Министерство образования и науки Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
“Московский физико-технический институт (государственный
университет)”
Факультет инноваций и высоких технологий
Кафедра “Банковских информационных технологий”

ИССЛЕДОВАНИЕ И ВЫЯВЛЕНИЕ МЕТРИК ЭФФЕКТИВНОСТИ КОММУНИКАЦИЙ ВНУТРИ КОМПАНИИ НА ОСНОВЕ МЕТАДАННЫХ ЭЛЕКТРОННОЙ ПОЧТЫ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ АНАЛИЗА СОЦИАЛЬНЫХ СЕТЕЙ И АНАЛИЗА ТЕКСТА

Выпускная квалификационная работа
(магистерская диссертация)

Направление подготовки: 01.04.02 Прикладные математика и информатика

Выполнила:
студентка 194в группы Емельянова Анна Александровна

Научный руководитель:
д.ф.-м.н., профессор Теренин Алексей Алексеевич

Москва 2017

Содержание

1. Аннотация

Данная работа посвящена исследованию пользовательского поведения в использовании рабочих почтовых сервисов. Главная задача этого исследования состоит в том, чтобы выявить метрики эффективности коммуникаций внутри компаний на основе метаданных электронной почты, а также изучить возможность нахождения лексических единиц, способных помочь понять причины возникновения каких-либо тенденций в поведении пользователей или которые можно рассматривать в качестве предикторов для какой-либо ситуации или проблемы.

2. Введение

2.1. Актуальность цели исследования

В крупных компаниях часто возникает проблема в неэффективном использовании почтовых сервисов, что ведет к загруженности почтового ящика, потери информации и не комфортному рабочему процессу. Существуют организации, в которых эффективные коммуникации являются залогом их успеха. Ярким примером являются компании по продаже автомобилей, где специально обученные менеджеры ведут иногда долгий и трудный разговор с покупателями в личных встречах, по телефону, по почте. Как выяснить, в случае часто проваленных сделок, их причину? Проанализировав коммуникационные данные между сотрудниками: почтовые переписки, переписки в социальных сетях, телефонные звонки и т.п., можно понять, в чем кроется весьма неочевидный источник проблем. Такое исследование поможет выявить важные показатели эффективной совместной работы, которые помогут менеджерам понять, как организовать и руководить группой сотрудников.

Несмотря на широко распространенное использование электронной почты, как ни странно, мало что известно о свойствах почтовой службы и о социально-поведенческих закономерностях электронной почты, сформированных на крупных предприятиях. Социальные сети предприятия демонстрируют несколько особенностей по сравнению с другими популярными социальными сетями, так как сообщения электронной почты следуют из повседневного сотрудничества, а поток информации тесно связан с основной организационной структурой.

Подробное изучение пользовательского взаимодействия может открыть некоторые закономерности. Подобные исследования проводил Питер Глур [1], в результате которых были выявлены интересные факты. Например, если в коллективе периодически менять лидера, то такая команда будет работать лучше, чем та, где лидер не меняется на протяжении долгого времени. В данной работе проводится исследование нахождение тенденций и поведенческих закономерностей в использовании электронных почтовых

сервисов компаний с целью выявления проблем и метрик качества коммуникаций.

2.2. Структура исследования

В данной работе проводится исследование использования почтовой службы и потока информации в двух компаниях. Первая часть исследования заключается в изучении анонимизированных данных компании №1 в целях изучения пользовательского поведения, выявления проблем и метрик эффективности коммуникаций. Изучаются отношения пользователей и их влияние на общий поток информации. Данный анализ дает ряд характеристик, которые могут служить ориентиром для будущего усовершенствованного почтового сервиса и более эффективного коммуницирования внутри компании.

Во второй части используются данные компании №2, которые содержат также и все текстовые сообщения. Проводится сравнение основных поведенческих характеристик пользователей почтового сервиса компании №1 и компании №2. Затем исследуется взаимосвязь между текстами сообщений, которые люди пишут в электронной почте на рабочем месте, и полученными ранее метриками. Такое исследование показывает, что определенные слова и фразы могут являться сильными индикаторами для рассматриваемых характеристик.

Полученные результаты и методы исследования, содержащиеся в данной работе, могут применяться для улучшения рабочего процесса в других компаниях.

3. Используемые методы

Данный раздел посвящен краткому описанию основных методов, которые применялись в исследовательской работе и показали наилучшие результаты в достижении поставленных целей.

3.1. Логистическая регрессия

Логистическая регрессия является одним из статистических методов классификации с использованием линейного дискриминанта Фишера. В отличие от обычной регрессии, в методе логистической регрессии не производится предсказание значения числовой переменной исходя из выборки исходных значений. Вместо этого, значением функции является вероятность того, что данное исходное значение принадлежит к определенному классу.

Определение

Пусть объекты описываются *n* числовыми признаками $f_j\colon X \to \mathbb{R},\ j = 1,\ldots,n$. Тогда пространство признаковых описаний объектов есть $X = \mathbb{R}^n$. Пусть Y — конечное множество номеров (имён, меток) классов. Пусть задана обучающая выборка пар «объект, ответ» $X^m = \{(x_1,y_1),\ldots,(x_m,y_m)\}$.

Случай двух классов

Положим $Y = \{-1,+1\}$. В логистической регрессии строится линейный алгоритм классификации $a\colon\ X \to Y$ вида

$$a(x,w) = \operatorname{sign}\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right) = \operatorname{sign}\langle x,w\rangle,$$

где w_j — вес j-го признака, w_0 — порог принятия решения, $w = (w_0,w_1,\ldots,w_n)$ — вектор весов, $\langle x,w\rangle$ — скалярное произведение признакового описания объекта на вектор весов. Предполагается, что искусственно введен «константный» нулевой признак: $f_0(x) = -1$. Задача обучения линейного классификатора заключается в том, чтобы по выборке X^m

настроить вектор весов w. В логистической регрессии для этого решается задача минимизации эмпирического риска с функцией потерь специального вида:

$$Q(w) = \sum_{i=1}^{m} \ln\left(1+\exp(-y_i\langle x_i, w\rangle)\right) \to \min_w.$$

После того, как решение w найдено, становится возможным не только вычислять классификацию $a(x) = \text{sign}\langle x, w\rangle$ для произвольного объекта x, но и оценивать апостериорные вероятности его принадлежности классам:

$$\mathbb{P}\{y|x\} = \sigma\left(y\langle x, w\rangle\right), \quad y \in Y,$$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ — сигмоидная функция. Во многих приложениях апостериорные вероятности необходимы для оценивания рисков, связанных с возможными ошибками классификации.

Многоклассовая логистическая регрессия

Основное отличие многоклассовой логистической регрессии от бинарной заключается лишь в наличии некоторых ограничений.

Оценив условную вероятность каждого класса для объекта x_i с помощью следующей формулы:

$$P(y \mid x_i, w) = \frac{\exp\langle w_y, x_i\rangle}{\sum_{c \in Y} \exp\langle w_c, x_i\rangle}$$

воспользуемся принципом максимума правдоподобия, применив регуляризацию [2], и решим следующую задачу оптимизации:

$$Q(w) = \sum_{i=1}^{l} \log P(y_i \mid x_i, w) - \frac{1}{2C}\sum_{y \in Y} \|w_y\|^2 \to \max_w$$

3.2. Метод максимизации правдоподобия: общий вид

В общем виде метод максимума правдоподобия записывается следующим образом. Пусть некоторая случайная величина x имеет распределение $F(x,\theta), X_n$ — выборка размера n:

$$X \sim F(x,\theta),$$

$$X^n = (X_1,\dots,X_n)$$

Тогда функция правдоподобия имеет вид:

$$L(X^n,\lambda) = \prod_{i=1}^{n} P(X = X_i,\theta)$$

Поскольку при логарифмировании не меняются положения максимумов функции, удобно работать не с самим правдоподобием, а с логарифмом правдоподобия:

$$\ln L(X^n,\lambda) = \sum_{i=1}^{n} \ln P(X = X_i,\theta)$$

Оценкой максимального правдоподобия называется величина:

$$\hat{\lambda}_{OMП} = \arg\max_{\lambda} \ln L(X^N,\lambda)$$

В случае непрерывной случайной величины метод максимального правдоподобия записывается аналогично:

$$X \sim F(x,\theta),$$

$$L(X^n,\lambda) = \prod_{i=1}^{n} f(X = X_i,\theta),$$

$$\hat{\lambda}_{OMП} = \arg\max_{\lambda} \ln L(X^N,\lambda)$$

3.3. Преобразование Бокса-Кокса

В реальности часто приходится иметь дело с данными, которые по тем или иным причинам не проходят тест на нормальность, например, тест Шапиро-Уилика [3]. В этой ситуации есть два выхода: либо обратиться к непараметрическим методам, либо воспользоваться специальными методами, позволяющими преобразовать исходную «ненормальную статистику» в «нормальную». Среди множества таких методов преобразований одним из лучших (при неизвестном типе распределения) считается преобразование Бокса-Кокса.

Описание

Для исходной последовательности длиной N:

$$X = x_0, x_1, x_2, \ldots, x_{N-1}$$

Однопараметрическое Бокс-Кокс преобразование определяется следующим образом:

$$x_i(\lambda) = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda}, \lambda \neq 0 \\ \ln(x_i), \lambda = 0 \end{cases}$$

где $i = 0,1,2,\ldots,N-1$

Это преобразование имеет единственный параметр - лямбда. При значении лямбда равном нулю осуществляется логарифмическое преобразование входной последовательности, при значении лямбда отличном от нуля – степенное. Если параметр лямбда равен единице, то закон распределения исходной последовательности не изменяется, хотя при этом последовательность получит сдвиг за счет вычитания единицы из каждого ее значения.

В зависимости от значения лямбда, преобразование Бокса-Кокса включает в себя следующие частные случаи:

$$\lambda = -1.0, \;\; x_i(\lambda) = \frac{1}{x_i}$$

$$\lambda = -0.5, \;\; x_i(\lambda) = \frac{1}{\sqrt{x_i}}$$

$$\lambda = 0.0, \;\; x_i(\lambda) = ln(x_i)$$

$$\lambda = 0.5, \;\; x_i(\lambda) = \sqrt{x_i}$$

$$\lambda = 2.0, \;\; x_i(\lambda) = {x_i}^2$$

При использовании Бокс-Кокс преобразования необходимо, чтобы все значения входной последовательности были положительными и отличными от нуля. Если входная последовательность не удовлетворяет этим требованиям, то ее можно сдвинуть в положительную область на величину, гарантирующую "положительность" всех ее значений. Для того, чтобы избежать во входных данных появления отрицательных или равных нулю значений, необходимо находить минимальное значение входной последовательности и вычитать его из каждого ее элемента, дополнительно осуществляя сдвиг на небольшую величину, равную 1e-5. Такое дополнительное смещение необходимо для гарантированного сдвига последовательности в положительную область, в случае, если минимальное ее значение равно нулю.

Для последовательностей, которые содержат только положительные значения, такого сдвига можно и не делать, но для того чтобы в процессе преобразования при возведении в степень снизить вероятность получения излишне больших величин, и для "положительных" последовательностей будем используется тот же алгоритм сдвига. Таким образом, любая входная последовательность после сдвига будет располагаться в положительной области, и иметь при этом близкое к нулю минимальное значение.

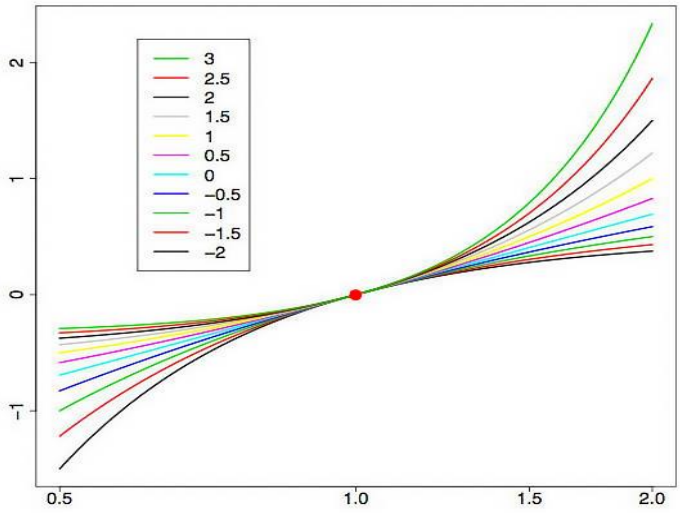


Рис.1. Бокс-Кокс преобразование при различных значениях параметра лямбда

На **рис.1** показано, как выглядят кривые Бокс-Кокс преобразования при различных значениях параметра лямбда. Горизонтальная шкала на графике представлена в логарифмическом масштабе. Как видим, при изменении параметра лямбда "хвосты" исходного распределения могут быть или "растянуты", или "сжаты". Верхняя кривая соответствует значению лямбда=3, а нижняя значению лямбда=-2.

Для того чтобы в результате Бокс-Кокс преобразования закон распределения результирующей последовательности был максимально приближен к нормальному закону, необходимо выбрать оптимальное значение параметра лямбда.

Одним из способов определения оптимальной величины этого параметра является максимизация логарифма функции правдоподобия:

$$f(x,\lambda) = -\frac{N}{2}\ln\left[\sum_{i-0}^{N-1}\frac{\left(x_i(\lambda)-\bar{x}(\lambda)\right)^2}{N}\right]+(\lambda-1)\sum_{i=0}^{N-1}\ln(x_i)$$

$$\bar{x}(\lambda) = \frac{1}{N}\sum_{i=0}^{N-1}x_i(\lambda)$$

То есть необходимо выбрать такое значение параметра лямбда, при котором данная функция принимает максимальное значение. Находить максимум логарифма функции правдоподобия можно разными способами. Например, методом простого перебора. Для этого необходимо в выбранном диапазоне, изменяя с небольшим шагом величину параметра лямбда, вычислять значение функции правдоподобия. И в качестве оптимального значения лямбда выбрать то, при котором величина функции правдоподобия окажется максимальной. При этом величина шага будет определять точность вычисления оптимального значения параметра лямбда. Чем меньше шаг, тем выше точность, но при уменьшении шага пропорционально будет увеличиваться и требуемый объем вычислений. Для повышения эффективности вычислений могут быть использованы различные алгоритмы поиска максимума/минимума функции, генетические алгоритмы и так далее.

4. Исследования и результаты.

4.1. Компания №1

Компания №1 является компанией из группы компаний Сбербанк. Изучаемые данные этого предприятия анонимизированные.

4.1.1. Используемые данные

Имеются метаданные почты за один месяц, размером в 77901 строк, которые имеют следующие основные признаки:

- Timestamp - дата отправления или получения письма
- Sender - отправитель
- Recipients - получатели
- MessageSubject - тема письма
- TotalBytes - размер письма в байтах
- RecipientCount - количество получателей
- MessageId - индетификатор уникального письма

На **рис.1** приведен пример одной строки имеющихся данных

Рис.1 Пример одной строка данных компании №1

```
Timestamp                                          19.05.2016 02:03:19
Sender                  99f912969da4765194bdbd392118e7f9@8d192ad005fa7...
$_.recipients           9a07926a0c2034df187d91cd270f91ee@8d192ad005fa7...
MessageSubject                      55d3697c4c4b03bfb6d60ebae7e6c425
InternalMessageId                                             858559
clientid                                                         NaN
ConnectorId                      SRVEX01\ORT Mail Relay to Customers
$_.recipientstatus                                               NaN
TotalBytes                                                    390153
RecipientCount                                                     1
RelatedRecipientAddress                                          NaN
$_.reference                                                     NaN
ReturnPath              99f912969da4765194bdbd392118e7f9@8d192ad005fa7...
MessageInfo                                              10I: NTS:
EventId                                                    RECEIVE
MessageId               <04d61213-32bb-44e1-8655-86a714ac0b75@SRVEX01....
Name: 0, dtype: object
```

После предварительной обработки из данных были получены следующие статистические показания:

- Всего пользователей: 137
- Количество отправителей: 130
- Количество получателей: 120
- Всего писем: 30171
- Количество тем писем: 10250
- 7 человек не проявляли никакой активности, т.е. 7 человек будут считаться” мертвыми душами”

4.1.2. Результаты исследования данных

Чтобы охарактеризовать общую рабочую нагрузку, рассмотрим объем почтового трафика по количеству элементов электронной почты за 1 месяц, на который распространяется имеющийся набор данных. На **Рис.2** левый график отображает этот объем для всего предприятия, на нем представлено количество отображаемых электронных писем независимо от количества отправителей или получателей на электронную почту. График показывает, что рабочая нагрузка составляет примерно 1.5 тысячи электронных писем в день. Загрузка электронной почты также зависит от количества получателей на электронную почту, то есть сколько раз отправляемый элемент электронной почты должен быть реплицирован. Эта перспектива представлена на правом графике **рис.2** в котором проанализировано среднее количество писем на пару отправителей и получателей в день.

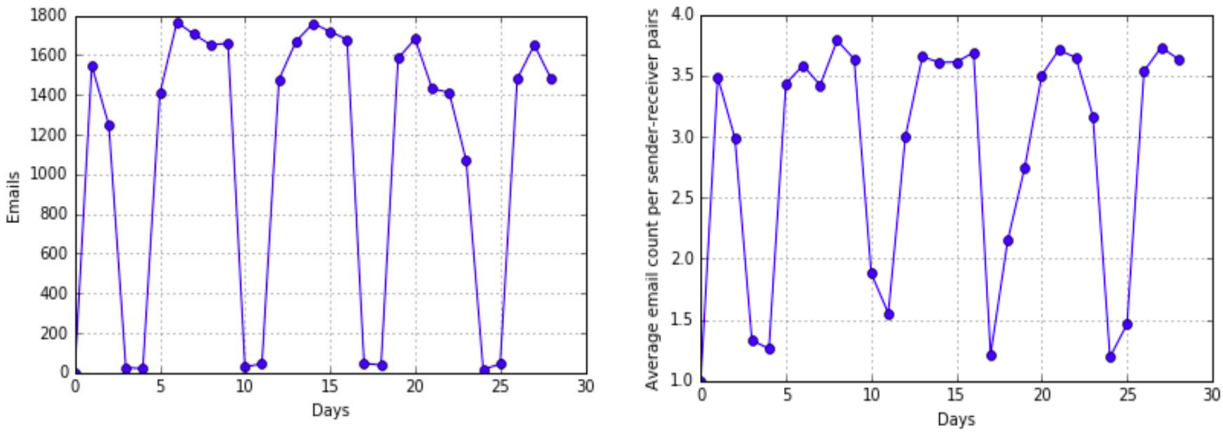


Рис.2 Объем почтового трафика и среднее количество писем в день на пару людей

Далее рассмотрим распределение размеров электронной почты. На левом графике **рис.3** представлена гистограмма размеров электронной почты для имеющихся данных, на правом - комплементарная функция кумулятивного распределения (CCDF [4]). Гистограмма размера электронной почты имеет пики около 3 КБ и 10 КБ, медиана 20 КБ распределения составляет примерно 1 МБ. CCDF размера электронной почты имеет вид степенного закона в широком диапазоне, охватывающий размеры электронной почты порядка от 10 КБ до 3 МБ. До 2 МБ, CCDF экспоненциально ограничена срезом, составляющим примерно 23 МБ. В целом, результаты показывают большое разнообразие размеров электронной почты. Тем не менее, большинство сообщений электронной почты представляют собой текстовые разговоры, обозначенные небольшой медианой, но также имеется немало сообщений, содержащих большие файлы.

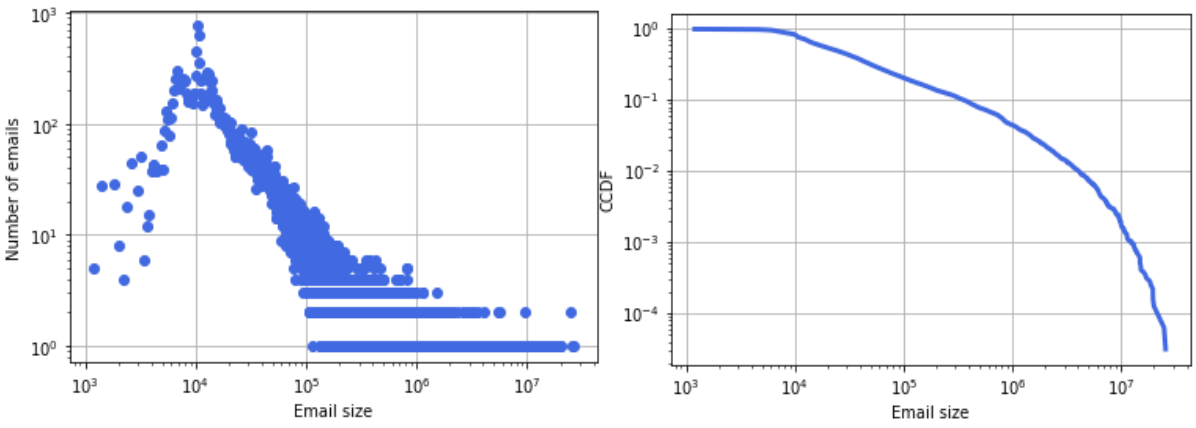


Рис.3. Гистограмма и CCDF размера электронной почты

Какова информационная нагрузка, создаваемая пользователями и навязанная пользователям из электронной почты? Для решения этого вопроса рассмотрим среднее количество отправленных и полученных в день писем электронной почты для каждого пользователя. На **рис.4** и на **рис.5** представлены соответствующие гистограммы и CCDF, показывающие распределения всех электронных писем, полученных и отправленных пользователями.

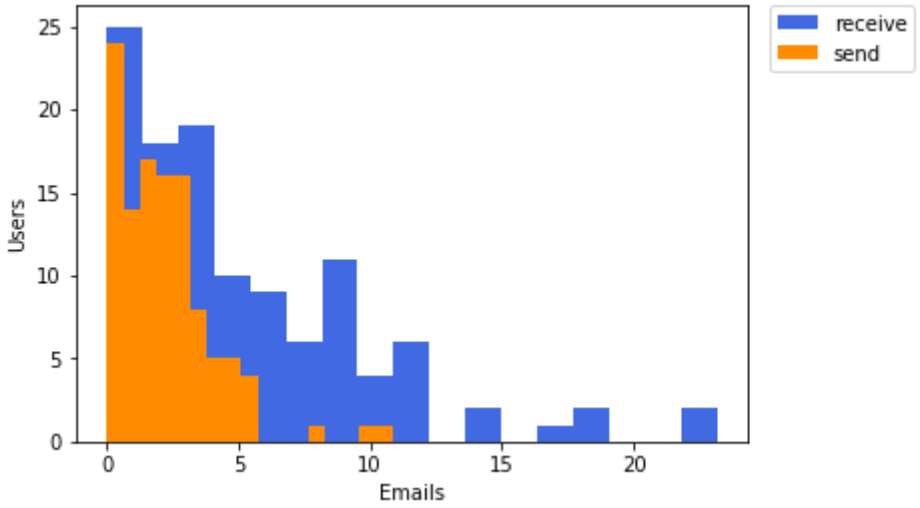


Рис.4. Гистограмма среднего количество отправленных и полученных в день писем для каждого пользователя

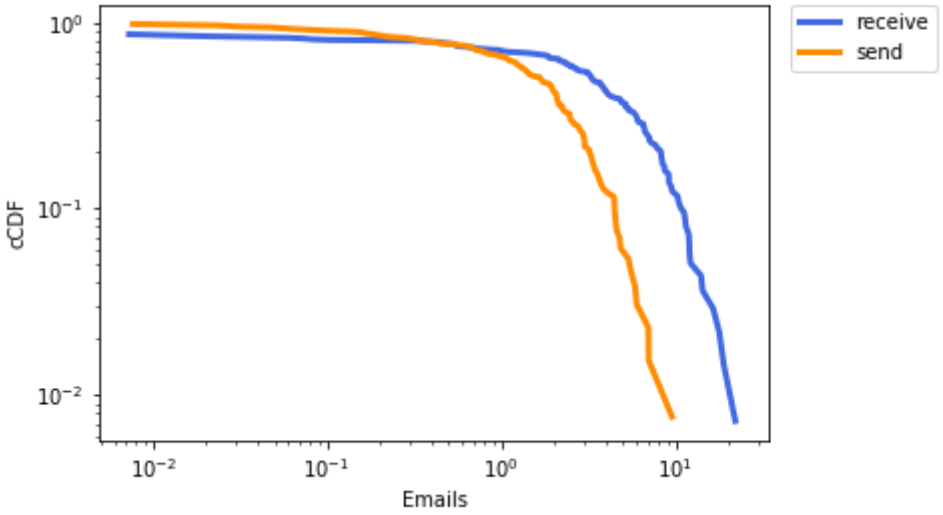


Рис.5. CCDF среднего количество отправленных и полученных в день писем для каждого пользователя

Среднее число писем, отправленных пользователем в день, составляет приблизительно 2 с 90% квантилью из примерно 6 писем в день. Эти цифры означают, что большинство корпоративных пользователей не генерируют значительную информационную нагрузку. CCDF отправленных писем экспоненциально сокращается примерно до 6 писем в день. Рассматривая сторону получателя, среднее количество полученных писем на пользователя составляет около 5 электронных писем в день с 90% квантилью из примерно 14

электронных писем в день. Соотношение полученных отправленных сообщений составляет примерно 2,5. В целом, эти наблюдения показывают, что пользователи довольно разнообразны в количестве обрабатываемых электронных писем. На **рис.6** график распределения полученных и отправленных писем за все время, показывает, что отправка электронных писем приведет к большому количеству полученных электронных писем. Например, если построить график линейной зависимости полученных от отправленных писем, то число полученных писем составляет 500 для пользователей, отправивших 200 писем.

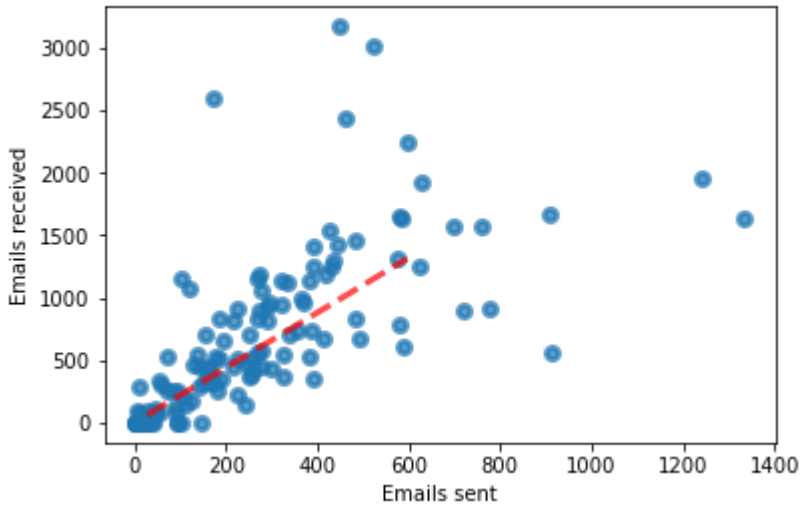


Рис.6. График распределения полученных и отправленных писем за все время

Одной из самых распространенных проблем в корпоративных почтах является наличие большого количества длинных тредов. Под тредом понимается цепочка писем между двумя или несколькими людьми на одну и ту же тему, включающие в себя в качестве отправителей и получателей тех людей, которые начали эту цепочку. Из распределения длин тредов на **рис.7** видно, что имеется достаточного много больших тредов. Каково пользовательское поведение в тредах? Для этого рассмотрим несколько графиков.

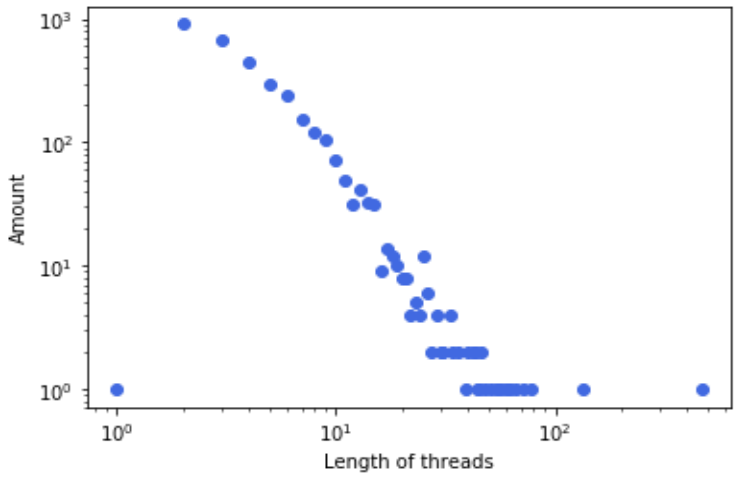


Рис.7. График распределение длин тредов

На **рис.8** слева показано распределение зависимости времени реакции ответов на письма в треде в зависимости от его длины. Распределение показывает, что чем длиннее тред, тем быстрее реагирует человек, можно предположить, что люди ведут переписки внутри корпоративной почты. На **рис.8** справа показано, что с ростом объема письма, уменьшается длина треда. Можно предположить, что длинные письма люди не читают. Из за отсутствия текстовой информации в данных компании №1 этот же вопрос будет рассмотрен в следующей части работы, в исследовании данных компании №2, где эта информация имеется.

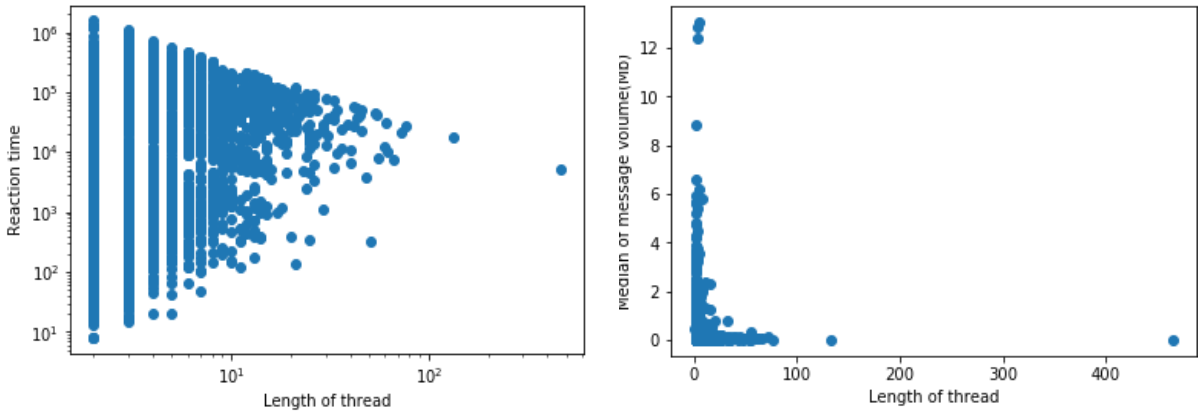


Рис.8. Распределения зависимости времени реакции ответов на письма и среднего веса письма в треде в зависимости от его длины

Почту люди используют не только для ведения переговоров, но также часто требуется пересылать файлы. Поэтому рассмотрим количество писем больших размеров, которые имеют файлы. На **рис.9** из графика слева видно, что есть множество тредов, в которых количество крупных писем приближается к размеру треда. Это говорит о том, что в письмах пересылается большое количество файлов. Как известно, в обычном почтовом ящике трудно отслеживать изменение файлов, поэтому для такой компании было бы важно иметь ресурс для более удобного всеобщего пользования файлами и отслеживания за их изменениями. График справа на **рис.9**, описывающий распределение количества больших писем в треде, показывает, что сотни файлов пересылаются несколько десятков раз. Эти показания подтверждают, что нужен инструмент для более эффективного манипулирования файлами.

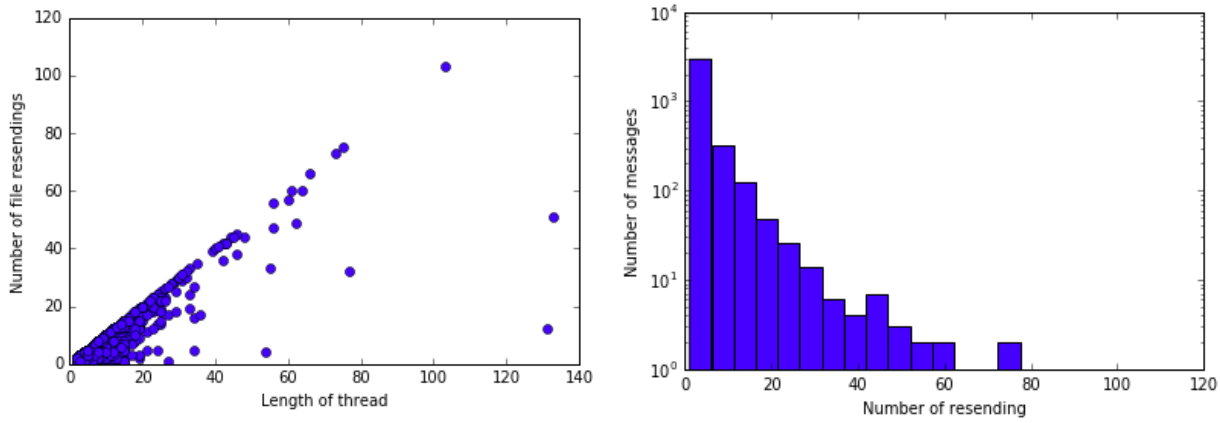


Рис.9. Распределение количества крупных писем в зависимости от длины треда и распределение количества больших писем в треде

Для более полного представления о пользовательском поведении внутри корпоративного почтового сервиса рассмотрим несколько графиков, характеризующих активность пользователя в отношении обработки электронной почты. Рассмотрим распределения интервалов времени между отправками и приемами электронной почты, то есть время, прошедшее между двумя последовательными событиями отправки и получения на пользователя соответственно. На **рис.10** показаны соответствующие CCDF распределения.

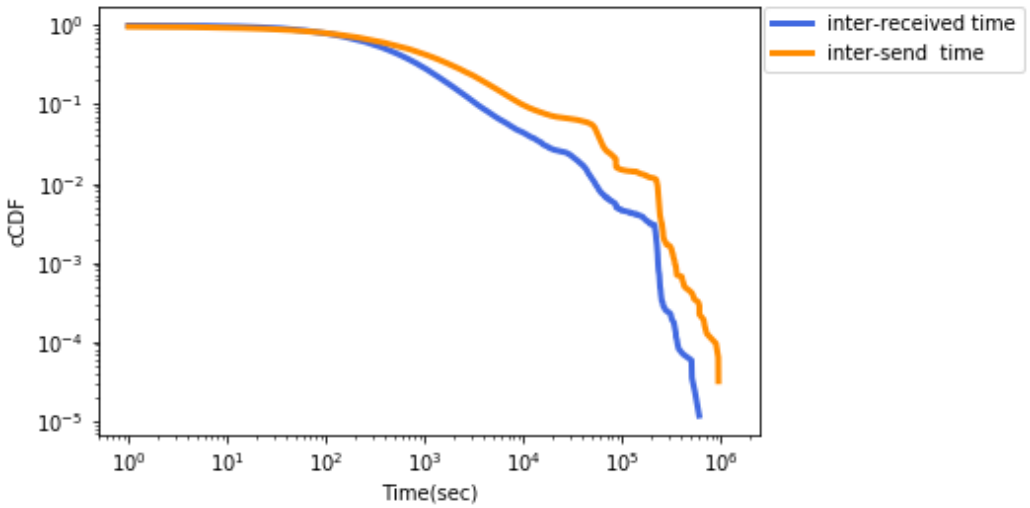
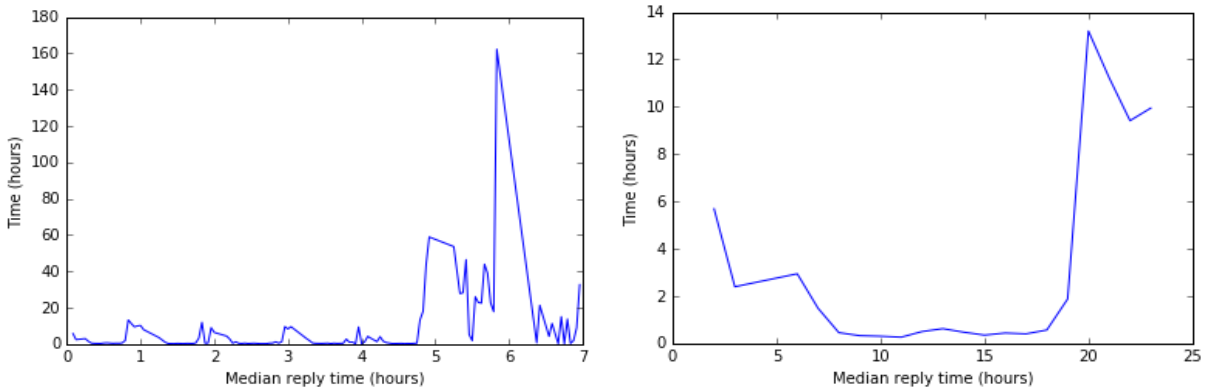


Рис.10. CCDF распределения времени между приемами и отправками сообщений

Медианное время между отправкой и получением составляет около 11 мин и 6 мин соответственно. Распределение интервалов времени между приемами сообщений сосредоточено на меньших значениях, чем распределение интервалов времени между их отправками, что является естественным порядком, поскольку отправленные электронные письма для каждого пользователя происходят от одного человека, в то время как полученные происходят в основном от нескольких лиц. Распределение времени между передачами представляет особый интерес, поскольку оно отражает деятельность человека. В частности, небольшое значение медианного времени между получениями электронной почты и существенно большее среднее время между отправками электронной почты (примерно 2 часа) показывают, что распределение интервалов времени между отправками состоит из большого количества отправок с малой длительностью и нескольких с большей длительностью. Это отражается и в CCDF распределении интервалов времени между получениями электронной почты, который проявляет медленный спад до порядка полудня и далее убывает более быстро. Качественно схожие распределения характеризуют различные аспекты человеческой деятельности [5].

Наконец, для более полного представления о пользовательском поведении, рассмотрим характер обработки электронной почты в зависимости от времени суток. **Рис.11** показывает сильную еженедельную периодичность с временем отклика, которое значительно больше для электронных писем, отправленных в конце рабочей недели или в выходные дни. В течение рабочей недели время ответа на электронные письма естественно меняется в диапазоне от 1 часа до полудня. Стоит отметить, что в рабочее время среднее время отклика менее 1 часа. Наблюдения показывают, что электронные письма имеют более высокую вероятность ответа на них в начале или середине рабочего дня.

Рис.11. Среднее время отклика пользователя на электронные письма в зависимости



от времени суток, в которое было получено исходное электронное письмо.

Итоги

Изучив подробно пользовательское поведение внутри почтового сервиса, были выявлены некоторые тенденции в поведении пользователей, а также замечены такие проблемы, как наличие длинных тредов, передача больших объемов данных, нагрузка на получателя, ведущая к более долгой реакции. Эти показатели могут быть рассмотрены как метрики эффективной коммуникации. Далее будет рассмотрена возможность нахождения лексических единиц, которые могут быть использованы как “флаги”, указывающие на возникновение одной из перечисленных проблем.

4.2. Компания №2

Корпорация Энрон (англ. Enron Corporation) — американская энергетическая компания, обанкротившаяся в 2001 году. Штаб-квартира компании располагалась в Хьюстоне (штат Техас). Акции компании торговались на Нью-Йоркской фондовой бирже под тикером ENE. До банкротства в Энрон работало около 22 000 сотрудников в 40 странах мира, и она являлась одной из ведущих в мире компаний, в таких областях как производство электроэнергии, транспортировка газа, газоснабжение, связь и целлюлозно-бумажное производство. В непроизводственном секторе компания занималась торговлей фьючерсами и производными ценными бумагами. Декларируемая выручка за 2000 год составила порядка 101 млрд долларов. Журнал Fortune назвал Энрон «самой инновационной компанией Америки» в течение шести лет подряд. В конце 2001 года стало известно, что информация о финансовом состоянии компании в значительной степени была сфальсифицирована с помощью бухгалтерского мошенничества, известного как «Дело Энрон». 2 декабря 2001 года было объявлено о банкротстве компании. С тех пор Энрон стал популярным символом умышленного корпоративного мошенничества и коррупции.

4.2.1. Используемые данные

Enron Corpus [8] - это большая база данных более 600 000 писем, созданных 158 сотрудниками корпорации Enron и приобретенных Федеральной комиссией по регулированию энергетики во время ее расследования после краха компании. Копия базы данных впоследствии была приобретена и выпущена Эндрю МакКаллум, ученым в компьютерных науках из Массачусетского университета в Амхерсте. Корпус «уникален» тем, что он является одним из общедоступных массовых коллекций «реальных» электронных писем, которые легко доступны для изучения, поскольку такие собрания обычно связаны многочисленными правовыми ограничениями, которые делают их труднодоступными.

После обзора истории компании [7] Энрон было решено использовать в исследовательской работе данные с 2000 по 2001 год. В этот период за компанией не было замечено никаких противозаконных действий и поведение людей ничем не отличалось от поведения людей в других компаниях. После предварительной подготовки, данные имеют размер в 66 тыс. строк и следующие основные признаки:

- Date - дата и время отправления письма
- From - отправитель
- To - получатели
- Subject - тема письма
- Content - содержимое письма
- MessageId - индетификатор уникального письма

На **рис.12** приведен пример исследуемых данных.

Рис.12. Пример исследуемых данных

	Date	From	To	Subject	content
0	2000-01-01 14:36:00	['sally.beck@enron.com']	['fernley.dyson@enron.com']	Happy New Year - No Y2K Fear!	We are wrapping up several hours in the office...
1	2000-01-01 19:17:00	['lenos@ucy.ac.cy']	['gordon.sick@rogroup.com']	Program attached; March NY RO Conference/Parti...	The current version of the conference program ...
2	2000-01-02 13:12:00	['andrew.parsons@enron.com']	['philippe.bibi@enron.com', 'mark.palmer@enron...	Summary of Y2K Glitches	Following please find a summary of the minor Y...
3	2000-01-03 06:17:00	['emoler@velaw.com']	['steven.kean@enron.com', 'cynthia.sandherr@en...	Options Memo	Here's the options memo you requested. Happy ...

Основные статистические показатели:

- Количество пользователей: 5500
- Всего писем: 65982
- Количество тем писем: 35908

4.2.2. Пользовательское поведение компании №2

Для данных компании Enron рассмотрим два графика. На **Рис.13** слева показано распределение объемов писем, где объемом письма является количество слов, входящих в сообщение.

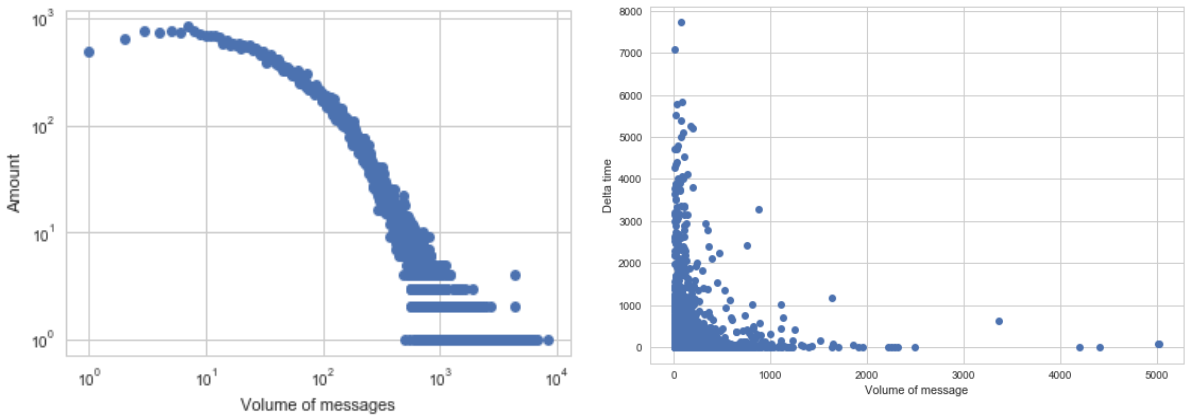


Рис.13. Распределение объемов писем и зависимость времени реакции ответа на почту от длины треда

На **рис.13** справа проиллюстрирована зависимость времени реакции ответа на почту от длины треда. Из этих графиков можем сделать вывод, что и в почтовом сервисе компании Enron, присутствовали такие проблемы как длинные треды и объемные письма. Используем эту информацию и наличие текстовых сообщений для дальнейшего исследования.

4.2.3. Исследование взаимосвязи текста сообщений и метрик

Ранее было установлено, что уменьшение таких характеристик как объем сообщения, длина треда и время ответа на письмо ведет к более эффективным коммуникациям внутри компании. Но было бы интересно посмотреть, какие темы ведут к длинным перепискам, о чем пишут в больших сообщениях, и что заставляет людей отвечать долго. Таким образом были проведены исследования по решению задач поиска корреляций этих характеристик по

тексту сообщений. Далее описаны методы, которые показали наилучшие результаты в решении поставленных задач. На **рис.14** изображено облако слов, визуализирующие более частые темы, на которые говорят в сообщениях корпуса Enron.

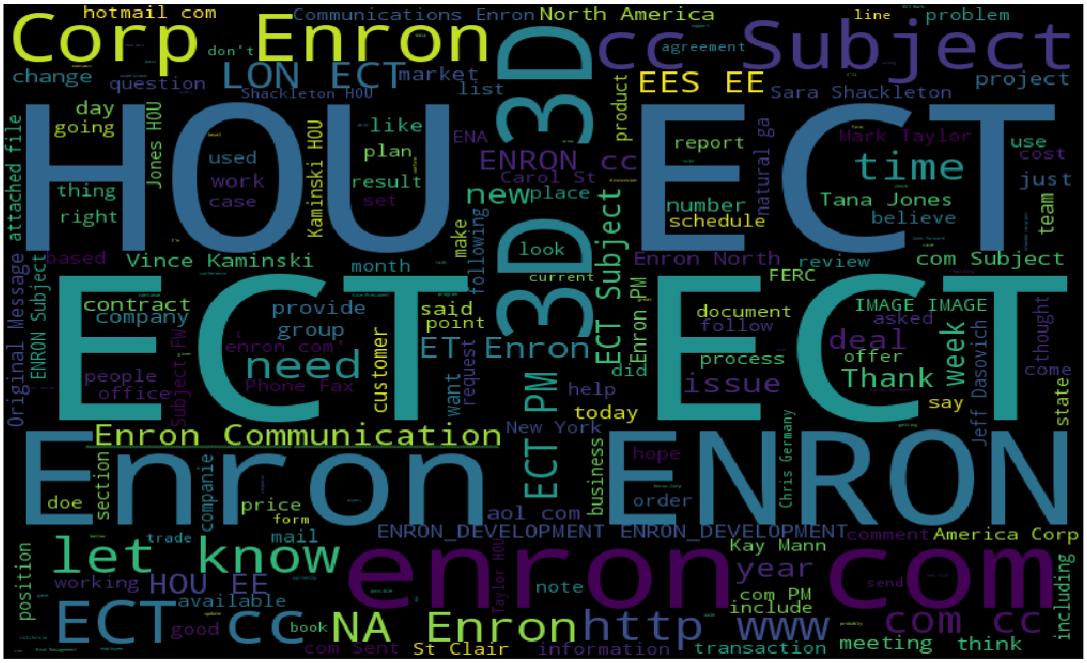


Рис.14. Визуализация частотности слов в корпусе Enron

4.2.3.1. Предобработка текста.

Предобработка текста является самой трудоемкой и важной частью в задачах автоматического анализа текста. Этот этап не имеет единого алгоритма для всех задач, связанных с анализом текстов, он зависит от типа извлекаемой информации. Одной из главных проблем анализа текстов является большое количество слов в документе. Если каждое из этих слов подвергать анализу, то время по-иска новых знаний резко возрастает и вряд ли будет удовлетворять требовани-ям пользователей. В то же время очевидно, что не все слова в тексте несут полезную информацию. Кроме того, в силу гибкости естественных языков формально различные слова (синонимы и т. п.) на самом деле означают оди-наковые понятия. Таким образом, удаление неинформативных слов, а

также приведение близких по смыслу слов к единой форме значительно сокращают время анализа текстов. Были проведены следующие действия на этапе предварительной обработки текста:

- Токенизация - это разбиение непрерывной строки на отдельные «слова» по определенным правилам:
    - приведение текста к нижнему регистру
    - замена всех знаков препинания и прочих символов на пробелы
    - объявление токенов, которыми в данном случае являются униграммы, биграммы и триграммы.
- Удаление стоп-слов. Стоп-словами называются слова, которые являются вспомогательными и несут мало информации о содержании документа. В случае имеющихся данных это такие слова как “at,” “it” , “here.”.
- Лемматизация — морфологический поиск. Он заключается в преобразовании каждого слова к его нормальной форме. В первую очередь слово проверяется по словарю. Если оно там есть, то оно приводится к указанной ему форме. Иначе по определённому алгоритму выводится способ изменения данного слова, на основании него делаются выводы о начальной форме.

4.2.3.1. Извлечение признаков из текста

Было проведено несколько способов формирования векторов-признаков и два из них показали примерно одинаковые результаты. Самым простым и быстрым из них оказался основанным на извлечении признаков из текста с помощью TF-IDF метода. О нем и будет рассказано ниже.

Каждое сообщение является документом, так, если слово часто встречается в тексте, и оно не является стоп-словом, то, скорее всего, оно важно. Но есть и вторая тонкость. Если слово встречается в других документах реже, чем в данном, то и в этом случае, скорее всего, оно важно для текста. По этому слову можно отличить этот текст от остальных. Такой учет слова

позволяет сделать TF-IDF метод, который позволяет извлечь значение признака для слова wи текста xпо следующей формуле:

$$TF - IDF(x, w) = n_{dw} log \frac{l}{n_w}$$

Эта формула состоит из двух частей. n_{dw}- доля вхождений “слова” w в документ d. Если “слово” часто встречается в тексте, то оно важно для него, и значение признака будет выше. Второй множитель называется IDF (inverse document frequency, обратная документная частота), это отношение l, общего количества документов, к n_w, числу документов в выборке, в которых слово w встречается хотя бы раз. Если это отношение велико, “слово” редко встречается в других документах, значение признака будет увеличиваться. Если слово встречается в каждом тексте, то значение признака будет нулевым ($log \frac{l}{l} = 0$).

Формирование матрицы обучения

В данном исследовании вместо “слов” используются униграммы, биграммы и триграммы. Признаками являются TF-IDF-коэффициенты, вычисленные для каждого токена. Количество сформировавшихся признаков с помощью метода, который описан выше, получается равным примерно 600 тыс. Таким образом, каждое сообщение является вектором размером примерно 600 тыс. элементов. Для понижения размерности может быть использован метод главных компонент [6], но это ресурсоемкий процесс.

Альтернативный метод извлечения признаков и формирования матрицы

В качестве альтернативного метода формирования матрицы для обучения, может быть использован TF-IDF метод в сочетании с программным инструментом анализа семантики естественных языков word2vec [13]. Он основан на обучении нейронных сетей. Word2vec принимает большой текстовый корпус в виде всех сообщений, которые рассматриваются, в качестве входных данных и сопоставляет каждому слову вектор, выдавая координаты

слов на выходе. Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (а, следовательно, имеющие схожий смысл), в векторном представлении будут иметь близкие координаты векторов-слов.

Для каждого сообщения составляется вектор следующим образом:

1. Для каждого слова определяется его вектор из обученной модели word2vec.
2. TF-IDF- коэффициент слова умножается на его вектор.
3. Все получившиеся вектора в сообщении складываются.

Минус в этом методе заключается в его ресурсоемкости, плюс - на выходе получается матрица намного меньшей размерности, чем при методе TF-IDF. В целом, конечный результат в задачах предсказания практически не меняется.

4.2.3.2. Предсказание "хвоста" треда по тексту

В этом разделе описывается подход для нахождения слов и фраз в сообщениях, влекущих за собой длинные треды. Для этого решается задача предсказания длины "хвоста" треда по тексту сообщения. И для начала, в данных были выделены треды длиной не менее двух сообщений. Так, для данной задачи используются данные размером в 7818 строк, содержащие 2918 тредов.

Чтобы найти слова и фразы, которые предвещают длинную переписку, необходимо сгенерировать значения целевой переменной для их предсказания по тексту, который интерпретирует матрица, описание формирования которой рассмотрено в пункте *4.2.3.1.* Значения целевой переменной генерируются следующим образом. В заранее отсортированных по времени отправки сообщений данных, рассматривается каждый тред, в котором каждому сообщению (сверху-вниз) присваивается значение $L - n$, где L - длина треда, $n \in [1, L]$. На **рис.15** показан пример разметки целевой переменной "tail" на

одном треде.

	clean_content	tail
230	go figuring position far	6
231	know mean position defunct	5
239	idea rely tell u position	4
243	without calling respective trader know position	3
8540	tana last day please let know limit changed ge...	2
8543	letter way mike moscoso right interoffice mail...	1
8544	many thanks tc enron north america corp tana j...	0

Рис.15. Пример целевой разметки для предсказания “хвоста” треда по тексту сообщения

Достаточно неплохой результат предсказания “хвоста” треда по тексту показала гребневая регрессия, которая применялась из-за способности отбирать признаки, в силу наличия регуляризатора [9]. На **рис.16** продемонстрированы истинные значения объектов на отложенной выборке, размер которой равен четверти всех рассматриваемых данных, и предсказанные моделью значения этих объектов. График показывает, что предсказанные значения хорошо повторяют распределение истинных так, что среднеквадратичное отклонение получается равным 1,704. Но мы также можем заметить, что предсказанные значения распределяются по трем интервалам определенным образом.

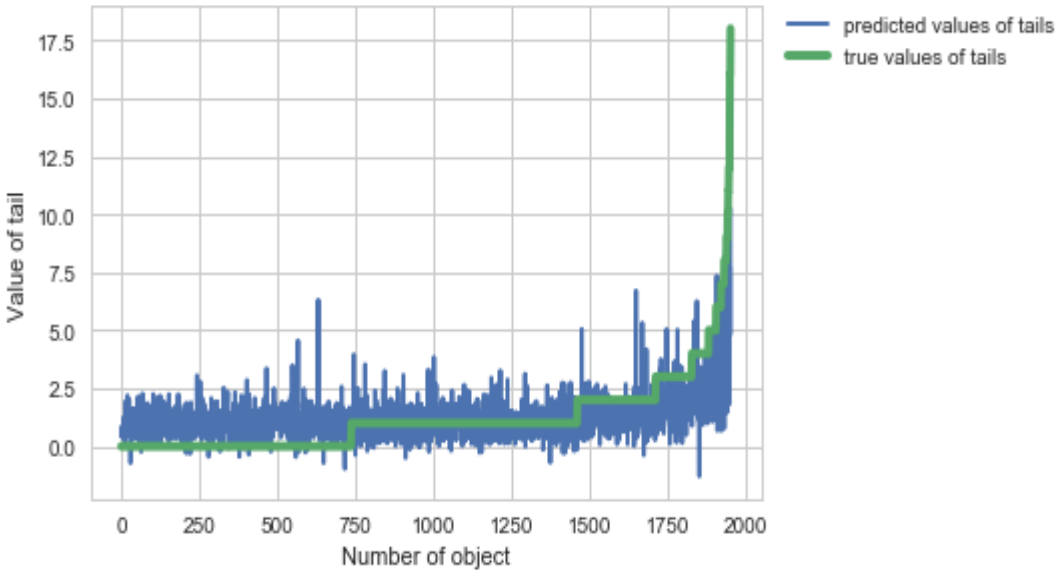


Рис.16. Истинные и предсказанные значения “хвостов” тредов на отложенной выборке

Например, на объектах 0 - 750 предсказанные значения имеют один характер распределения, на объектах 750-1500 - другой, более приближенный к истинным значениям. Это наводит на мысль, что можно разбить значения “хвостов” тредов на диапазоны, задав им категории и решив задачу классификации.

Для решения задачи классификации удобнее всего использовать логистическую регрессию, так как это позволит легко посмотреть на признаки и их коэффициенты, отвечающие с большей или меньшей степенью за тот или иной класс. Таким образом, было найдено такое разбиение целевых значений, при котором модель мультиклассовой логистической регрессии давала наилучшее качество. Качественные результаты итерационного подбора наилучших показателей продемонстрированы в **табл.1**.

диапазоны оптимальных длин тредов для классификации

1 категория	2 категория	3 категория
L≤4	4<L≤ 9	L>9

точность на отложенной выборке

0.98	0.73	0.82

полнота на отложенной выборке

0.99	0.58	0.65

accuracy_score = 0.96

Табл.1. Результаты решения задачи классификации по предсказанию категории

В качестве альтернативного варианта применяемой модели может быть метод опорных векторов с полиномиальным ядром [12]. Этот метод дает ненамного результаты лучше, чем мультиклассовая логистическая регрессия, но для того, чтобы получить признаки и их значимость, потребуется сделать несколько дополнительных действий: вычислить опорные вектора и провести итеративный процесс вычисления расстояний между опорными векторами и полученными признаками.

В **табл.2** приведены лексические единицы, имеющие наибольшие коэффициенты, указывающие на принадлежность к тому или иному классу.

thanks	fabian answer question	really enjoyed
hou ect[1]	know mean position	lunch tommorow still
requirement meeting	check broker monday	great see wed
enron broker fee	missing deal buy	fun play tues
bank need turn	lunch work	western hub best
financial transaction	next wednesday vacation	daughter college

Табл.2. Лексические единицы, предсказывающие диапазон длины треда

[1] Enron Capital and Trade Resources (ECT).
Houston Natural Gas (HOU)

В таблице можем увидеть, что в коротких тредах содержаться письма преимущественно делового характера. Разговоры, касающиеся работы могут быть в тредах длиной от 2 до 5 сообщений. Далее, чем длиннее тред, тем более неформальные беседы ведутся в переписках.

4.2.3.3. Предсказание объема сообщения по его тексту

Для того, чтобы узнать, о чем пишут в больших сообщениях, была решена регрессионная задача предсказания объема текста, выражающегося в количестве слов в сообщение, по его тексту. Так, кроме предварительной обработки данных и текста, рассмотренных выше, никаких дополнительных преобразований не требуется. Для более быстрой работы алгоритмов были взяты данные, которые использовались в пункте *4.2.3.2.*, то есть размер данных составляет 7818 строки.

Итак, генерация целевых переменных не составляет труда: каждому сообщению присваивается значение, равное количеству слов этого сообщения. Матрица признаков - интерпретация текста методом TF-IDF (пункт *4.2.3.1.).*

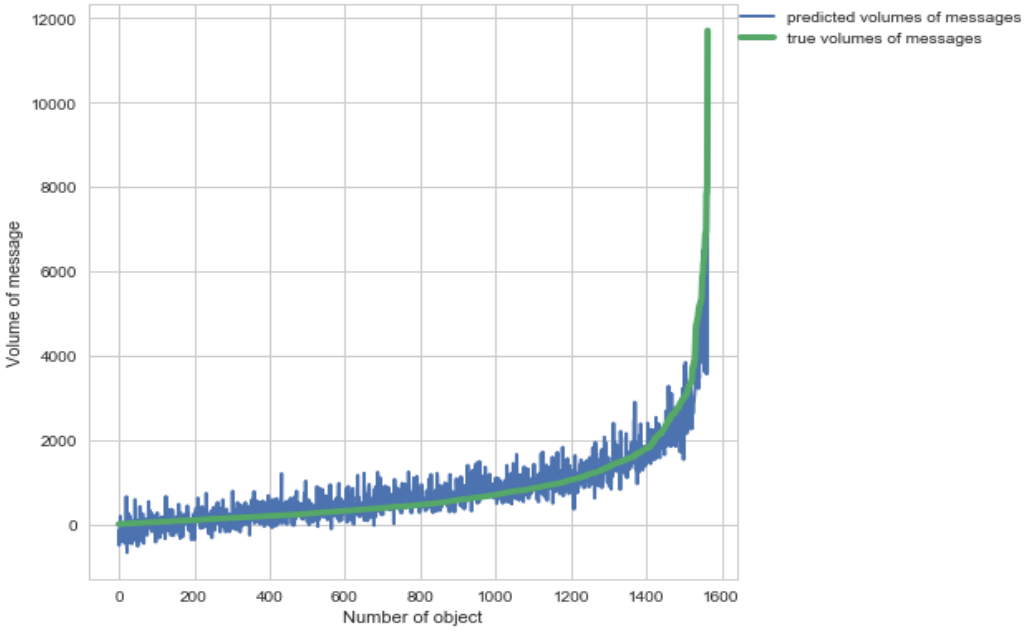


Рис.17. Истинные и предсказанные значения объемов писем на отложенной выборке

Используя и в этой задачи модель гребневой регрессии с подобранными

параметрами, получим, изображенную на **рис.17**, картину распределения истинных и предсказанных значений на отложенной выборке, составляющей четверть всей выборки. Распределение предсказанных значений снова достаточно неплохо приближено к истинным, но на объектах 1700 - 2000, которые имеют большие значения, модель срабатывает плохо и, соответственно, получается большое среднеквадратичное отклонение. Это не удивительно, проанализировав гистограмму распределения целевой переменной на **рис.18**, увидим, что она описывает степенной закон.

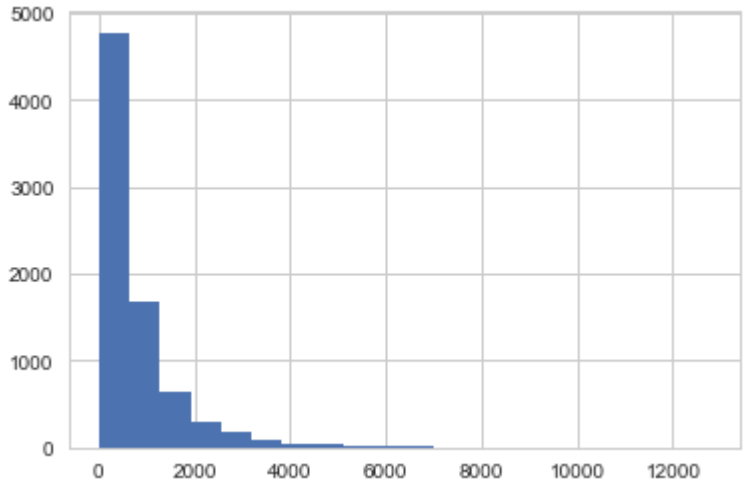


Рис.18. Гистограмма распределения объемов сообщений

Чтобы приблизить распределение данных к нормальному закону, воспользуемся однопараметрическим преобразованием Бокса-Кокса (раздел 3.3.). Формула однопараметрического преобразования зависит от параметра лямбда:

$$x_i(\lambda) = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(x_i), & \lambda = 0 \end{cases}$$

В данном случае, $\lambda = 0.148$. Результат преобразования изображен на **рис.19.**

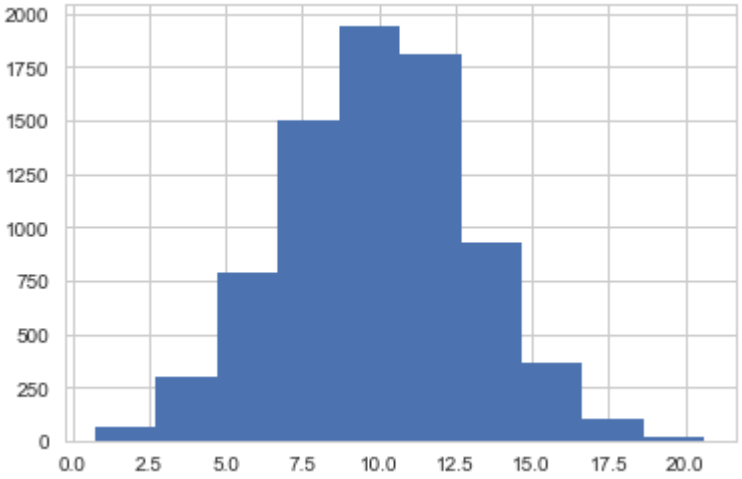


Рис.19. Гистограмма распределения объемов писем после Бокс-Кокс преобразования

На **рис.20** продемонстрировано распределение истинных и предсказанных значений, полученных с помощью гребневой регрессии после изменения целевой переменной. Среднеквадратичная ошибка на отложенной выборке равна 0.61.

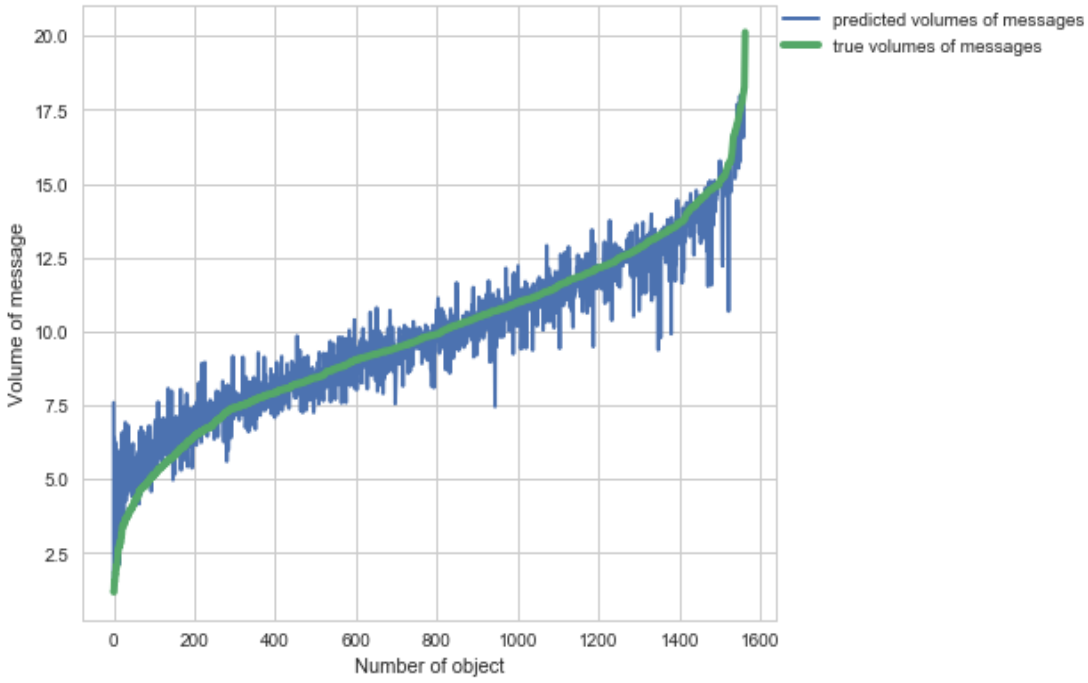


Рис.20. Истинные и предсказанные значения объемов писем на отложенной выборке после преобразования целевой переменной

Стоит отметить, что если основная цель заключается именно в предсказании объема, то после получения предсказанных моделью значений, необходимо сделать пересчет, используя ту же формулу преобразования Бокса-Кокса и полученное ранее значение λ. Но в данном случае преследуется другая цель - поиск лексических единиц, которые мы и получили и которые представлены в **табл.3.**

forwarded	6.99736	guy view	-3.44155
scheduled attend	4.270511	fine	-3.36713
development enron	4.195731	call around	-3.26370
gas	4.15108	update yet	-3.25324
business	4.14826	heard anything	-2.91181
company	3.78439	deal ticket	-2.71791
team	3.70668	sound great	-2.56855
meeting profile	3.48081	game week	-2.51890

Табл.3. Лексические единицы и их коэффициенты, предсказывающие большой объем сообщения

Итоги

Из таблицы **табл.3**. можно сделать вывод, что люди часто использует пересылку (forwarded) и в больших письмах пишут на темы рабочего характера. Чем меньше письмо, тем более неформальный формат общения оно несет.

Отметим, что неформальный формат общения также наблюдался и в больших тредах. Это соотносится с наблюдаемой ранее характеристикой, где с увеличением длины треда уменьшается объем письма.

Но в **табл.3** и **табл.2** наблюдаются слова свойственные только корпоративной культуре компании Enron. Таким образом нельзя утверждать,

что найденные слова являются индикаторами длинных тредов и объемных писем в любой компании. Чтобы определить лексические единицы, инициирующие, указанные метрики более точно, необходимо провести дополнительную работу для выявления дополнительных “стоп-слов”, характерных для рассматриваемой организации.

5. Выводы.

В данной исследовательской работе удалось изучить пользовательское поведение внутри корпоративной почты двух компаний. Кроме того, были обнаружены схожие закономерности. Были выделены такие проблемы как большое количество тредов и больших писем. Это влечет за собой загрузку электронной почты и нагрузку на пользователя внутри почтового сервиса, что ведет к неэффективному коммуницированию и ухудшению рабочего процесса.

Также удалось убедиться в том, что описанные методики автоматической обработки текста, могут использоваться для поиска лексических единиц, в качестве индикаторов каких-либо событий или могут являться дополнительной информацией для понимания некоторых процессов и явлений. Но для более точных результатов поиска таких слов и фраз, которые могли бы применяться как метрики качества и в других компаниях, требуется дополнительная работа.

6. Список используемых материалов:

[1] What Email Reveals About Your Organization. Peter A. Gloor. Winter 2016.

[2] Understanding machine learning algorithm. From Theory to Algorithms. Shai Shalev-Shwartz The Hebrew University, Jerusalem Shai Ben-David University of Waterloo, Canada. 2014.

[3] https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

[4] Computational Statistics. Gentle, J.E. 2009

[5] K. J. Delaney and V. Vara. Will Social Features Make Email Sexy Again? The Wall Street Journal, October 18, 2007.

[6] http://setosa.io/ev/principal-component-analysis/

[7] B. Klimt and Y. Yang. Introducing the Enron corpus. In First conference on email and anti-spam (CEAS), 2004.

[8] https://www.cs.cmu.edu/~./enron/

[9] Email Information Flow in Large-Scale Enterprises Thomas Karagiannis and Milan Vojnovic Microsoft Research May 2008 Technical Report MSR-TR-2008-76

[10] Networks, Crowds, and Markets: Reasoning about a Highly Connected World David Easley Dept. of Economics Cornell University Jon Kleinberg Dept. of Computer Science Cornell University

[11] T. Karagiannis, J.-Y. L. Boudec, and M. Vojnovic. Power Law and Exponential Decay of Inter Contact Times between Mobile Devices. In Proc. of ACM Mobicom 2007, pages 183–194, Montreal, Canada, 2007

[12] Training and Testing Low-degree Polynomial Data Mappings via Linear SVM. Yin-Wen Chang. Cho-Jui Hsieh. Kai-Wei Chang

[13] https://code.google.com/archive/p/word2vec/

[14] T. Joachims. Text categorization with support vector machines: Learning with many relevant features.Machine Learning: ECML-98, pages 137–142, 1998.

[15] J. Hancock, C. Landrigan, and C. Silver. Expressing emotion in text-based communication. In Proc. CHI

[16] J. Holmes and S. Schnurr. Politeness, humor and gender in the workplace: negotiating norms and identifying contestation. Journal of Politeness Research. Language, Behaviour, 2005.