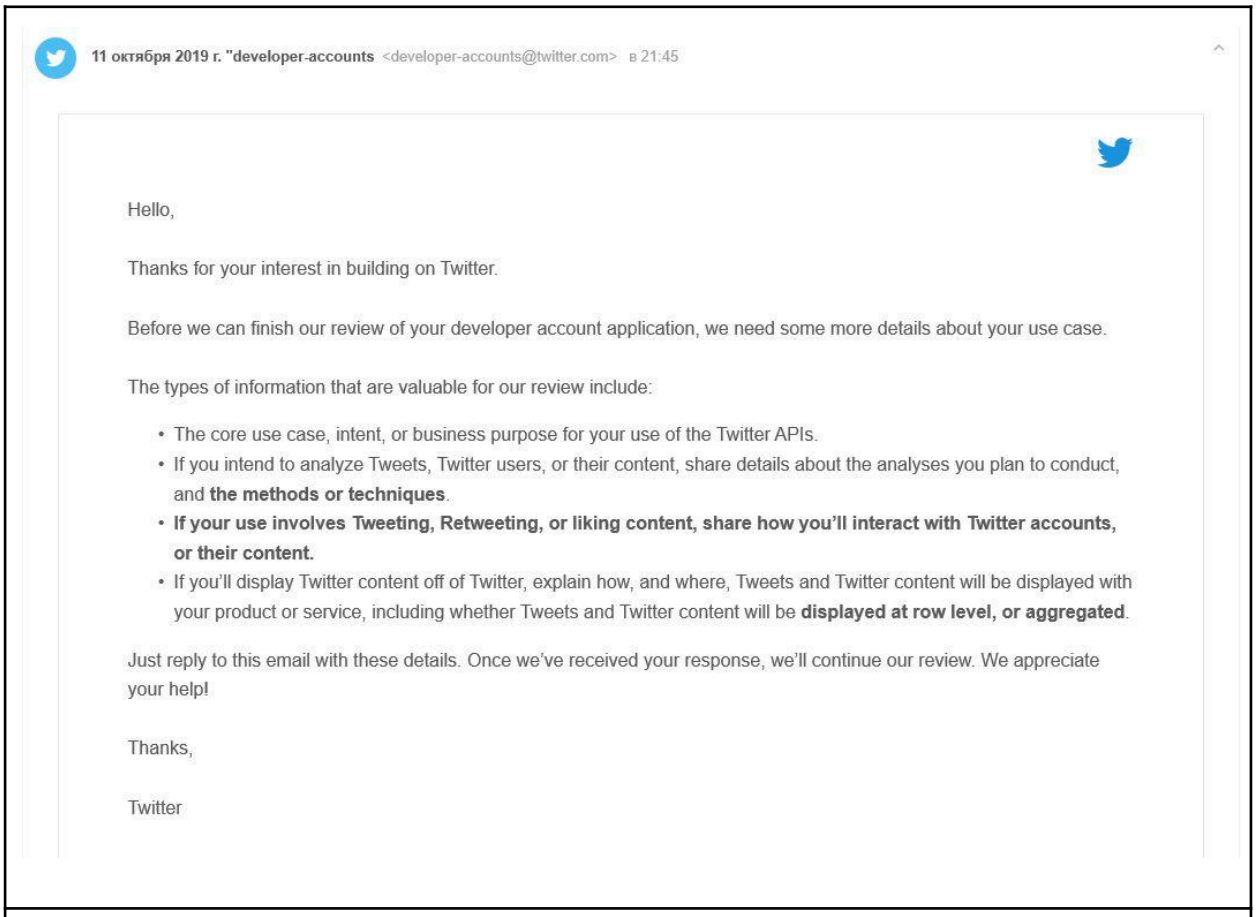
Список использованной литературы

1. Борискина Ольга Олеговна, Шилихина Ксения Михайловна Корпусные исследования политического дискурса в лингвистике // Полит. наука. 2017. №2.
2. Сахарова Ольга Сергеевна, Федорова Ирина Александровна Основные коммуникативные тактики в публичных выступлениях женщин-политиков (на примере предвыборного выступления Хиллари Клинтон) // Вестник ТГПУ. 2017. №6 (183).
3. Документация — Морфологический анализатор pymorphy2. [Электронный ресурс] URL: https://pymorphy2.readthedocs.io/en/latest/user/index.html (дата обращения 24.04.20)
4. Matplotlib User's Guide. [Электронный ресурс] URL: https://matplotlib.org/contents.html (дата обращения 24.04.20)
5. Pandas Documentation. [Электронный ресурс] URL: https://pandas.pydata.org/docs/ (дата обращения 24.04.20)
6. Python 3.7.7 Documentation. [Электронный ресурс] URL: https://docs.python.org/3.7/ (дата обращения 24.04.20)
7. Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, pp 320-332 (2015)
8. Twitter Developer API reference index. [Электронный ресурс] URL: https://developer.twitter.com/en/docs/api-reference-index (дата обращения 24.04.20)
9. The Twitter Ruby Gem Documentation. [Электронный ресурс] URL: https://www.rubydoc.info/gems/twitter (дата обращения 24.04.20)

Приложения

Приложение А. Письма с собеседования

Первая и самая трудная и объёмная часть собеседования, которая производилась на сайте, не сохранилась, но вторая часть, с уточнениями по email, осталась в виде переписки и приведена здесь.



11 октября 2019 г. "developer-accounts <developer-accounts@twitter.com> в 21:45

Hello,

Thanks for your interest in building on Twitter.

Before we can finish our review of your developer account application, we need some more details about your use case.

The types of information that are valuable for our review include:

- The core use case, intent, or business purpose for your use of the Twitter APIs.
- If you intend to analyze Tweets, Twitter users, or their content, share details about the analyses you plan to conduct, and **the methods or techniques**.
- **If your use involves Tweeting, Retweeting, or liking content, share how you'll interact with Twitter accounts, or their content.**
- If you'll display Twitter content off of Twitter, explain how, and where, Tweets and Twitter content will be displayed with your product or service, including whether Tweets and Twitter content will be **displayed at row level, or aggregated**.

Just reply to this email with these details. Once we've received your response, we'll continue our review. We appreciate your help!

Thanks,

Twitter

рис.1. Запрос на уточнение предоставляемой информации

Re: Twitter developer account application [ref:_00DA0K0A8._5004A1n53lB:ref]

BI **bikeroleg@yandex.ru** bikeroleg@yandex.ru 11.10.19 в 22:1

1 получатель

Язык письма — английский. Перевести на русский? Перевести

1. No business purposes, I need it for studying now and for possible hobby projects ahead.

2. I will analyze tweets by finding the dependence of keywords in them and the amount of likes/retweets/replies. Users' profiles or any of their content except raw text will not be analyzed.

3. I will use tweeting but won't directly interact with users. I don't plan to like or retweet something with the API yet, although I would need these functions ahead to check if this changes the results of the research. That includes liking/retweeting just 3-5 manually selected tweets.

4. No content will be shown off Twitter, neither tweets nor any other content, statistics only.

рис.2. Ответ на запрос

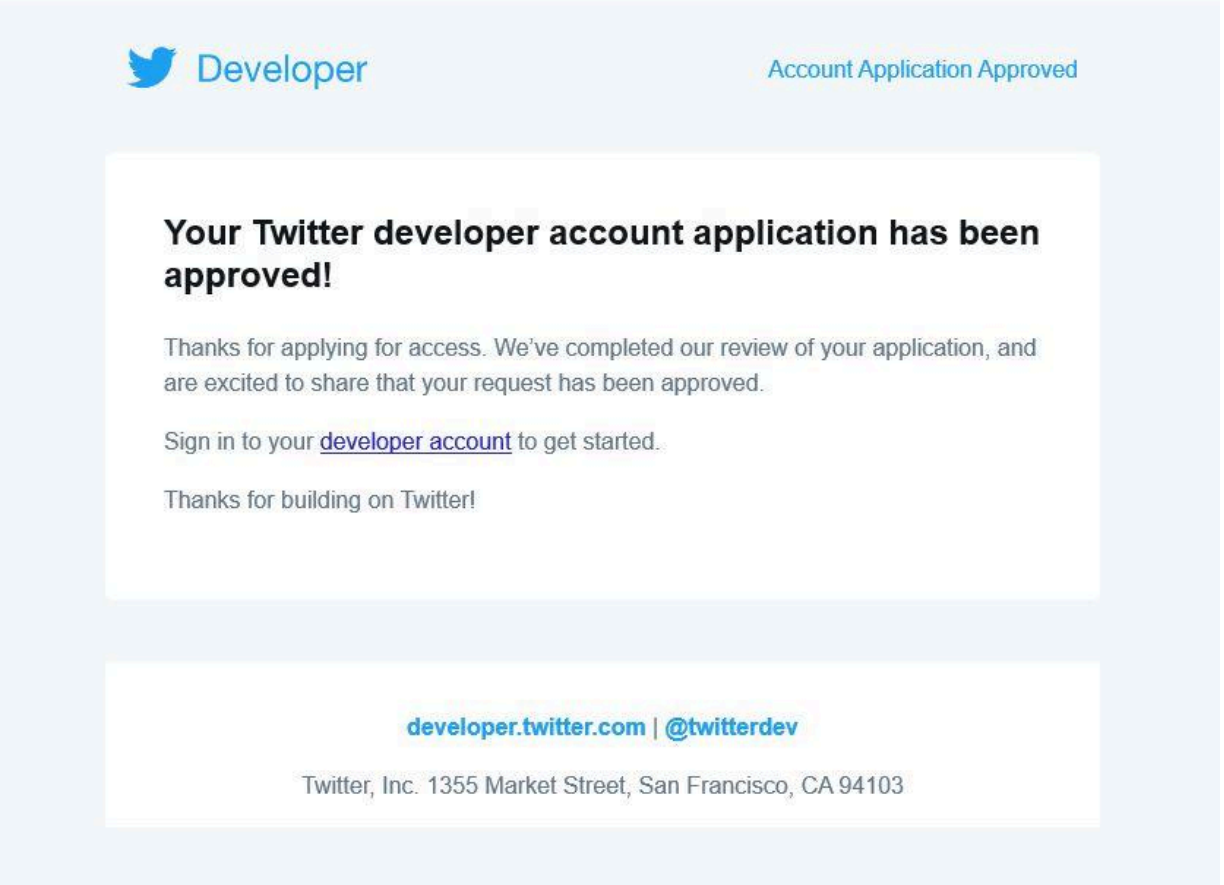


рис. 3. Подтверждение успешного прохождения собеседования

Приложение Б. Список стоп-слов

и	было	там	кто	два	нельзя
в	вот	потом	этот	об	такой
во	от	себя	это	другой	им
не	меня	ничего	говорил	хоть	более
что	еще	ей	того	после	всегда
он	нет	может	потому	над	конечно
на	о	они	этого	больше	всю
с	из	тут	какой	тот	между
со	ему	где	совсем	через	весь
как	теперь	есть	ним	эти	который
а	когда	надо	здесь	нас	год
то	даже	ней	этом	про	очень
все	ну	для	один	всего	самый
она	вдруг	тебя	почти	них	ещё
так	ли	их	мой	какая	вообще
его	если	чем	тем	много	сколько
но	уже	была	чтобы	разве	
да	или	сам	нее	сказала	
ты	ни	чтоб	кажется	три	
к	быть	без	сейчас	эту	
у	был	будто	были	моя	
же	него	человек	куда	впрочем	
вы	до	чего	зачем	хорошо	
за	вас	раз	сказать	свою	
бы	нибудь	тоже	всех	этой	
по	опять	себе	никогда	перед	
только	уж	под	сегодня	иногда	
ее	вам	будет	можно	лучше	
мне	сказал	ж	при	чуть	
	ведь	тогда	наконец	том	

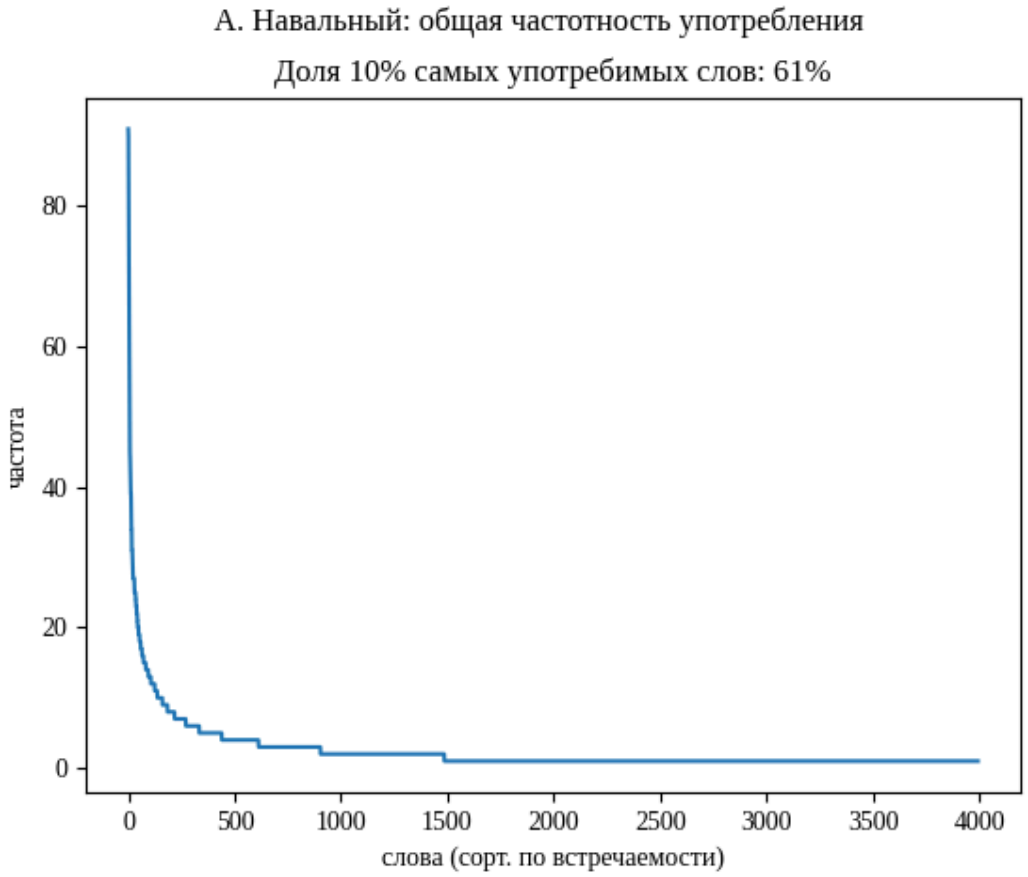
Приложение В. Программный репозиторий

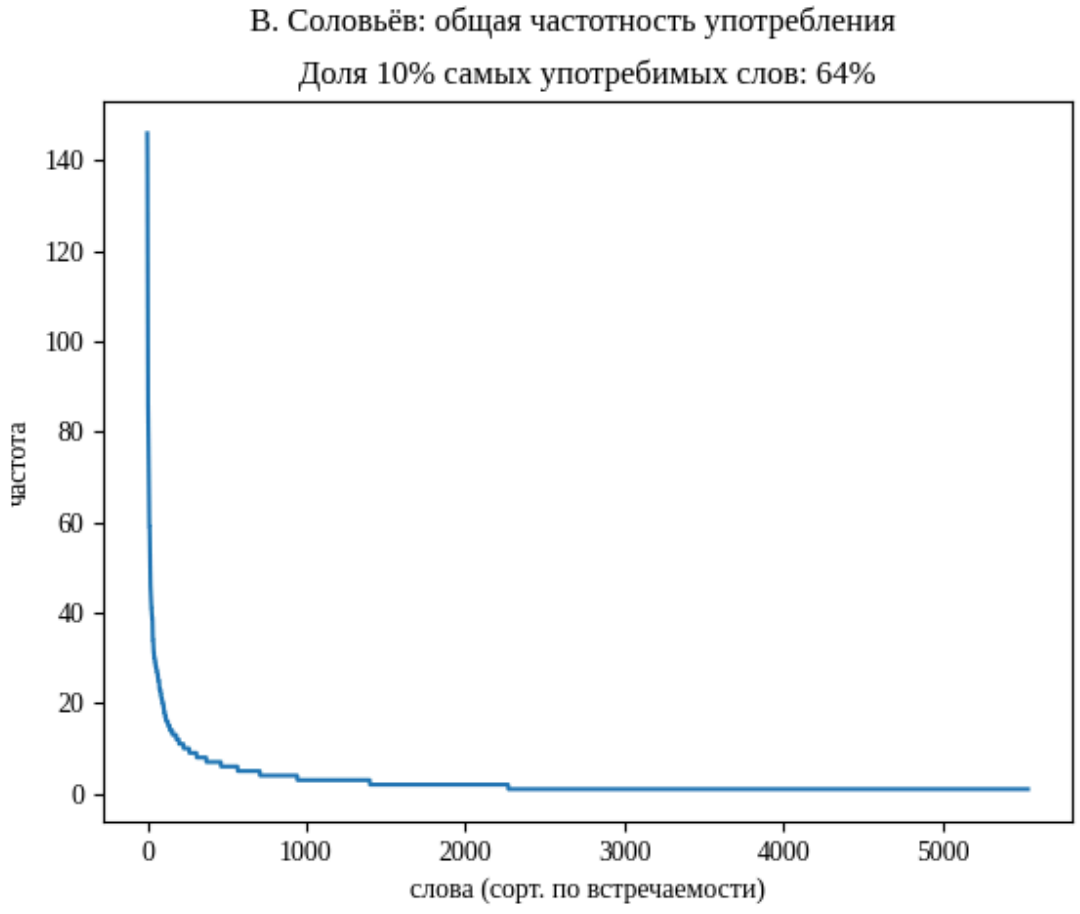
Исходный код программного решения, а также база данных с использовавшимися твитами и готовые результаты в формате CSV, TXT и XLS находятся в открытом доступе на GitHub-репозитории по следующему адресу: https://github.com/bikeroleg/iproject

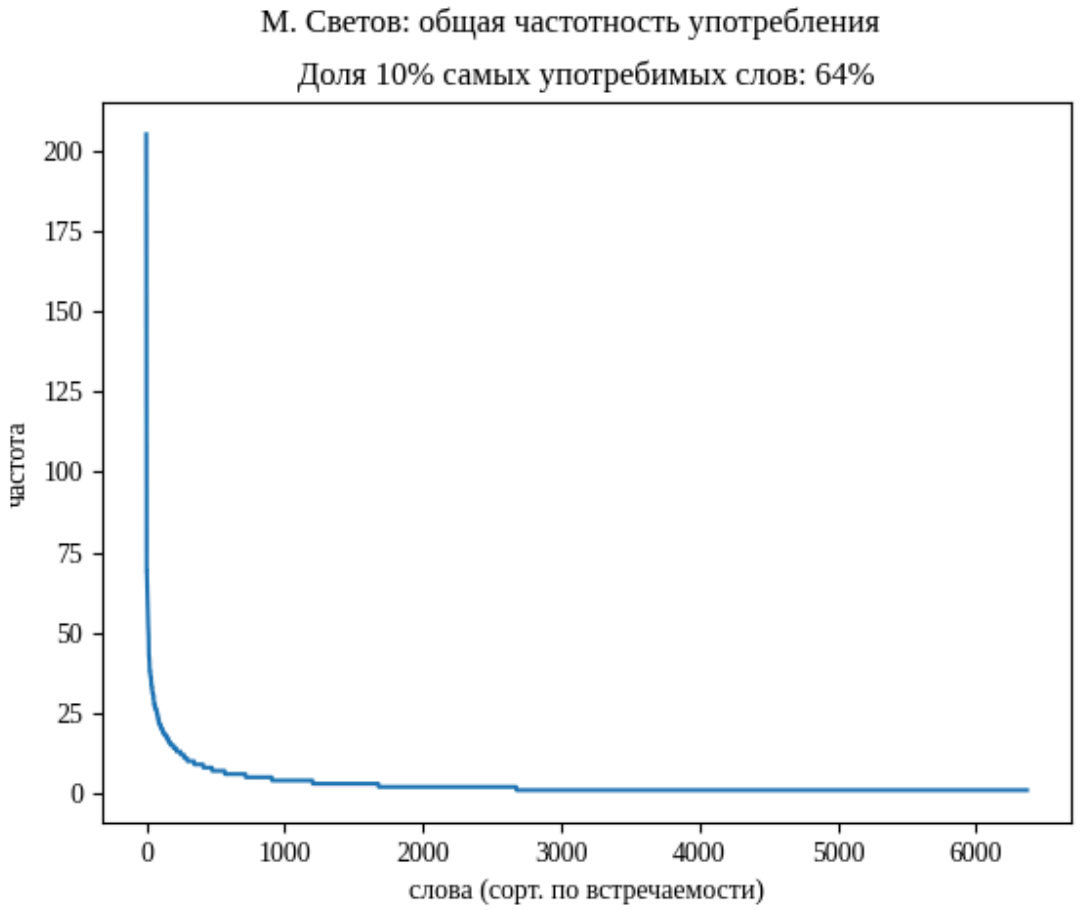
Краткая навигация по репозиторию:

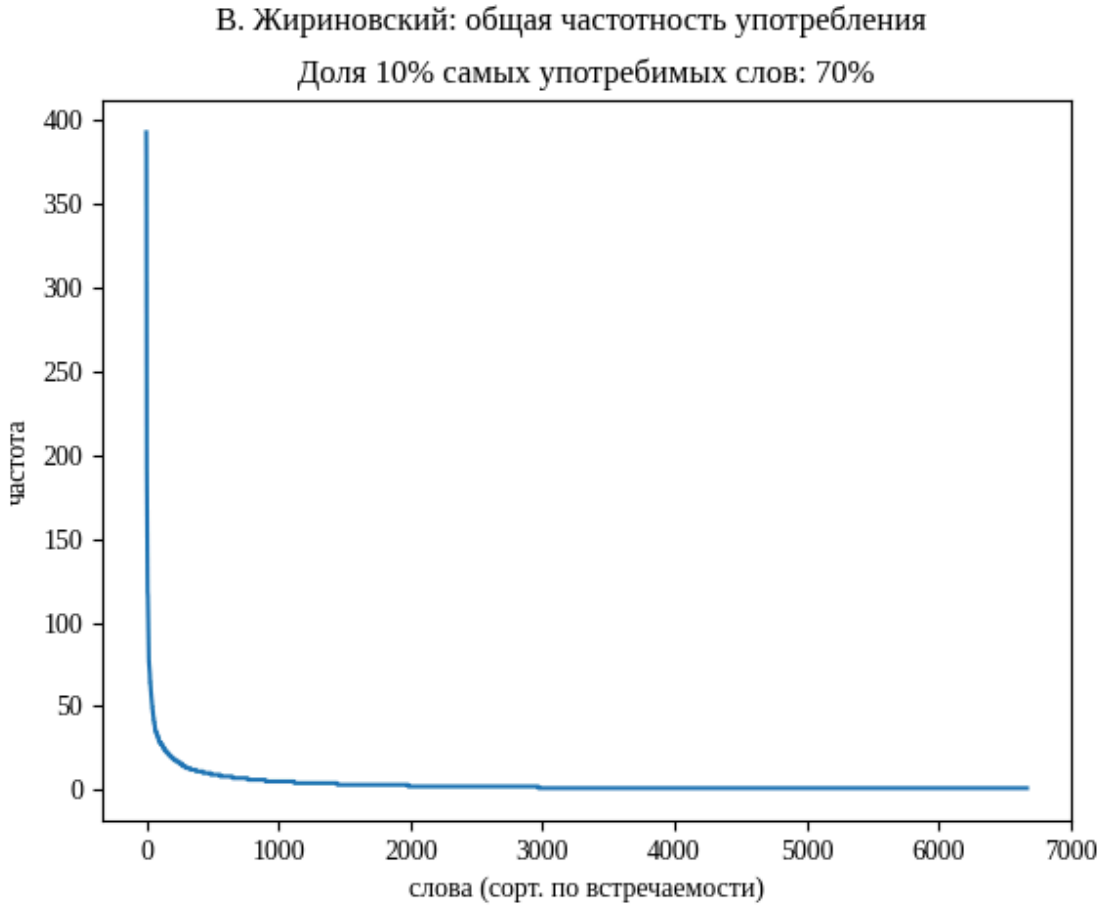
- words.xls — книга Excel с итоговыми данными, представленными в более удобном виде. Для того, чтобы просто ознакомиться с данными исследования, достаточно посмотреть лишь её.
- tweets.db — БД формата SQLite, содержащая исходные тексты твитов, псевдонимы их авторов и ссылки на эти твиты.
- iproject.rb — основная программа
- database.rb — модуль, отвечающий за работу с БД
- sklonyator/sklonyator.py — программа, разбивающая тексты в БД на слова и приводящая их к исходной форме
- sklonyator/sort.py — сортировщик по частям речи
- text2csv.rb — программа для преобразования данных, оформленных в виде текстового файла, в файл формата csv, использующегося для хранения данных в научных исследованиях.
- navalny/, zhirinovsky/, svetov/, soloviev/ — директории, где собраны данные, относящиеся к каждому человеку
- директория **iproject_graphs** содержит набор для построения графиков:
    - graphs.py — скрипт python для построения графиков (для общей массы слов). При запуске создаёт окна с графиками, где с ними можно взаимодействовать, приближать, перемещать и сохранять.
    - others/other_graphs.py — скрипт для построения графиков для отдельных частей речи; эти графики не были приведены в этом исследовании.

Приложение Г. Графики






М. Светов: общая частотность употребления
Доля 10% самых употребимых слов: 64%
частота
200
175
150
125
100
75
50
25
0
0
1000
2000
3000
4000
5000
6000
слова (сорт. по встречаемости)


В. Жириновский: общая частотность употребления
Доля 10% самых употребимых слов: 70%
частота
400
350
300
250
200
150
100
50
0
0
1000
2000
3000
4000
5000
6000
7000
слова (сорт. по встречаемости)