МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра дискретной математики и алгоритмики

ГУЛИН
Кирилл Иванович

ИССЛЕДОВАНИЕ МЕТОДОВ ВЫДЕЛЕНИЯ СМЫСЛОВЫХ ЕДИНИЦ ИЗ ДЕЛОВОЙ ПЕРЕПИСКИ

Дипломная работа

Научный руководитель:
Ю. В. Свирид

Допущен к защите
«_____» _____ 2021 г.
Зав. кафедрой дискретной математики
и алгоритмики, доктор физ.-мат. наук,
профессор В. М. Котов

Минск, 2021

Содержание

РЕФЕРАТ

Дипломная работа, 33 с., 23 рис., 12 источников.

Ключевые слова: ОБРАБОТКА ТЕКСТОВ, МАШИННОЕ ОБУЧЕНИЕ, ЛАТЕНТНОЕ РАЗМЕЩЕНИЕ ДИРИХЛЕ, МЕТОДЫ ПОНИЖЕНИЯ РАЗМЕРНОСТИ, ВЕКТОРНОЕ ПРЕДСТАВЛЕНИЕ ТЕКСТОВ, СТАТИСТИЧЕСКИЙ АНАЛИЗ, МЕТОДЫ КЛАСТЕРИЗАЦИИ, НЕЙРОННЫЕ СЕТИ.

Объект исследования — наборы данных деловых электронных переписок.

Цель работы — анализ деловых электронных переписок методами машинного обучения.

Методы исследования — латентное размещение Дирихле, методы кластеризации, методы обработки текстов, методы понижения размерности, методы получения векторных представлений.

Работа посвящена исследованию и анализу деловых электронных переписок, в частности, переписок Хиллари Клинтон и переписок сотрудников корпорации Enron. В результате работы был произведен статистический анализ электронных переписок. Были обнаружены закономерности в исходных данных. Также была разработана кластеризация содержаний электронных писем, в результате которой получились интерпретируемые результаты, что показало эффективность разработанных методов.

РЭФЕРАТ

Дыпломная праца, 33 с., 23 рыс., 12 крыніц.

Ключавыя словы: АПРАЦОЎКА ТЭКСТАЎ, МАШЫННАЕ НАВУЧАННЕ, ЛАТЭНТНАЕ РАЗМЯШЧЭННЕ ДЫРЫХЛЕ, МЕТАДЫ ЗНІЖЭННЯ ПАМЕРНАСЦІ, ВЕКТАРНАЕ ПРАДСТАЎЛЕННЕ ТЭКСТАЎ, СТАТЫСТЫЧНЫ АНАЛІЗ, МЕТАДЫ КЛАСТАРЫЗАЦЫІ, НЕЙРОНАВЫЯ СЕТКІ.

Аб'ект даследавання — наборы дадзеных дзелавых электронных перапісак.

Мэта работы — аналіз дзелавых электронных перапісак метадамі машыннага навучання.

Метады даследавання — латэнтнае размяшчэнне Дырыхле, метады кластарызацыі, метады апрацоўкі тэкстаў, метады зніжэння памернасці, метады атрымання вектарных уяўленняў.

Праца прысвечана даследаванню і аналізу дзелавых электронных перапісак, у прыватнасці, перапісак Хілары Клінтан і перапісак супрацоўнікаў карпарацыі Enron. У выніку працы быў выраблены статыстычны аналіз электронных перапісак. Былі выяўлены заканамернасці ў зыходных дадзеных. Таксама была распрацавана кластарызацыя зместаў электронных перапісак, у выніку якой атрымаліся інтэрпрэтаваныя вынікі, што паказала эфектыўнасць распрацаваных метадаў.

ABSTRACT

Diploma thesis, 33 p., 18 fig., 12 sources.

Keywords: TEXT PROCESSING, MACHINE LEARNING, LATENT DIRICHLET ALLOCATION, DIMENSION REDUCTION METHODS, TEXT EMBEDDINGS, STATISTICAL ANALYSIS, CLUSTERIZATION METHODS, NEURAL NETWORKS.

The object of research is business e-mail datasets.

Objective: analysis of business e-mails using machine learning methods.

Research methods — latent Dirichlet allocation, clustering methods, text processing methods, dimension reduction methods, methods for obtaining text embeddings.

The work is devoted to the research and analysis of business e-mails, in particular, the e-mails of Hillary Clinton and the e-mails of employees of the Enron corporation. As a result of the work, a statistical analysis of emails was carried out. Patterns were found in the original data. Also, the clustering of the contents of e-mails was developed, as a result of which interpretable results were obtained, which showed the effectiveness of the developed methods.

Введение

С ростом доступности электронных документов и быстрым ростом всемирной паутины задача автоматической категоризации документов стала ключевым способом классификации и группирования информации и знаний любого рода. Для правильной классификации электронных документов, онлайн-новостей, блогов, электронной почты и электронных библиотек необходимы интеллектуальный анализ текста (англ. *Text Mining*), машинное обучение (англ. *Machine Learning*) и методы обработки текстов на естественном языке (англ. *Natural Language Processing, NLP*).

Современные системы обработки текстов на естественном языке могут анализировать неограниченные объемы текстовых данных. Они могут понимать суть сложных контекстов, расшифровывать двусмысленности языка, извлекать ключевые факты и взаимосвязи. Учитывая огромное количество неструктурированных данных, которые создаются каждый день, от электронных медицинских карт до сообщений в социальных сетях, обработка текстов на естественных языках стала критически важной для эффективного анализа текстовых данных.

Глава 1
Используемые данные

1.1 Электронные письма Хиллари Клинтон

В 2015 году Хиллари Клинтон (американский политик, государственный секретарь США в 2009-2013 гг., кандидат в президенты США в 2016 г.) была вовлечена в большое количество споров по поводу использования личных учетных записей электронной почты на негосударственных серверах во время ее пребывания на посту государственного секретаря США. Некоторые политические эксперты утверждают, что использование Клинтон личных учетных записей электронной почты для ведения дел госсекретаря является нарушением протоколов и федеральных законов, обеспечивающих надлежащий учет деятельности правительства.

Был подан ряд исков о свободе информации из-за того, что Государственный департамент США не опубликовал полностью электронные письма, отправленные и полученные на личные аккаунты Клинтон. На сегодняшний день Государственным департаментом США опубликовано почти 7000 страниц отредактированных электронных писем Клинтон.

Документы были опубликованы в формате PDF. Платформа *Kaggle* очистила и нормализовала выпущенные документы и разместила их для публичного анализа. Мы будем основываться именно на датасете, опубликованном *Kaggle*.

1.2 Электронные письма корпорации Enron

Enron являлась государственной корпорацией штата Орегон со штабквартирой в городе Хьюстон. До объявления о банкротстве в декабре 2001 года Enron была седьмой по величине корпорацией в США. В феврале 2002 года Федеральная комиссия по регулированию в области энергетики (Federal Energy Regulatory Commission, FERC) начала всестороннее расследование торговой деятельности Enron на рынках электроэнергии Калифорнии. Согласно FERC, Enron получила некоторую информацию о рынке, недоступную для ее конкурентов. Прибыль Enron превысила 500 миллионов долларов в 2000 и 2001 годах. Расследование пришло к выводу, что многие торговые стратегии, используемые Enron, нарушают рыночные отношения, утвержденные Федеральной комисиией для Калифорнии. С июня 2002 года Министерство юстиции США возбудило уголовные дела против 30 человек, включая Джеффри Скиллинга, бывшего президента и генерального директора Enron, а также других руководителей высшего звена. Обвинения включали сговор, мошенничество с ценными бумагами и инсайдерскую торговлю.

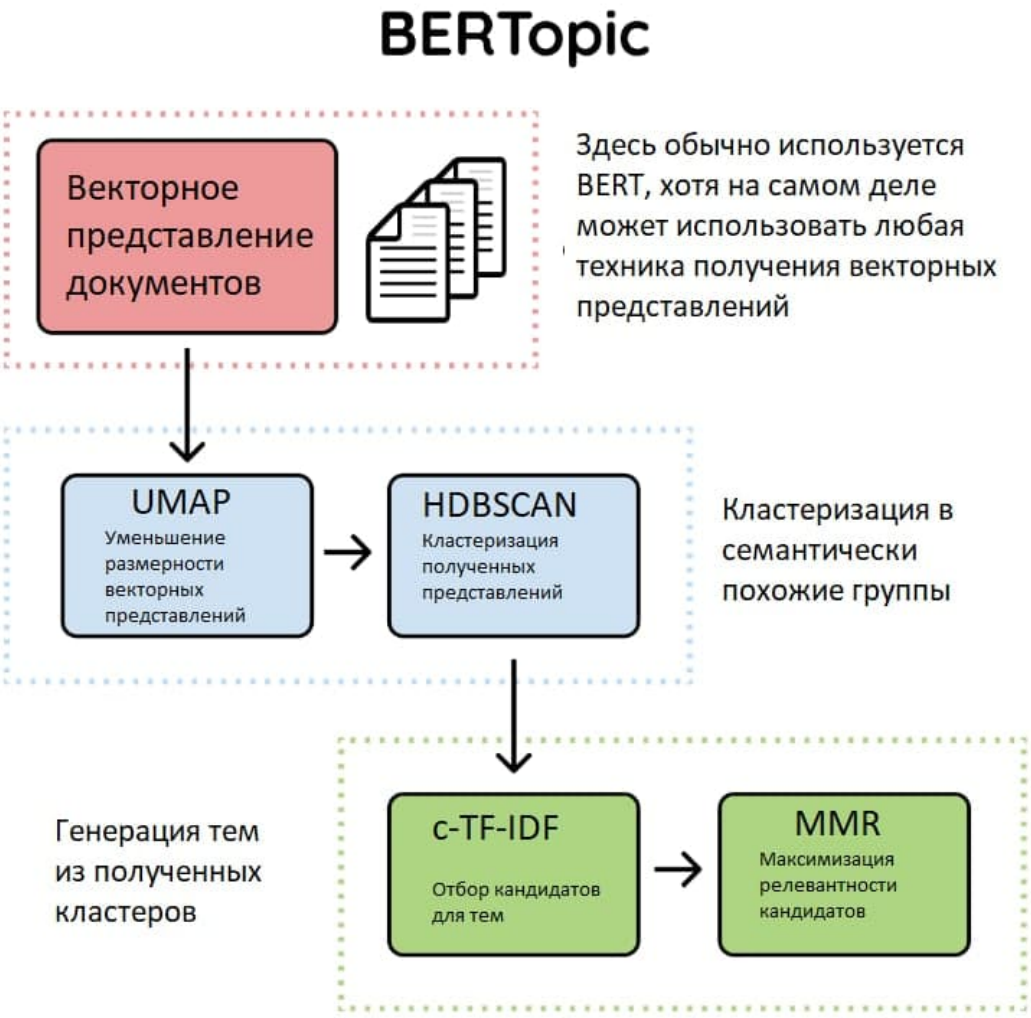
Исходный набор данных Enron был обнародован и размещен в Интернете Федеральной комиссией во время расследования энергетического кризиса в Западной Европе 2000-2001 годов. Позднее набор данных был приобретен Лесли Кельблинг из Массачусетского технологического института, где было выявлено несколько проблем с целостностью данных. Вскоре после этого группа исследователей из SRI International, некоммерческой корпорации, основанной Стэнфордским университетом, во главе с Мелиндой Гарвасио провела серьезную очистку и удаление вложений и отправила ее профессору Уильяму Коэну из Университета Карнеги-Меллона, который сделал публикацию на своей веб-странице. В документе с анализом базы данных Enron, представленном на конференции 2004 года, делается вывод о том, что набор данных Enron «подходит для оценки методов классификации электронной почты».

Используемый набор данных состоит из архивных электронных писем сотрудников Enron, в основном руководителей высшего звена и трейдеров, из которых были удалены вложения. Также некоторые из писем были удалены по требованию сотрудников корпорации. Общее количество писем – более полумиллиона.

Глава 2
Обзор используемых технологий

2.1 BERTopic

BERTopic – это метод моделирования тем, который использует трансформеры и c-TF-IDF для создания плотных кластеров, позволяющих легко интерпретировать темы, сохраняя при этом важные слова в описаниях тем [1].



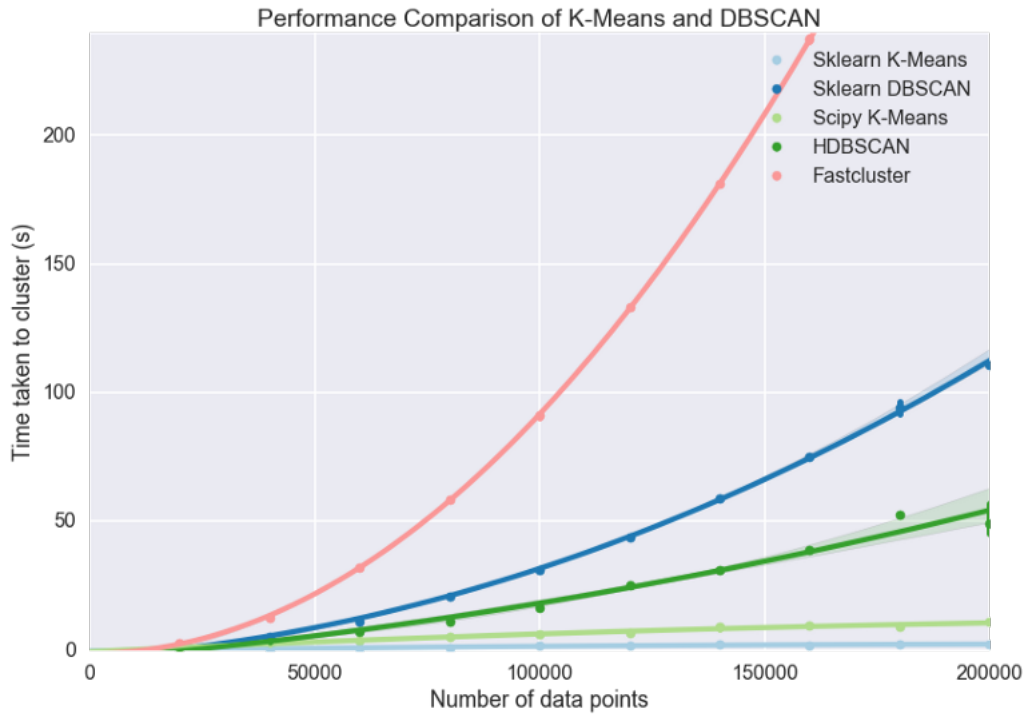
Первый шаг, который нам нужно сделать – это преобразовать документы в числовые данные. Для этой цели мы используем BERT, поскольку он извлекает различные векторные представления слов в зависимости от контекста. Это одна из лучших моделей в настоящее время для многих задач, связанных с текстом. BERT получил награду за лучшую работу на ежегодной конференции североамериканского отделения компьютерной лингвистики 2019 года [2] [3].

Мы хотим убедиться, что документы с похожим смыслом сгруппированы вместе, чтобы мы могли найти темы в этих кластерах. Перед этим нам сна-

чала нужно снизить размерность векторных представлений слов, поскольку многие алгоритмы кластеризации плохо справляются с высокой размерностью. UMAP – это один из немногих алгоритмов уменьшения размерности, он является наиболее эффективным, поскольку он сохраняет значительную часть многомерной локальной структуры в более низкой размерности.

После уменьшения размерности встраиваемых документов мы можем кластеризовать документы с помощью HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). HDBSCAN основан на алгоритме DBSCAN и, как и другие алгоритмы кластеризации, используется для группировки данных [4].

Помимо того, что он обычно показывает лучшее качество, он также быстрее, чем обычный DBSCAN. Ниже приведен график нескольких алгоритмов кластеризации. При отметке в 200 000 объектов DBSCAN занимает примерно вдвое больше времени, чем HDBSCAN. Стоит отметить, что по мере увеличения количества объектов разница в производительности будет и дальше увеличиться в пользу HDBSCAN:



HDBSCAN алгоритм кластеризации, который довольно хорошо работает с UMAP, поскольку UMAP поддерживает большую локальную структуру даже в пространстве меньшей размерности. Более того, HDBSCAN не переносит отдельные точки в кластеры, поскольку считает их выбросами.

Теперь мы сгруппировали похожие документы вместе, которые должны представлять темы, из которых они состоят. Что мы хотим узнать из созданных нами кластеров – это то, что отличает один кластер по своему содержанию от другого. Как мы можем извлечь темы из сгруппированных документов? Чтобы решить эту проблему, используется классовый вариант TF-IDF (c-TF-IDF), который позволил бы извлечь то, что делает каждый набор до-

кументов уникальным по сравнению с другим. Интуиция, лежащая в основе метода, заключается в следующем: когда мы применяем TF-IDF как обычно к набору документов, мы в сравниваем важность слов среди всех документов, а в классовом варианте теперь у нас есть одно значение важности для каждого слова в кластере, которое можно использовать для создания темы. Если мы возьмем несколько самых важных слов в каждом кластере, то получим хорошее представление о кластере и, следовательно, теме.

Чтобы создать эту оценку классового TF-IDF, нам нужно сначала создать один документ для каждого кластера документов:

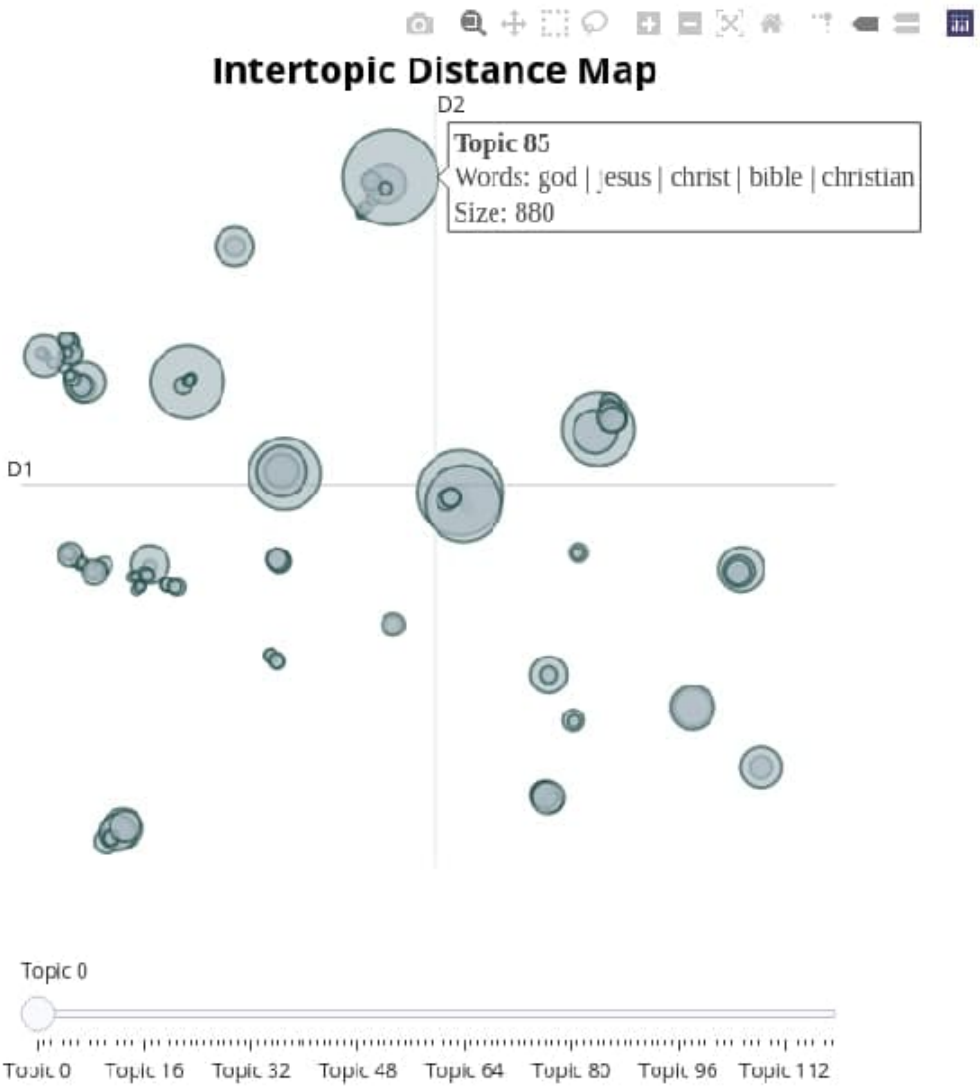
```
docs_df = pd.DataFrame(data, columns=["Doc"])
docs_df['Topic'] = cluster.labels_
docs_df['Doc_ID'] = range(len(docs_df))
docs_per_topic = docs_df.groupby(['Topic'], as_index = False) \
.agg({'Doc': ' '.join})
```

Затем мы применяем TF-IDF на основе классов:

$$\text{c-TF-IDF}_i = \frac{t_i}{w_i} \cdot \log \frac{m}{\sum_j^n t_j}$$

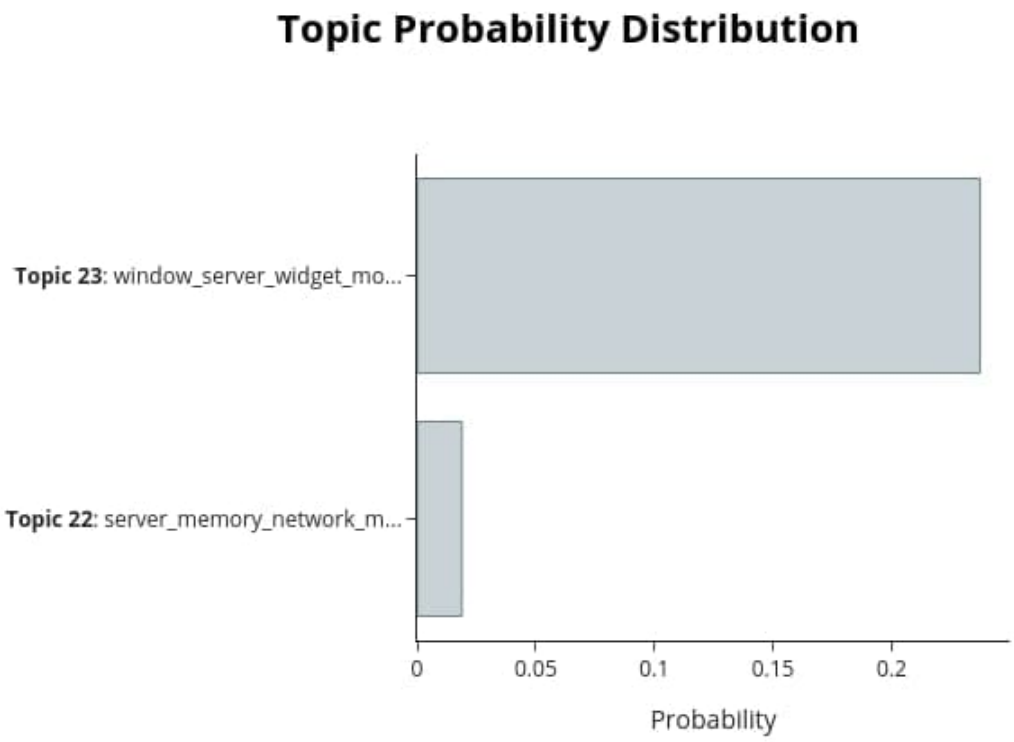
Где частота каждого слова t извлекается для каждого класса i и делится на общее количество слов w в классе. Это действие можно рассматривать как форму регуляризации частых слов в классе. Затем общее количество документов m делится на общую частоту слова t по всем n классам.

Теперь у нас есть одно значение важности для каждого слова в кластере, которое можно использовать для создания темы. После обучения нашей модели мы можем итеративно пройти, возможно, сотню тем, чтобы получить хорошее представление о темах, которые были извлечены. Однако это занимает некоторое время и не имеет глобального представления. Вместо этого мы можем визуализировать темы. Для этого используется представление тем в 2D с помощью UMAP, а затем визуализируем два измерения с помощью графического представления, чтобы мы могли создать интерактивное представление. Пример визуализации от авторов алгоритма:



На картинке видно, что кластеры довольно равномерно распределились по пространству и кластера действительно очень интерпремируемы: например, на рисунке показан кластер религиозных слов.

Мы также можем рассчитать вероятность того, что темы могут быть найдены в документе. Эти вероятности означают, насколько BERTopic уверен в том, что определенные темы могут быть найдены в документе:



Важно понимать, что распределение вероятностей не указывает на распределение частотности тем в документе. Это просто показывает, насколько BERTopic уверен в том, что в документе можно найти определенные темы.

Глава 3
Предобработка данных

Машинное обучение является мощным и эффективным инструментом при реализации алгоритмов классификации, маршрутизации, обработки и поиска документов, однако, определяющее значение в этих процессах имеет качество исходных данных [6]. Именно поэтому проведение подготовки исходных документов, их предварительная обработка позволяет значительно повысить точность результатов, получаемых в ходе применения машинного обучения.

3.1 Предобработка электронных писем Хиллари Клинтон

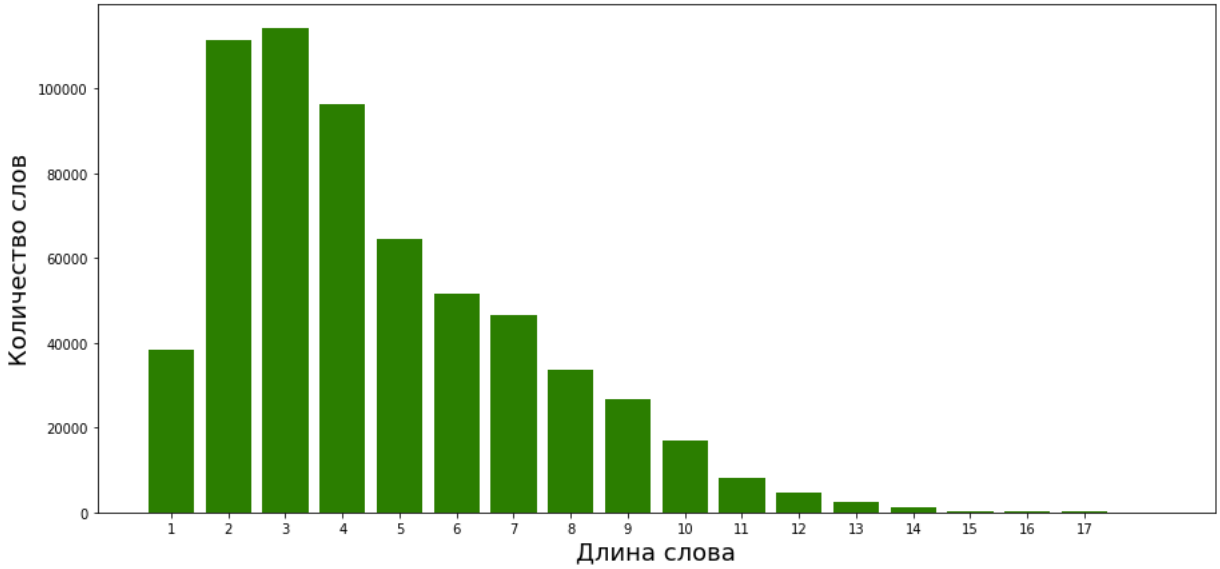
Этап предобработки можно разбить на 6 шагов.

1. На вход поступает множество документов определенных форматов (txt, doc или pdf, как в нашем случае). Выбирается библиотека программного кода в зависимости от формата исходного документа и осуществляется извлечение данных из документа в виде неформатированного текста. Этот шаг уже произведен платформой *Kaggle*. Общее количество электронных писем — 7945.

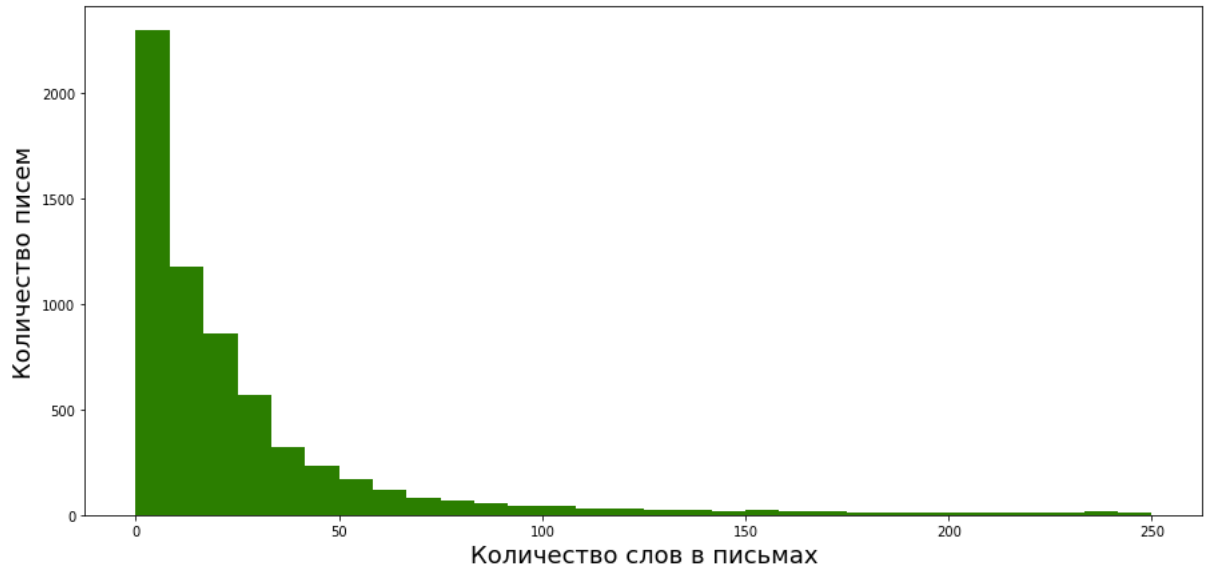
2. В текстовом файле, взятом из *Kaggle*, могут быть пропущенные данные (например, в связи с плохим качеством pdf-файла). Такие данные пропускаются и нами не обрабатываются. После осуществления этого шага остается 6742 писем.

3. Текст каждого электронного письма проходит процесс нормализации — удаляются знаки препинания, выделяются отдельные слова. После этого каждое слово приводится в нижний регистр.

   На этом шаге для дальнейшего анализа можно посмотреть на различного рода статистики. Ниже приведена гистограмма распределения количества слов каждой длины:



А ниже приведена гистограмма распределения количества электронных писем каждой длины:



4. Дальше происходит фильтрация текста по стоп-листу — набору коротких слов (артиклей, предлогов, местоимений), не несущих большой смысловой нагрузки, что приводит к сокращению объема текста и повышению его смысловой ценности.

5. Следующим шагом происходит лемматизация — процесс приведения слов к леммам, т. е. нормальным словесным формам. Для реализации лемматизации можно использовать библиотеку программного кода *spaCy* [7], позволяющую привести все слова к нормальной форме. Полученный после выполнения лемматизации набор слов уже может использоваться для проведения машинного обучения и решения конкретных задач.

6. Индексация — построение некоторой числовой модели текста, которая

переводит текст в удобное для дальнейшей обработки представление.

3.2 Предобработка электронных писем корпорации Enron

3.2.1 Выделение метаданных из сырого текста писем

Набор данных *Enron* также требует предварительной обработки. К примеру, так выглядит необработная информация одного электронного письма:

```
Message-ID: <18782981.1075855378110.JavaMail.evans@thyme>
Date: Mon, 14 May 2001 16:39:00 -0700 (PDT)
From: phillip.allen@enron.com
To: tim.belden@enron.com
Subject:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: Tim Belden <Tim Belden/Enron@EnronXGate>
X-cc:
X-bcc:
X-Folder: \\Phillip_Allen_Jan2002_1\\Allen, Phillip K.\\\'Sent Mail
X-Origin: Allen-P
X-FileName: pallen (Non-Privileged).pst

Here is our forecast
```

Конечно, анализировать данные (в том числе метаинформацию о письме) в таком формате бессмысленно. Для обработки мы будем использовать библиотеку *email* [12]. Данная библиотека позволяет из сырых данных выделить вспомогательную информацию о письме, в частности, библиотека позволяет выделить следующие интересные нам атрибуты:

- полное содержание письма,
- дата отправки,
- адрес получателя,
- адрес отправителя,
- тема письма,
- логин отправителя.

3.2.2 Выделение содержания писем

После этого требуется также привести содержание письма в приемлимый для дальнейшего обучения вид. Например, для письма с содержанием ниже мы хотим выделить только единицы, имеющие отношение к сути письма.

```
Forwarded by Phillip K Allen/HOU/ECT on 09/12/2000 11:22 AM

Michael Etringer

09/11/2000 02:32 PM

To: Phillip K Allen/HOU/ECT@ECT
cc:
Subject: Contact list for mid market

Phillip,
Attached is the list. Have your people fill in the columns
highlighted in yellow. As best can we will try not to overlap
on accounts.
Thanks, Mike
```

Выделение этих единиц происходит в соответствии со следующей последовательности шагов:

1. Перевод всех символов в нижний регистр.
2. Удаление всех слов, содержащих цифры. Такие слова не несут смысловой нагрузки и, соответственно, влияют на качество обучения в худшую сторону (подавляющее большинство таких слов было получено отправителями по ошибке).
3. Удаление единиц, соответствующих информации о пересланных письмах.
4. Удаление единиц, соответствующих информации о вложениях в письме.
5. Удаление единиц, соответствующих почтовым адресам, присутствующим в тексте писем.
6. Удаление единиц, соответствующих корпоративным именам пользователей.
7. Удаление единиц, соответствующих ссылкам в сети Интернет.

8. Удаление единиц, не несущих смысловой составляющей, в заголовке письма.

Шаги 3-7 выполняются с использованием регулярных выражений.

В результате, для примера выше, дальнейшая работа будет производиться со следующим текстом:

```
contact list for mid market. phillip, attached is the list.
have your people fill in the columns highlighted in yellow.
as best can we will try not to overlap on accounts. thanks, mike'
```

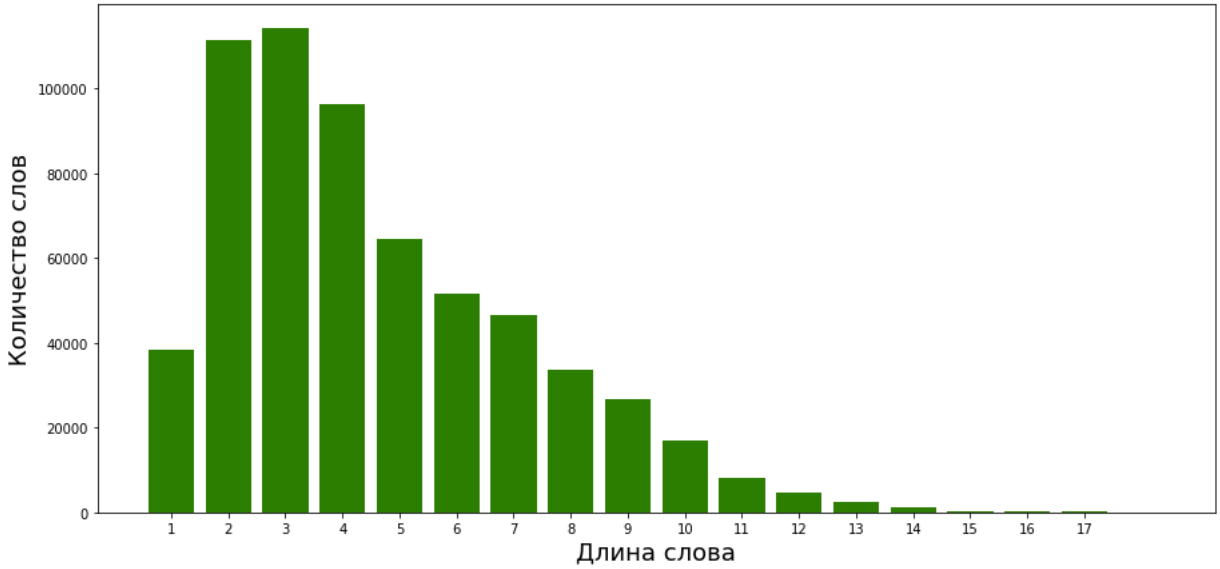
Глава 4
Анализ электронных писем

Прежде чем приступать к исследованию данных методами машинного обучения, может быть полезно посмотреть на различного статистики.

4.1 Анализ электронных писем Хиллари Клинтон

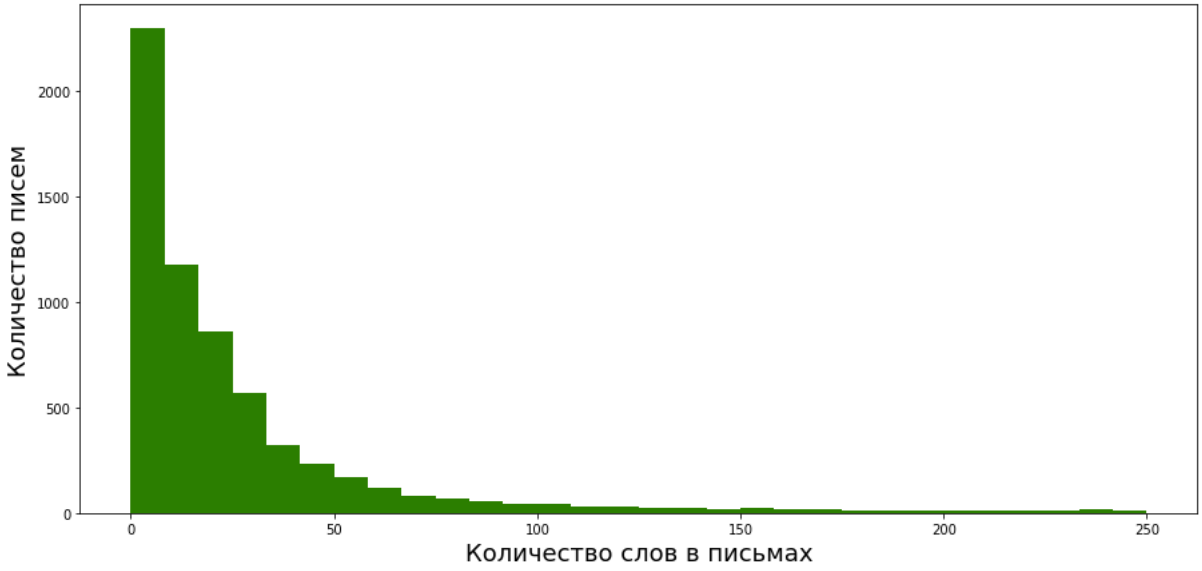
4.1.1 Количества слов

- Гистограмма распределения количества слов каждой длины:



  Гистограмма выглядит вполне естественным образом, много коротких слов (например, местоимений, предлогов).

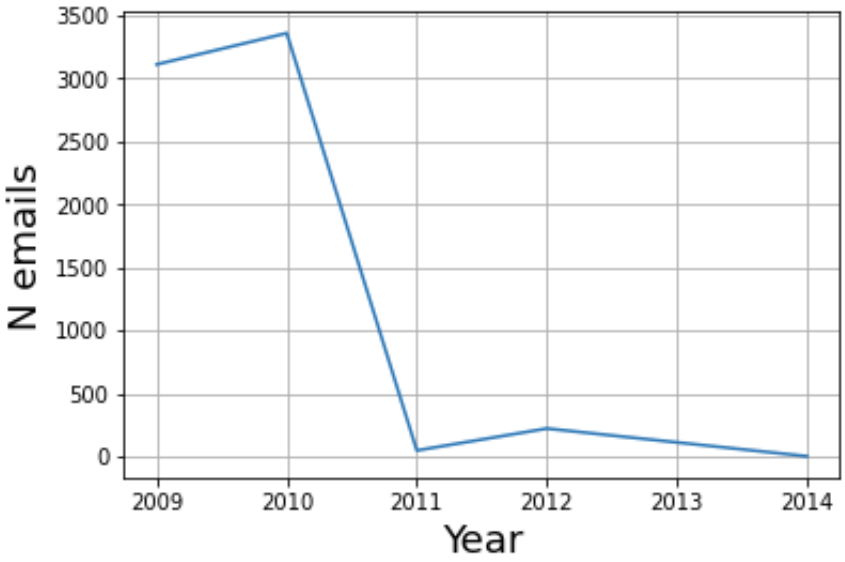
- Гистограмма распределения длин (в количестве слов) писем:



Гистограмма соответствует интуитивным ожиданиям – более длинные письма пишутся реже.

4.1.2 Время отправки писем

- Количество отправленных писем по годам:



На графике можно заметить странную аномалию с нулем писем в 2011 году. Вероятнее всего, это связано с особенностями набора данных.

- Количество отправленных писем по дням недели:

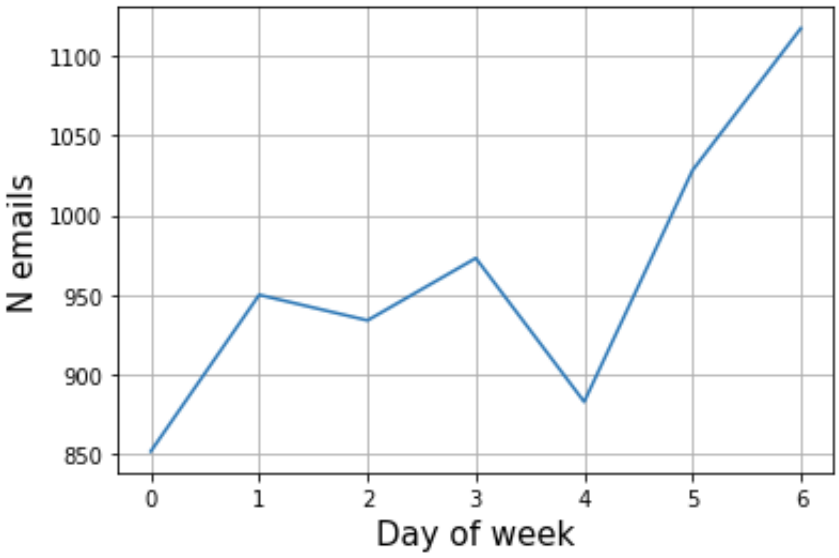
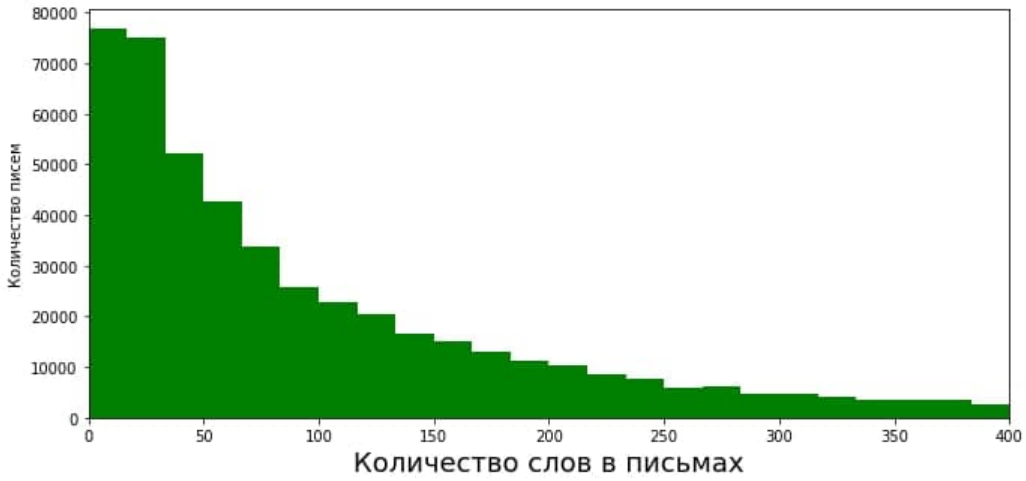


График выглядит слегка неестественно (в отличие от *Enron*). Можно попытаться интерпретировать это как особенности одного отдельного человека, занимающего специфичным видом деятельности.

4.2 Анализ электронных писем корпорации Enron

4.2.1 Длины писем

Гистограмма распределения длин (в количестве слов) писем:



Гистограмма, как и в случае писем Клинтон, соответствует интуитивным ожиданиям – более длинные письма пишутся реже.

4.2.2 Время отправки писем

- Количество отправленных писем по годам:

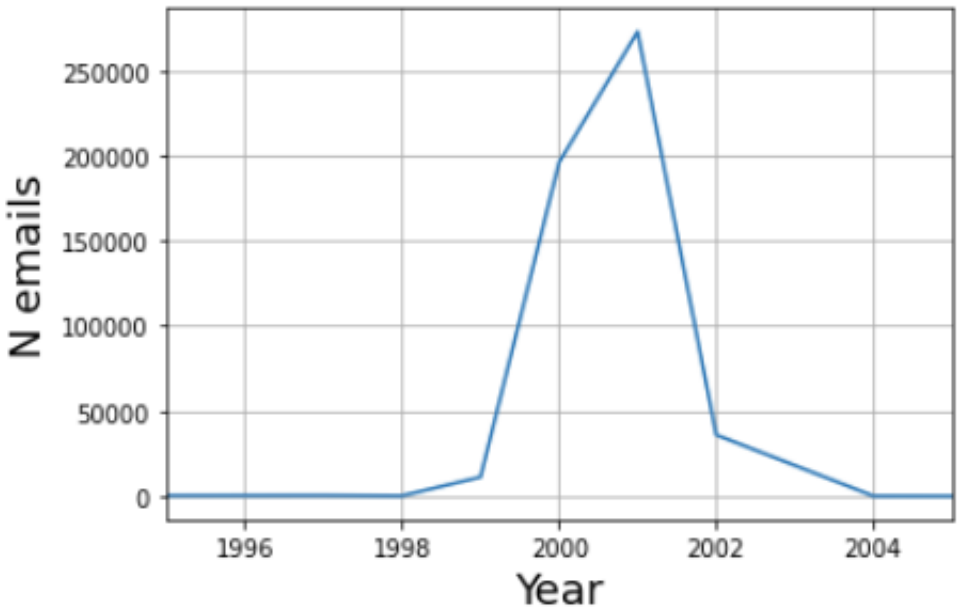
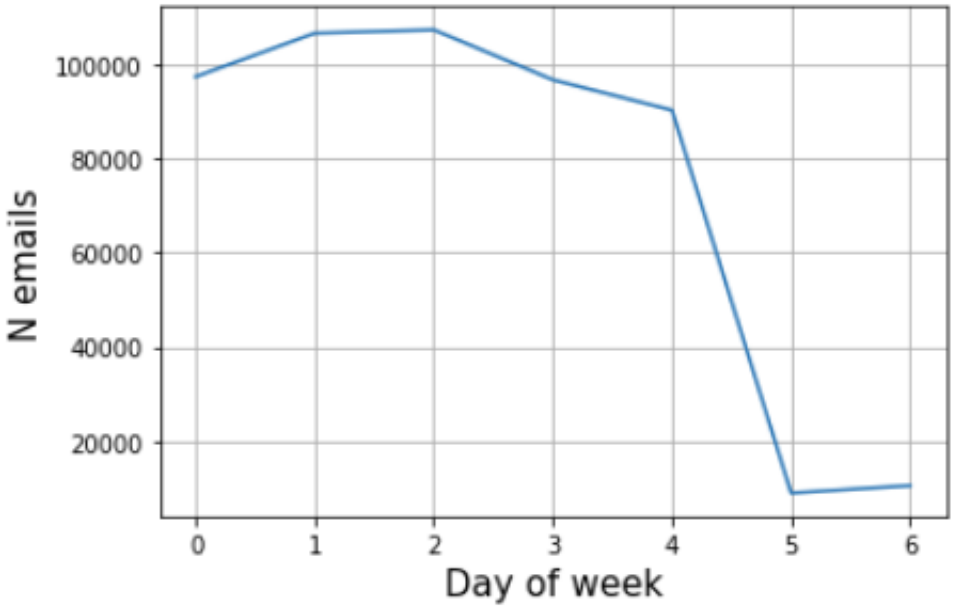


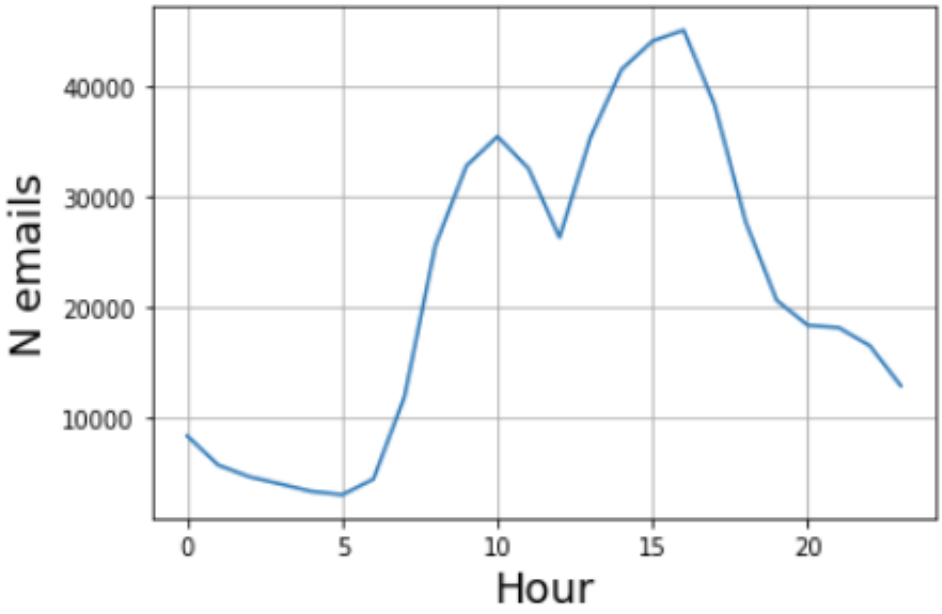
  График соответствует наибольшей активности компании в 2000-2001 годах и банкротству к концу 2001 года.

- Количество отправленных писем по дням недели:



Этот график также выглядит естественно — наибольшее число писем во вторник и среду, в середине рабочей недели, наименьшее — в выходные дни.

- Количество отправленных писем по времени суток:



На графике вышем видим наибольшую продуктивность во вторую половину дня, низкую активность в ночные часы, а также аномалию в самом разгаре дня, объсняющуюся обеденным перерывом.

4.2.3 Частотность слов

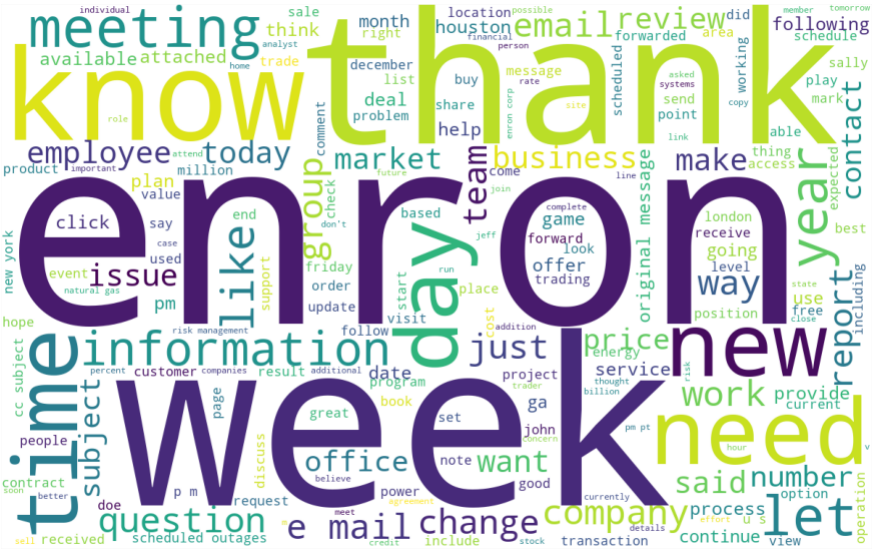
Для интерпретации самых часто используемых слов использовались так называемые облака слов.

Облако слов – это метод визуализации данных, используемый для представления текстовых данных, в которых размер каждого слова указывает его частоту или важность. Важные точки текстовых данных могут быть выделены с помощью облака слов. Облака слов широко используются для анализа данных с веб-сайтов социальных сетей.

- Облако слов, построенное по словам из тем электронных писем:

Слова, встречающиеся в темах писем: `meeting`, `conference room`, `update`, `approval`,`access request`, `enron`.

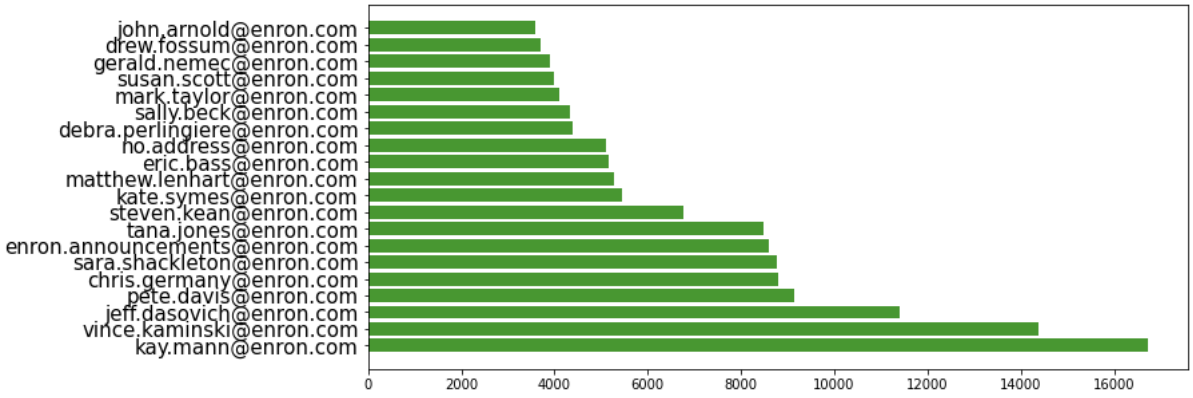
- Облако слов, построенное по словам из содержания электронных писем:



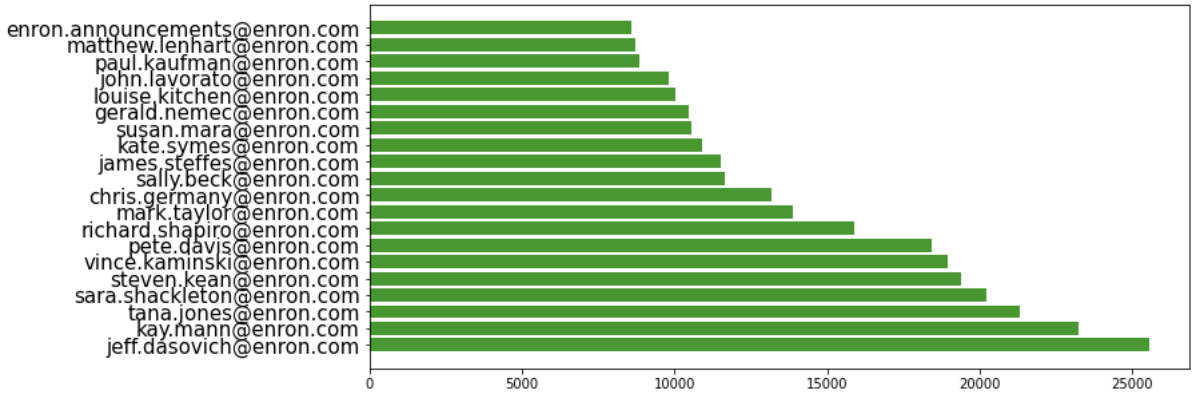
Слова, встречающиеся в содержании писем: `enron`, `week`, `thank`, `know`, `new`.

4.2.4 Отправители и получатели писем

- 20 адресов, с которых было отправлено наибольшее количество электронных писем:



- 20 адресов, на которые было отправлено наибольшее количество электронных писем:



  Как видим, в графиках распределения получателей и отправителей писем много различий — некоторые люди пишут писем меньше, чем получают и наоборот.

- Теперь посмотрим количество писем между фиксированной парой собеседников. Рассмотрим только электронные письма, отправленные на один адрес электронной почты, так как они могут быть более важными личными сообщениями.

Отправитель	Получатель	Количество
pete.davis	pete.davis	9141
vince.kaminski	vkaminski@aol.com	4308
enron.announcements	all.worldwide	2206
enron.announcements	all.houston	1701
kay.mann	suzanne.adams	1528
vince.kaminski	shirley.crenshaw	1190
steven.kean	maureen.mcvicker	1014
kay.mann	nmann@erac.com	980
kate.symes	evelyn.metoyer	915
kate.symes	kerri.thompson	859

Здесь интересно, что некоторые люди отправляют сами себе много электронных писем.

Глава 5
Исследование данных

5.1 Тематическое моделирование

Тематическая модель (англ. *topic model*) — модель коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции. Алгоритм построения тематической модели получает на входе коллекцию текстовых документов. На выходе для каждого документа выдаётся числовой вектор, составленный из оценок степени принадлежности данного документа каждой из тем. Размерность этого вектора, равная числу тем, может либо задаваться на входе, либо определяться моделью автоматически.

Тематическое моделирование (англ. *topic modeling*) — построение тематической модели.

Задача построения тематической модели звучит следующим образом. Задана коллекция текстовых документов D. Каждый документ d из коллекции D представляет собой последовательность слов $W_d = (w_1, \ldots, w_{n_d})$ из словаря W, где n_d — длина документа d. Предполагается, что каждый документ может относиться к одной или нескольким темам. Темы отличаются друг от друга различной частотой употребления слов. Требуется найти эти темы, то есть определить

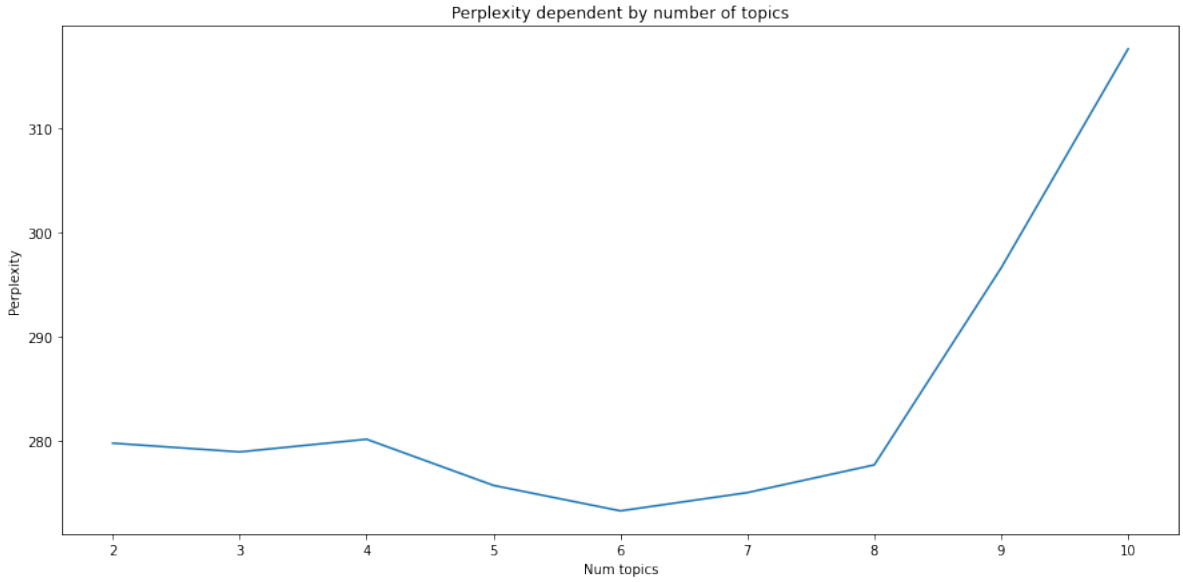
- число тем;
- распределения частот слов, характерное для каждой темы;
- тематику каждого документа — в какой степени он относится к каждой из тем.

Данная задача может рассматриваться как задача одновременной кластеризации документов и слов по одному и тому же множеству кластеров, называемых темами. Строится, так называемая, мягкая кластеризация, то есть один документ может принадлежать нескольким темам в различной степени.

Для тематического моделирования в качестве модели в данной работе используется латентное размещение Дирихле (англ. *latent Dirichlet allocation, LDA*) [8].

Для оценки качества данной модели используется перплексия (англ. *perplexity*) — оценка того, насколько хорошо вероятностная модель предсказывает выборку. Низкая перплексия указывает на то, что распределение вероятностей хорошо предсказывает выборку.

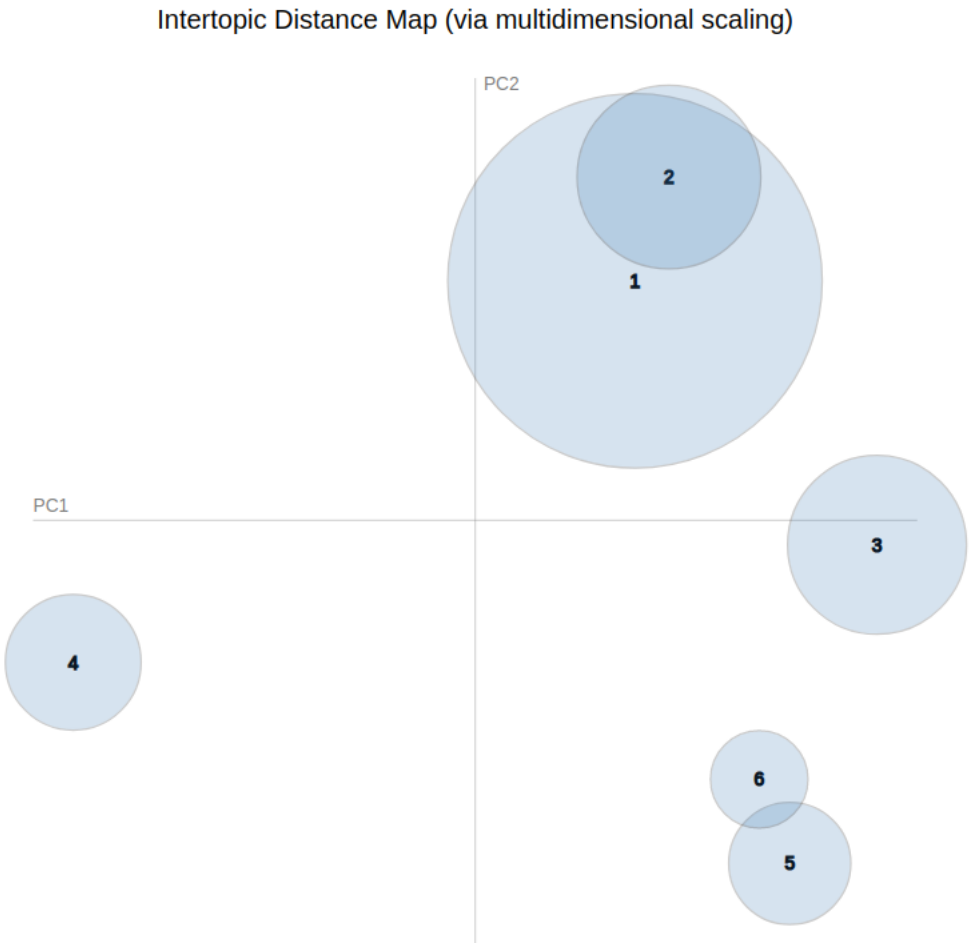
В зависимости от параметра модели, отвечающего за количество тем у распределения текстов, получилась следующая зависимость значения перплексии от количества тем:



Ниже приведены примеры слов, принадлежащие каждой из 6 (с оптимальным значением перплексии) тем:

Номер темы	Слова
1	obama, state, president, government, american, israel, policy, country
2	woman, say, work, health, year, senate, group, government, support, company
3	call, get, work, see, want, know, good, also, think, tomorrow
4	secretary, office, state, meet, room, department, arrive, route, depart, private
5	state, information, benghazi, department, doc, case, subject, iran, agreement, house
6	cheryl, gov, fyi, sullivan, state, friday, sunday, branch, wednesday, april, january

Распределение слов по темам:



5.2 Кластеризация слов из электронных писем

Кластерный анализ (англ. *Data clustering*) — задача разбиения заданной выборки объектов (ситуаций) на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Кластеризация (обучение без учителя) отличается от классификации (обучения с учителем) тем, что метки исходных объектов изначально не заданы.

В данной работе мы группируем похожие по смыслу слова с помощью векторного представления слов, полученных с помощью *Word2Vec* [9].

Word2Vec принимает большой текстовый корпус в качестве входных данных и сопоставляет каждому слову вектор, выдавая координаты слов на выходе. Сначала он генерирует словарь корпуса, а затем вычисляет векторное представление слов, «обучаясь» на входных текстах. Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (а следовательно, имеющие схожий смысл), будут иметь близкие (по косинусному расстоянию) векторы. Полученные векторные представления слов могут быть использованы для обработки естественного языка и машинного обучения.

Текстовый корпус, состоящий из слов из электронных писем, недостаточно большой, чтобы получить хорошие результаты. Поэтому мы использовали предобученный датасет, полученный из постов в Twitter [10], который был дообучен словами из электронных писем.

Полученные вектора кластеризуются с помощью алгоритма *K-Means*. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k. Действие алгоритма таково, что он стремится минимизировать среднеквадратичное отклонение на точках каждого кластера. Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров.

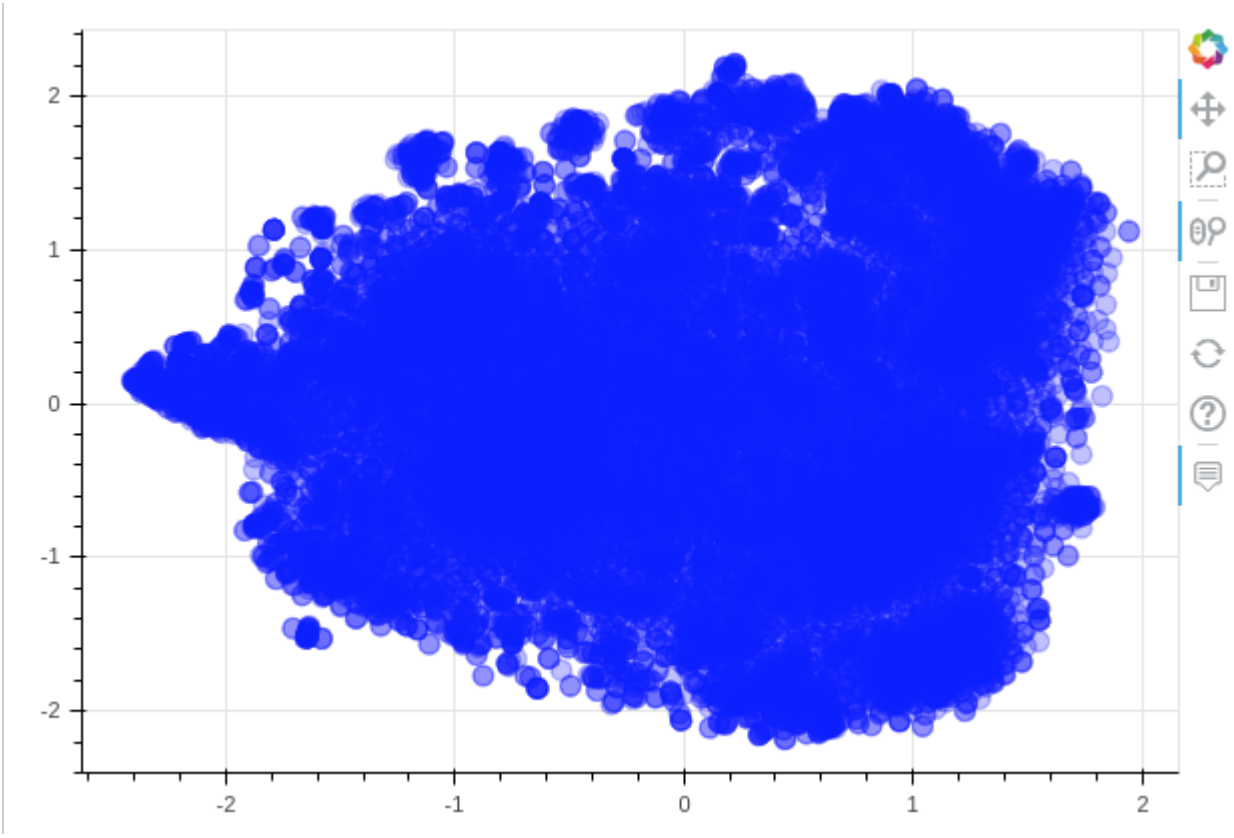
Результаты работы алгоритма. Ближайшие слова к «*obama*»:

Слово	Расстояние
romney	0.9429854154586792
barack	0.9073218107223511
president	0.8986026048660278
clinton	0.8913119435310364
hillary	0.8597259521484375
say	0.8407208323478699
hovv	0.8315389752388

Ближайшие слова к «*trump*»:

Слово	Расстояние
appropriator	0.7439741492271423
infighter	0.7368026971817017
zappos	0.7316897511482239
perkins	0.7260088920593262
donald	0.7180437445640564
buffett	0.7113708853721619
bloomberg	0.7067334651947021
clinton	0.7052138447761536

Так же была построена интерактивная проекция точек на *2D*-плоскость с помощью алгоритма *t-SNE* [11]. *t-SNE* — это техника нелинейного снижения размерности, хорошо подходящей для вложения данных высокой размерности для визуализации в пространство низкой размерности (двух- или трехмерное). В частности, метод моделирует каждый объект высокой размерности двух- или трёхмерной точкой таким образом, что похожие объекты моделируются близко расположенными точками, а непохожие точки моделируются с большой вероятностью точками, далеко друг от друга отстоящими.


2
1
0
-1
-2
-2
-1
0
1
2

Заключение

В данной работе были проведены эксперименты с исследованием текстов из электронных почты Хиллари Клинтон.

Основная проблема в исследовании — недостаточно большой размер датасета. Это приводит к проблеме с недостаточным уровнем обученности моделей. Она решается с помощью предобученных датасетов большего размера.

И тематическое моделирование, и кластеризация показали неплохие интерпретируемые результаты, о чем можно судить по представленным таблицам в соответствующих разделах.

Литература

1. BERTopic. https://maartengr.github.io/BERTopic/.

2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805.

3. Lena Voita. (Introduction to) Transfer Learning. https://lena-voita.github.io/nlp_course/transfer_learning.html.

4. Brendan Bailey. Lightning Talk: Clustering with HDBScan. https://towardsdatascience.com/lightning-talk-clustering-with-hdbscan-d47b83d1b03a.

5. Hillary Clinton's Emails, https://www.kaggle.com/kaggle/hillary-clinton-emails.

6. Обухов А. Д. Постановка задачи структурно-параметрического синтеза системы электронного документооборота научно-образовательного учреждения // Вестник ТГТУ. – 2016. – № 2. – С. 217–232. – DOI: 10.17277/vestnik.2016.02.pp.217-232.

7. Библиотека spaCy. https://spacy.io/.

8. David M. Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation // Journal of Machine Learning Research (3) 2003 pp. 993-1022.

9. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // In Proceedings of Workshop at ICLR. — 2013a.

10. GloVe: Global Vectors for Word Representation. https://nlp.stanford.edu/projects/glove/.

11. van der Maaten L.J.P., Hinton G.E. Visualizing Data Using t-SNE // Journal of Machine Learning Research. — 2008. — Ноябрь (т. 9).

12. Email — An email and MIME handling package. https://docs.python.org/3/library/email.html.