МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Кафедра дискретной математики и алгоритмики

АНАЛИЗ ДЕЛОВОЙ ПЕРЕПИСКИ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Курсовой проект

Гулина Кирилла
Ивановича
студента 4 курса,
специальность
“информатика“

Научный руководитель:
Свирид Юрий
Владимирович

Минск, 2020

Содержание

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
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 [1]. В данной работе мы будем основываться именно на датасете, опубликованном *Kaggle*.

Глава 2
Предобработка данных

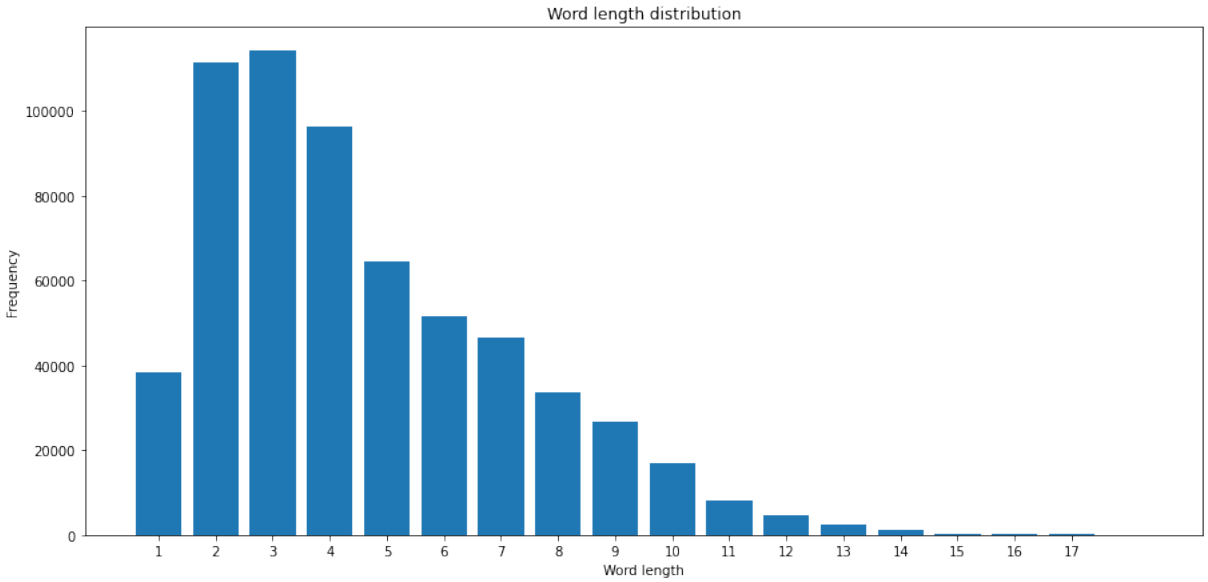
Машинное обучение является мощным и эффективным инструментом при реализации алгоритмов классификации, маршрутизации, обработки и поиска документов, однако, определяющее значение в этих процессах имеет качество исходных данных [2]. Именно поэтому проведение подготовки исходных документов, их предварительная обработка позволяет значительно повысить точность результатов, получаемых в ходе применения машинного обучения. Разобьем этот этап 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.

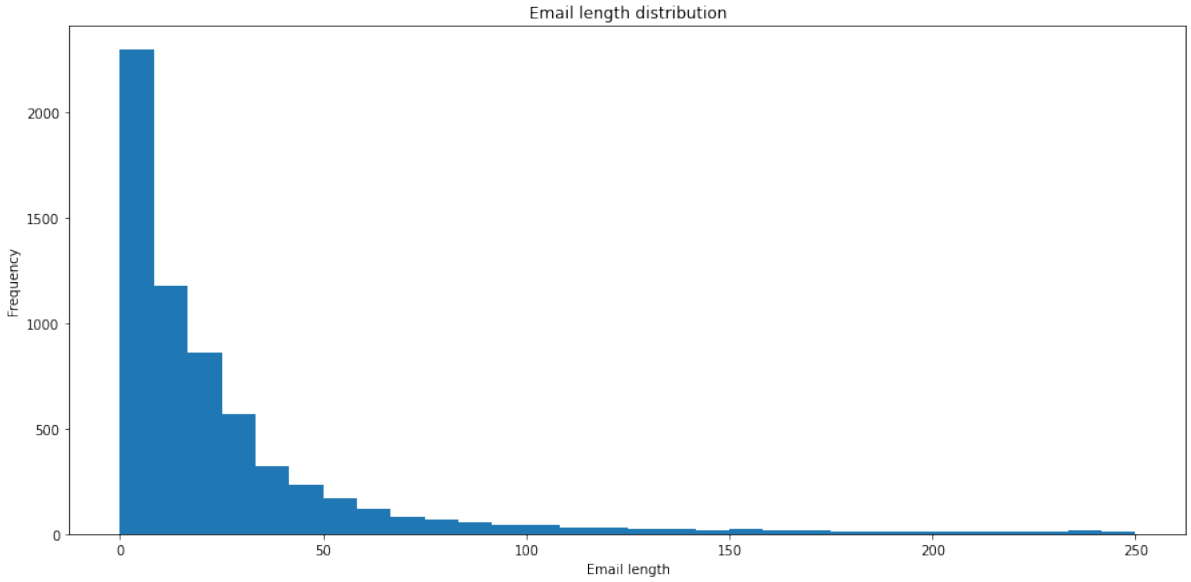
2. В текстовом файле, взятом из *Kaggle*, могут быть пропущенные данные (например, в связи с плохим качеством pdf-файла). Такие данные пропускаются и нами не обрабатываются. После осуществления этого шага остается 6742 писем.

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* [3], 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.

Глава 3
Исследование данных

3.1 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

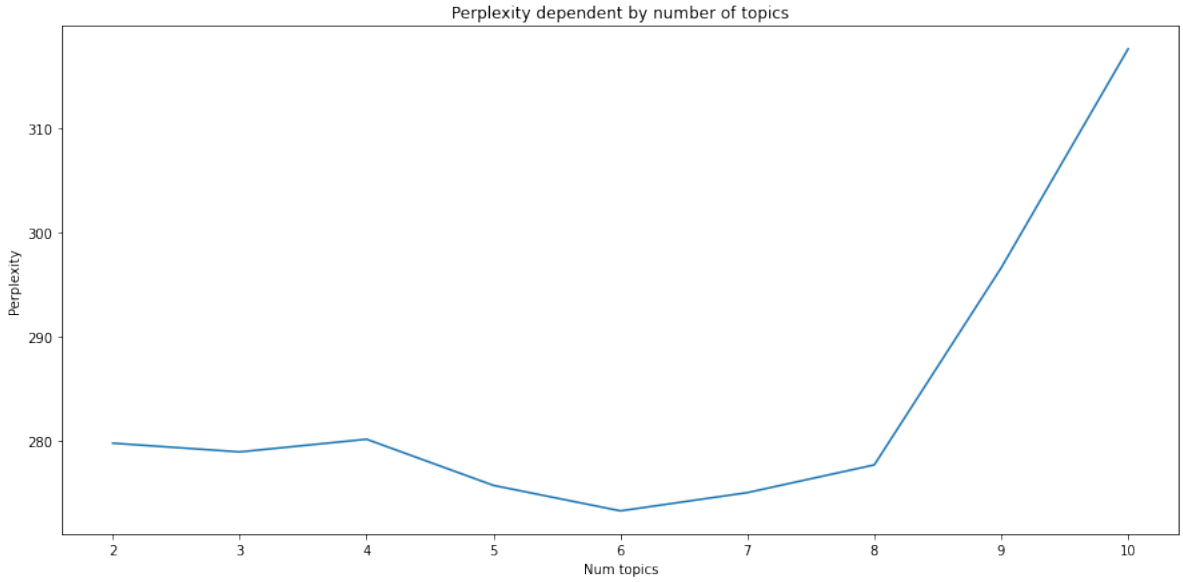
- число тем;
- распределения частот слов, характерное для каждой темы;
- тематику каждого документа — в какой степени он относится к каждой из тем.

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*) [4].

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.

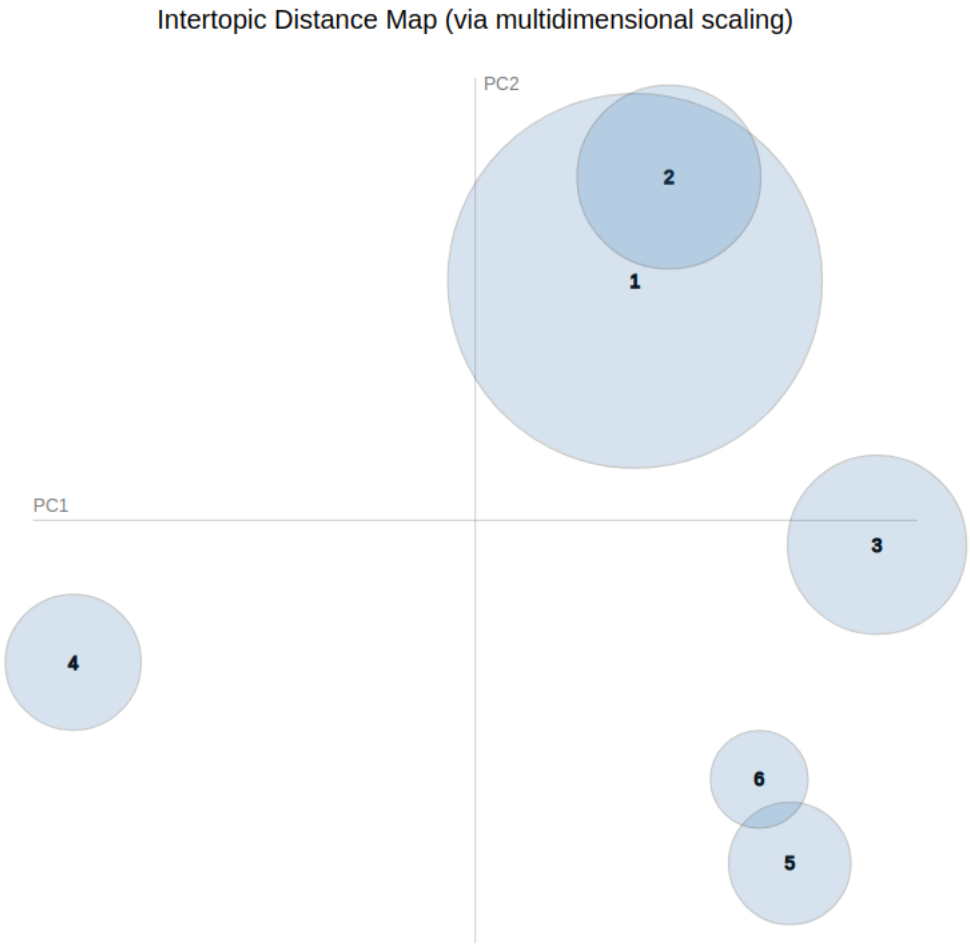
В зависимости от параметра модели, отвечающего за количество тем у распределения текстов, получилась следующая зависимость значения перплексии от количества тем:



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:



3.2 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* [5].

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 [6], 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.

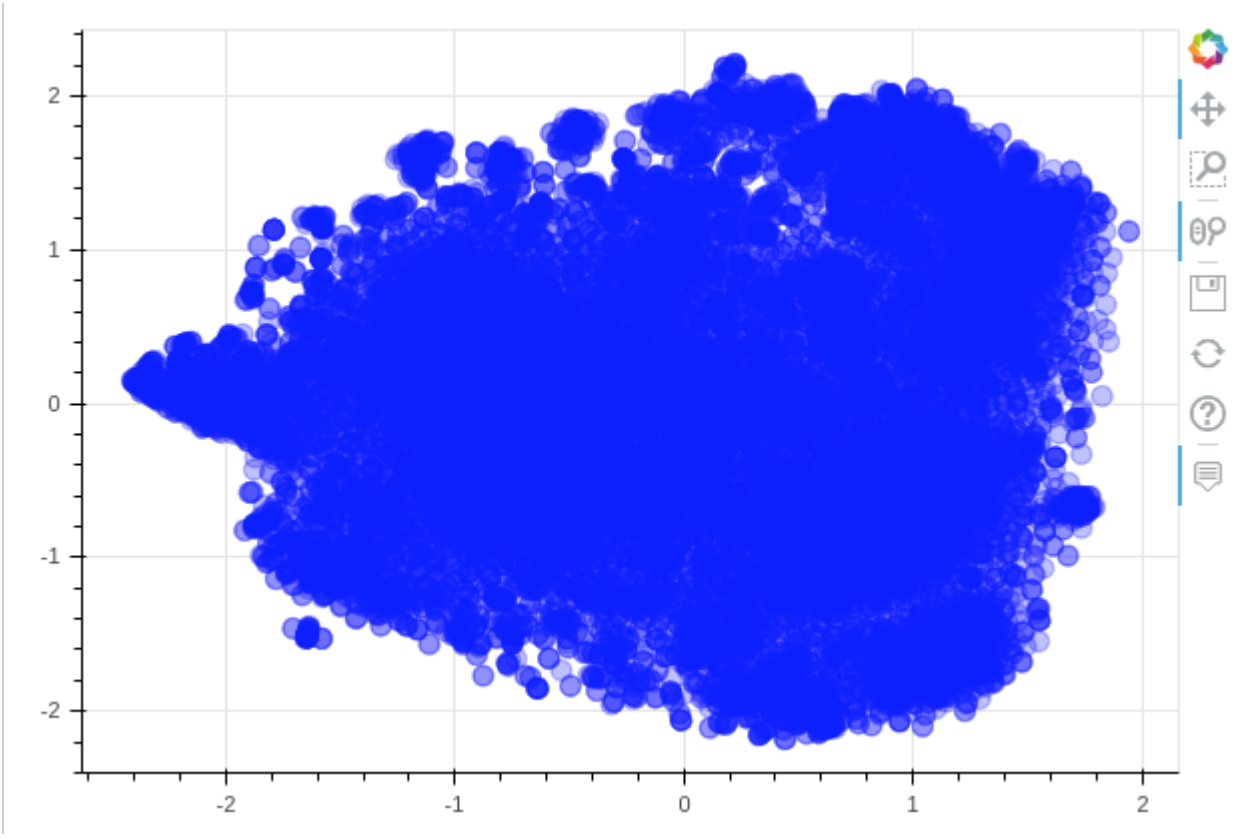
Результаты работы алгоритма. Ближайшие слова к «*obama*»:

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* [7]. *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. Hillary Clinton's Emails, https://www.kaggle.com/kaggle/hillary-clinton-emails.

2. 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.

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

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

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

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

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