Санкт-Петербургский государственный университет
Кафедра технологии программирования

Башарин Егор Валерьевич

Выпускная квалификационная работа бакалавра

Контекстная обработка данных социальных сетей

Направление 010400
Прикладная математика и информатика

Научный руководитель,
старший преподаватель
Попова С.В.

Санкт-Петербург
2016

Содержание

Введение

В настоящее время явление социальных сетей достаточно распространено. Социальные сети уверенно вошли в жизнь современного человека и теперь занимают в ней значимую часть. Главным образом они оказывают влияние на поведение, предубеждения, ценности и намерения человека, что отражается во всех сферах его деятельности. Оказываемое влияние, быстрый рост популярности и открытый доступ к контенту привлекли к социальным сетям внимание правительства, финансовых организаций и исследователей. Выделение ключевых концепций стало важным условием для порождения знаний и формулирования стратегий. Анализ полученных данных помогает исследователям улучшить понимание об информационных потоках, о формировании и распространении мнений, о связи ценностей и предубеждений пользователя и генерируемого им контента. Существенным барьером при использовании социальных сетей является необходимость выбора методологии для сбора, обработки и анализа информации, полученной с сайтов социальных сетей. Однако, существуют компании по производству программного обеспечения, разрабатывающие проприетарные системы сбора информации для визуализации данных, и исследователи, занимающиеся разработкой экспертных систем для анализа настроений [1].

Пользователи социальных сетей ежедневно публикуют данные о своей активности, чувствах и мыслях, выражая свое мнение и позицию. Это способствует появлению в социальных сетях групп пользователей (сообществ), имеющих общие интересы. Для выявления ключевых концепций и тематик присущих группе пользователей используется контекстная обработка генерируемого ими контента. В данной работе контекстная обработка данных основана на идеях и принципах тематического моделирования. Результаты такой обработки могут использоваться для мониторинга мнений и политических взглядов пользователей или для предсказания поведения рынка.

Постановка задачи

Целью данной работы является изучение методов контекстной обработки данных социальных сетей, в основе которых лежат принципы и идеи тематического моделирования. Под социальной сетью понимается веб-сайт или онлайн-сервис, который предназначен для поддержания социальных взаимоотношений при помощи Интернета.

Для того чтобы достичь поставленной цели предлагается выполнить следующий ряд задач:

1. Выбор источника данных
2. Загрузка и предварительная обработка данных
3. Выбор тематической модели
4. Построение тематической модели
5. Оценка качества модели

Обзор литературы

Тема данной работы тесно пересекается с информационным поиском, основы которого подробно рассмотрены в книге Кристофера Майнинга "Introduction to Information Retrieval" [3]. Особое внимание стоит уделить главам 2 и 18. В главе 2 описываются методы подготовки и предварительной обработки текстовой информации. Глава 18 сосредотачивает внимание на подходах латентно-семантического анализа, который является ценным инструментом в тематическом моделировании. В конце каждой главы приведены ссылки на литературу для более подробного изучения темы.

Вероятностное латентно-семантическое моделирование стало логичным продолжением идей латентно-семантического моделирования и нашло свое применение в тематическом моделировании. Это стало причиной появления вероятностных тематических моделей. Основные принципы вероятностного латентно-семантического анализа (probabilistic latent semantic analysis - pLSA) были описаны Томасом Хоффманом в 1999 году в статье [4]. Затем они были развиты Дэвидом Блеем в его статье 2003 года [5], в которой была введена и рассмотрена тематическая модель латентного размещения Дирихле (latent dirichlet allocation - LDA). Статья Д.Блея описывает основные преимущества LDA перед pLSA, а также методы построения и оценки качества тематической модели LDA. В статье Д.Блея 2012 года [7] рассматриваются связь LDA с другими вероятностными тематическими моделями, а также применение LDA в тематическом моделировании.

В техническом отчете Грегора Хейнрича "Parameter estimation for text analysis" [6] рассматриваются методы оценки параметров моделей для тематического анализа текстов. В отчете подробно разобраны темы, связанные с основными подходами оценки параметров, сопряженными распределениями и Байесовскими сетями, а также применение данных тем для построения тематической модели LDA.

Среди русскоязычной литературы следует обратить внимание на работы К. В. Воронцова. В работе [8] подробно описаны основные идеи вероятностного тематического моделирования. В первой части данной работы ставится задача тематического моделирования. Далее рассматриваются основные вероятностные тематические модели pLSA, LDA и их модифицированные версии, а также методы их построения и оценки.

Глава 1. Подготовка данных

1.1 Обзор социальных сетей

Несмотря на то, что социальные сети появлись около 20 лет назад, их популярность растет с каждым годом. На рисунке 1 показан график, отображающий рост числа пользоватлей социальных сетей во всем мире. По итогам 2015 года число пользователей социальных сетей превысило отметку в 2 миллиарда человек и по прогнозам их количество будет только расти [11]. Поэтому можно сделать вывод, что социальные сети прочно укрепляются в жизни совеременного человека, а их изучение становится актуальной проблемой.

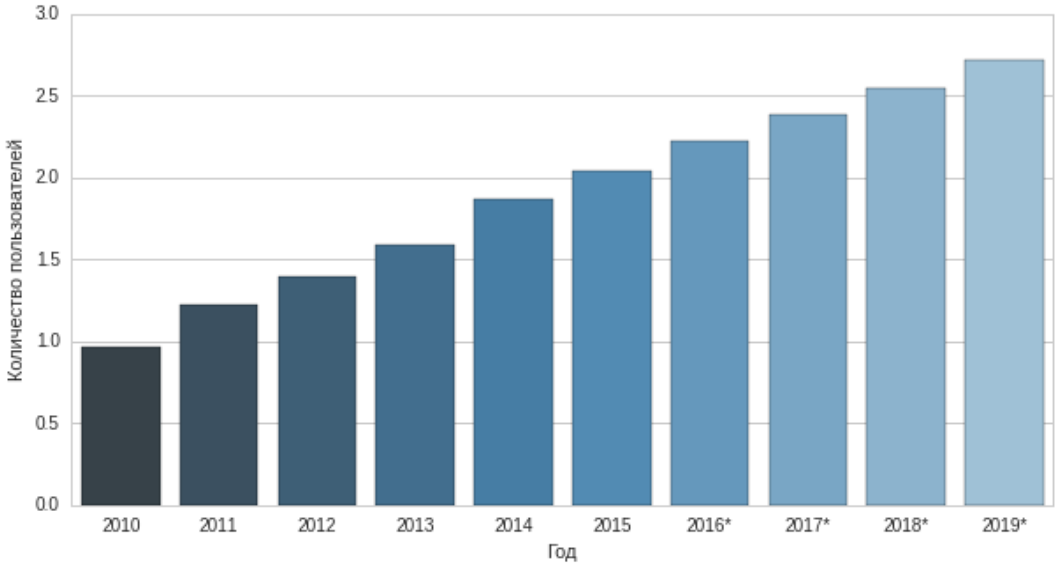

Рис. 1: Число пользователей социальных сетей по годам

Число социальных сетей довольно велико, и каждая из них предоставляет различные возможности для пользователей и преследует различные цели. На рисунке 2 представлен график, отражающий количество активных пользователей в самых популярных социальных сетях на апрель 2016 года [12]. На графике видно, что такие социальные сети как Facebook, WhatsApp, Facebook messenger и QQ пользуются наибольшей популярностью у пользователей. Также стоит обратить внимание на социальную сеть VKontakte, которая довольно популярна в российском сегменте интернета и насчитывает около 100 миллионов активных пользователей.

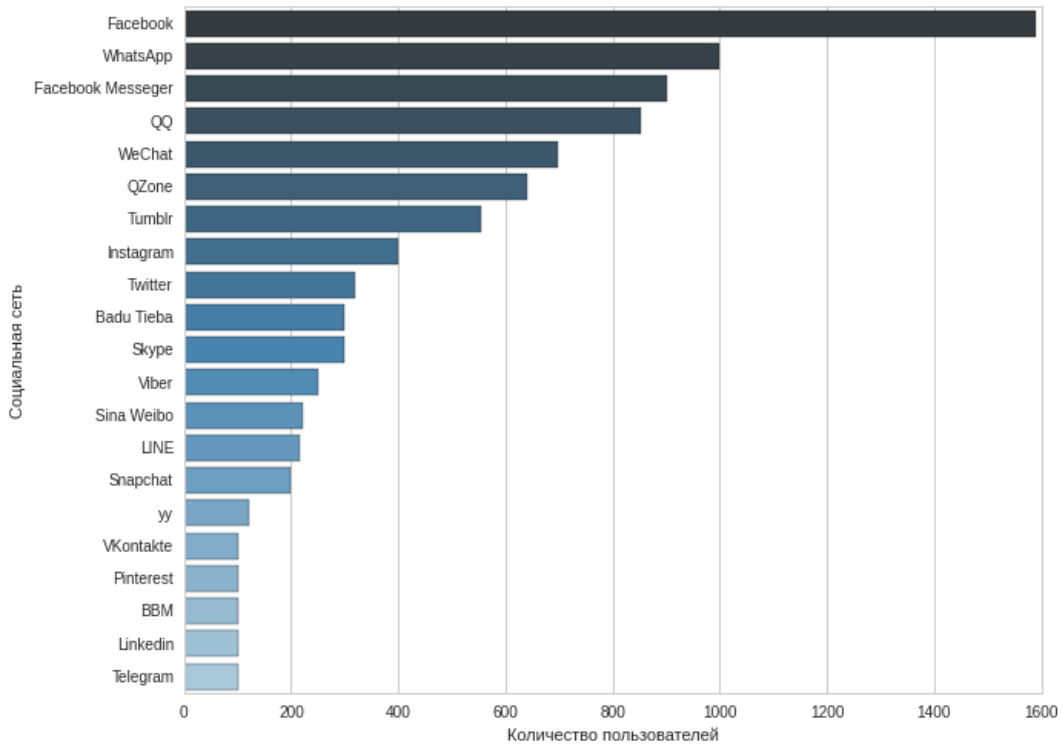

Рис. 2: Рейтинг самых популярных социальных сетей на апрель 2016 года

Социальные сети Facebook и Vkontakte предоставляют похожие возможности своим пользователям: создание профиля с фотографией и информацией о себе, обмен сообщениями с другими пользователями, создание сообществ, публикация сообщений на страницах других пользователей или сообществ, загрузка видеозаписей и фотографий и множество других функций для взаимодействия между пользователями. Такие социальные сети как WhatsApp, QQ, WeChat, Skype, Viber, Telegram в основном выполняют роль месседжеров и их предназначенение ограничивается обменом текстовой, аудио- и видео- информации между пользователями. Социальная сеть Instagram ориентирована на публикацию пользователями фотографий и видеозаписей. Особенность социальной сети Twitter - это возможность публикации коротких сообщений. Linkedin представляет собой социальную сеть, предназначенную для поиска и установления деловых связей.

1.2 Выбор социальной сети и загрузка данных

В качестве исходных данных рассмотрим публикации в сообществах социальных сетей. Такие сообщества, как правило, представляют собой одну или несколько веб-страниц. Каждое сообщество обладает определенной

тематической направленностью: спорт, музыка, политика, финансы и др. Возможность создания сообществ поддерживается такими социальными сетями как Facebook и Vkontakte. В данной работе рассматривается социальная сеть Vkontake, так как она наиболее популярна в российском сегменте Интернета.

Для того чтобы загрузить публикации из сообществ социальной сети Vkontakte был реализован программный модуль на языке программирования Python 2.7. Для получения доступа к информации о сообществах и их публикациям использовалась технология API Vkontakte [13], которая предоставляет методы для работы с данными социальной сети [14]. Число обращений к методам API имеет ограничение: не более 3 раз в секунду.

API (Application programming interface, интерфейс программировния приложений) представляет собой набор готовых классов, функций и структур, предоставляемых сервисом для использования во внешних программных продуктах.

Идентификатор категории	**Название категории**
0	Рекомендации
1	Новости
2	Спорт
3	Музыка
4	Развлечения
6	Бренды
7	Наука
8	Культура и искусство
9	Радио и телевидение
10	Игры и киберспорт
11	Магазины
12	Красота и стиль
13	Автомобили

Таблица 1: Категории сообществ Vkontakte

Для загрузки данных реализованный программный модуль делает запросы к методам API Vkontakte для выполнения следующих задач:

1. Получение информации о категориях сообществ с помощью метода API «groups.getCatalogInfo»

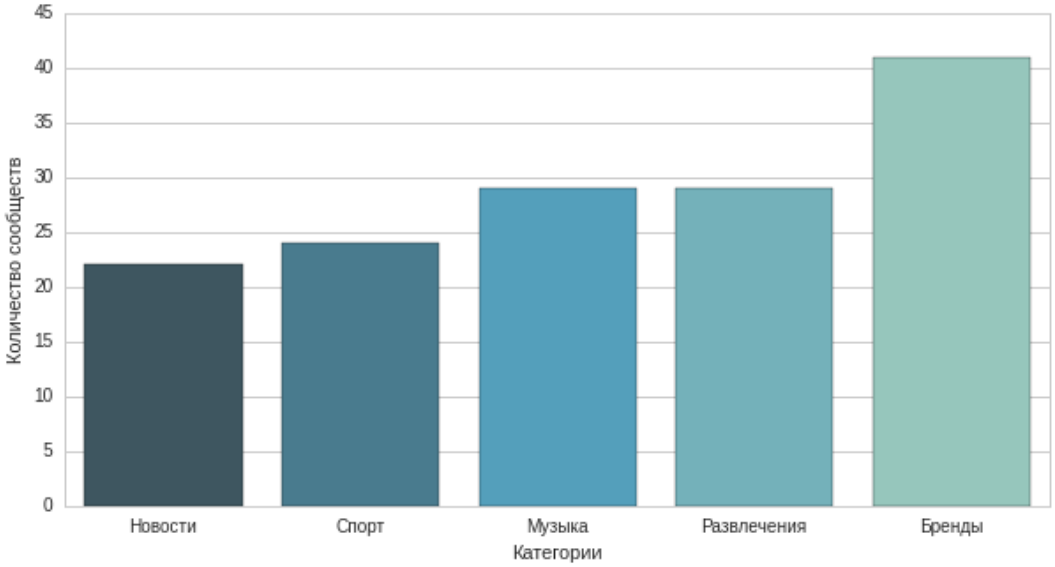

Рис. 3: Количество сообществ в категориях

2. Получение списка популярных сообществ для каждой категории с помощью метода API «groups.getCatalog»

3. Получение публикаций для каждого сообщества с помощью метода API «wall.get»

Информация о полученных категориях сообществ представлена в таблице 1. Из таблицы видно, что все сообщества социальной сети делятся на 13 категорий. Для дальнейшей работы из них были выбраны 5 категорий: «Новости», «Спорт», «Музыка», «Развлечения» и «Бренды». Для каждой из выбранных категорий был получен список популярных сообществ. Количество сообществ в каждой из категорий отображено на графике, представленном на рисунке 3. Общее число сообществ, для которых была получена информация равняется 145.

На последнем этапе работы программного модуля выполняется получение текстов всех публикаций из выбранных сообществ. На рисунке 4 изображен график, показывающий число публикаций в каждой категории. Общий размер скачанных данных составляет около 13 ГБ. График, изображенный на рисунке 5 отражает объем занимаемой памяти для каждой из категорий.

В результате работы программного модуля для каждого сообщества был создан файл, на первой строке которого записаны идентификатор и название сообщества, а на следующих строках размещены публикации этого сообщества (на одной строке одна публикация).

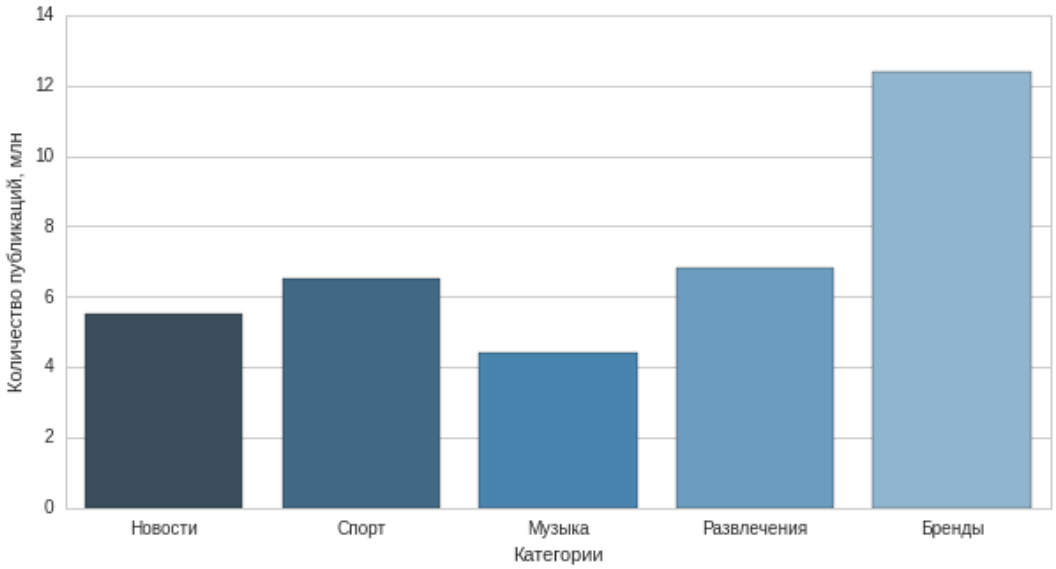

Рис. 4: Количество публикаций в категориях

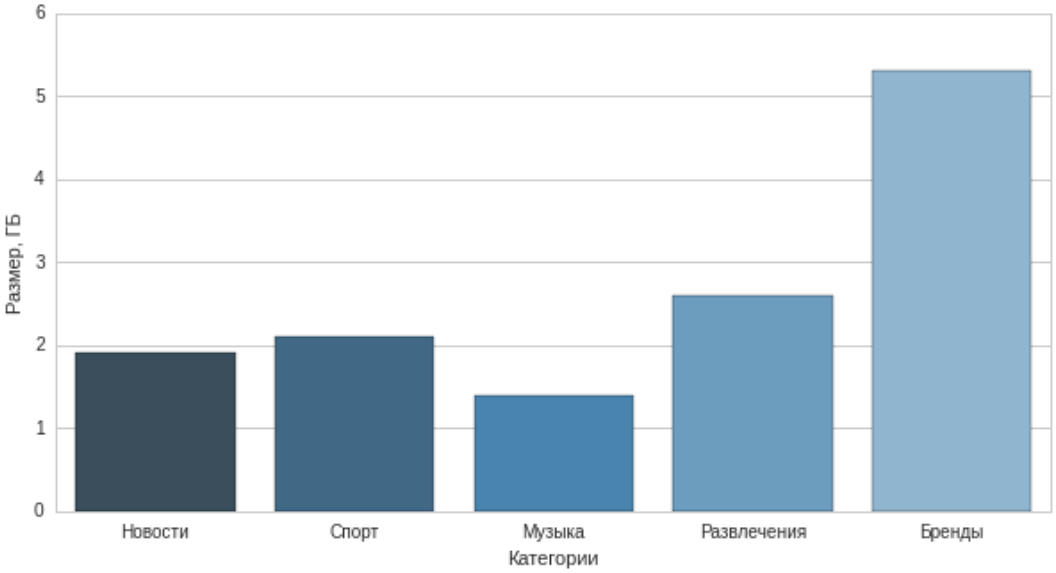

Рис. 5: Общий размер публикаций для каждой категории

1.3 Предварительная обработка данных

Перед тем как приступить к построению тематической модели необходимо провести предварительную обработку данных. Она необходима для того, чтобы избавиться от информации, которая не несет никакой смысловой нагрузки, а следовательно не оказывает заметного влияния на искомые тематики и концепции. Также предварительная обработка включает в себя уменьшение числа форм слов в тексте, так как обилие различных форм слова ведет к росту словаря и снижению качества модели.

В данной работе полагается, что к информации, которая не несет смысловой нагрузки относятся знаки препинания, эмотиконы [15], гиперс-

сылки, цифры и другие символы,не являющиеся элементами русского или английского алфавитов. К такой информации можно отнести и часто используемые слова (стоп-слова): предлоги, местоимения, союзы, числительные и частицы [3].

Стоит заметить, что многие популярные группы часто имеют одинаковые публикации, из-за чего возникает проблема дупликатов в коллекции публикаций. В данной работе эта проблема решается с помощью хеш-функций, вычисляемых для текста каждой публикации. Хэш-функция выполняет преобразование входного массива данных в выходную битовую строку фиксированной длины [10].

Сокращение числа форм слов в тексте достигается путем применения стемминга или лемматизации к словам. Алгоритм стемминга заключается в поиске неизменяемой части слова, в то время как алгоритм лемматизации более сложен и необходим для поиска нормальной формы слова. Нормальной формой слова в русском языке считается: для существительных — единственное число, именительный падеж; для глаголов и причастий — глагол в форме инфинитива; для прилагательных — мужской род, единственное число, именительный падеж. Как правило, для предварительной обработки текста выбирается один из этих алгоритмов: для русских текстов наиболее эффективна лемматизация, для английских текстов — стемминг [8]. В виду того, что в данной работе рассматриваются публикации сообществ русскоязычной социальной сети, предпочтение отдается алгоритмам лемматизации.

Для предварительной обработки данных был реализован программный модуль на языке Python 2.7.

Среди средств для лемматизации были рассмотрены два морфологических анализатора из программных пакетов pymorphy2 [16] и pymystem3 [17]. Морфологические анализатор представляют собой набор алгоритмов для сопоставления слов и их форм и выявления грамматических характристик слов. В данной работе с помощью морфологических анализаторов осуществляется приведение слов к их нормальной форме. В результате экспериментов выяснилось, что морфологический анализатор из пакета pymystem3 более эффективен, так как для определения нормальной формы слова учитываются окружающие его слова. Данный функционал отсутствует у морфологического анализатора из pymorphy2, поэтому в ре-

ализации данного программного модуля предпочтение отдано морфологическому анализатору из пакета pymystem3.

Вычисление хеш-функции осуществляется с помощью встроенной функция языка Python 2.7: hash(). Данная функция принимает на вход некоторый объект и вычисляет него хеш-значение. Стоит обратить внимание на то, что две разные строки могут иметь одинаковое значение хеш-функции. Так как вероятность такого события мала, данный эффект не окажет значительного влияния на результаты работы модуля.

Для получения списка стоп-слов русского языка в программном модуле использованы пакеты nltk и stop_words.

Работа программного модуля заключается в обработке всех файлов, полученных в разделе 1.2. В каждом файле последовательно считываются строки (текст публикации). Со всеми строками, кроме первой, выполняются следующие действия:

1. Вычисление хеш-функции;
2. Сравнение полученного значения хэш-функции строки со значениями хеш-функций ранее просмотренных строк. Если при сравнении найдутся равные значения, то данная строка удаляется из файла, иначе сохраняется значение хеш-функции и продолжается обработка строки;
3. Замена html-тегов в строке пробельными символами. Примеры html-тегов:
, <h1>.
4. Обработка строки морфологическим анализатором. Результатом обработки является строка, в которой все слова приведены к нормальной форме;
5. Разбиение строки по пробельному символу. Результатом разбиения будет список подстрок L;
6. Обработка каждой подстроки s списка L. Результатом обработки будет список обработанных подстрок L'. Алгоритм обработки подстрок рассмотрен ниже;
7. Получение результирующей строки путем конкатенации подстрок списка L' и пробельных символов.

Исходный текст	Результат обработки
Андрей Кощеев и Дмитрий Головин перед плей-офф побывали в гостях у воспитанников детского дома №3 в рамках акции #КлубДобрыхДел: http://basket.fc-zenit.ru/photo/gl6179/	андрей кощеев дмитрий головин плей-офф побывать гость воспитанник детский дом рамка акция
Правозащитник Оксана Труфанова рассказала [club27532693|"Известиям"] о том, что стало причиной бунта в челябинской колонии строго режима №6. http://izvestia.ru/news/540272#новости	правозащитник оксана труфанов рассказывать становиться причина бунт челябинский колония строго режим
У нас было много театра в последнее время. Вот теперь на музыку нажать решили. 26 февраля, например, «Сегодняночью» играют в клубе «J.Walker»	театр последний музыка нажимать решать февраль например сегодняночь играть клуб
День матча! Единая лига ВТБ #ЗенитУНИКС «Сибур Арена» 15:00 (СПб) «Матч ТВ» http://tickets.fc-zenit.ru/stadium.php	матч единый лига втб сибур арена спб матч
То прекрасное чувство, когда распаковал новый монитор :-) Крутейший 27"LED серии 3 0:) http://spr.ly/m3sv8sc	прекрасный чувство распаковывать новый монитор крутой led серия

Таблица 2: Пример работы программного модуля

Алгоритм обработки подстрок, рассмотренный в пункте 5, реализован в виде отдельной функций. Данная функция получает на вход строку и выполняет с ней следующую последовательность действий:

1. Удаление символов в начале и конце строки. Символ подлежит удалению, если он не является символом русского или английского алфавитов;

2. Если теперь строка содержит символы, которые не являются ни дефисом, ни символом русского или английского алфавитов, то функция возвращает пустую строку;

3. Распознавание аббревиатур. Строка, состоящая только из заглавных букв с длиной больше единицы и меньше семи символов распознается как аббревиатура. Если строка является аббревиатурой, то она приводится к нижнему регистру и возвращается функцией;

4. Если длина строки меньше трех, то функция возвращает пустую строку, иначе строка приводится к нижнему регистру;

5. Если строка является стоп-словом, то возвращается пустая строка, иначе возвращается сама строка.

Примеры работы программного модуля приведены в таблице 2 .

1.4 Результаты предварительной обработки

В результате предварительной обработки данных число публикаций значительно уменьшилось с 34 миллионов до 700 тысяч. Такой эффект объясняется тем, что изначально число дубликатов публкаций было довольно большим. Также предварительная обработка данных повлияла объем необходимой памяти для хранения публикаций. Общий объем занимаемой памяти уменьшился с 13 ГБ до 250 МБ.

Глава 2. Выбор и построение тематической модели

2.1 Тематическое моделирование

2.1.1 Основные сведения

Тематическое моделирование представляет собой способ построения тематической модели для коллекции текстовых документов. Тематическая модель предоставляет информацию о тематиках каждого документа и о множестве слов, образующих каждую тематику.

Тематические модели применяются в задачах тематического поиска, построения рекомендательных систем, выявления тематик и концепций в новостных потоках, а также для классификации и кластеризации документов.

В последнее время широкое распространение получили вероятностные тематические модели, которые основаны на том, что документ или термин может одновременно принадлежать разным тематикам. Вероятностная тематическая модель представляет документы в виде дискретного распределения на множестве тематик, а тематики в виде дискретного распределения на множестве терминов. Другими словами вероятностные тематические модели выполняют «мягкую» кластеризацию документов и терминов по кластерам-тематикам, что решает проблему синонимии и омонимии. Слова-синонимы употребляются в одинаковых контекстах, и поэтому с высокой вероятностью принадлежат одной тематике. Слова-омонимы употребляются в различных контекстах, из-за чего распределяются в различные тематики.

2.1.2 Постановка задачи вероятностного тематического моделирования

Пусть D — множество текстовых документов, W — множество терминов, употребляемых в них. Под термином понимается либо отдельное слово, либо словосочетание. Каждый документ $d \in D$ представлен последовательностью терминов $\{w_i\}_{i=1}^{n_d}$ из W. Один и тот же термин может встречаться в документе несколько раз.

Пусть Z - это конечное множесто тематик. Положим, что появление термина w в каждом документе d связано с некоторой, вообще говоря,

неизвестной тематикой $z \in Z$. Пользуясь этим представим множества документов в виде множества троек вида (d, w, z), выбранных случайно и независмо из дискретного распределения $p(d, w, z)$, которое задано на множестве $D \times W \times Z$. Независимость элементов выборки подразумевает, что порядок терминов в документе не важен для выявления тематик. Такое предположение носит название гипотезы "мешка слов".

Задачу вероятностного тематического моделирования можно определить следующим образом: построить вероятностную тематическую модель для коллекции документов D — значит определить множество тематик Z, распределения $p(w|z)$ для всех тематик $z \in Z$ и распределения $p(z|d)$ для всех документов $d \in D$.

2.1.3 Порождающая вероятностная модель

Помимо рассмотренных выше гипотез также используется гипотеза об условной независимости, которая указывает на то, что вероятность появления термина w при условии того, что выбрана тематика z описывается распределением $p(w|z)$ и не зависит от документа d. Это эквивалентно следующим равенствам:

$$p(w|d, z) = p(w|z);$$

$$p(d, w|z) = p(d|z)p(w|z)$$

Используя гипотезу условной независимости и определяния условной и полной вероятности, получаем:

$$p(w|d) = \sum_{z \in Z} p(w|z)p(z|d). \tag{1}$$

Равенство (1) описывает процесс порождения множества документов D, если известны распределения $p(w|z)$ и $p(z|d)$. Процесс построения тематической модели является обратной задачей и связан с поиском распределений $p(w|z)$ и $p(z|d)$ по известному множеству документов D.

2.2 Выбор тематической модели

Рассмотрим две вероятностные тематические модели pLSA и LDA, и сравним их. Для начала введем следующие обозначения:

$$\Phi = (\varphi_{wz})_{|W|\times|Z|}, \qquad \varphi_{wz} = p(w|z);$$
$$\Theta = (\vartheta_{zd})_{|Z|\times|D|}, \qquad \vartheta_{zd} = p(z|d).$$

где Φ — матрица терминов тематик, а Θ - матрица тематик документов. Стоит обратить внимание на то, что матрицы Φ и Θ являются стохастическими. Под стохастическом матрицей понимается матрица с нормированными столбцами и неотрицательными элементами.

Модели pLSA и LDA основаны на вероятностной модели появления пары «документ-слово», которая может быть представлена следующим образом:

$$p(d, w) = \sum_{z \in Z} p(w|z)p(z|d)p(d), \tag{2}$$

где $p(d)$ — это априорное распределение на множестве документов.

2.2.1 Вероятностое латентно-семантическое моделирование

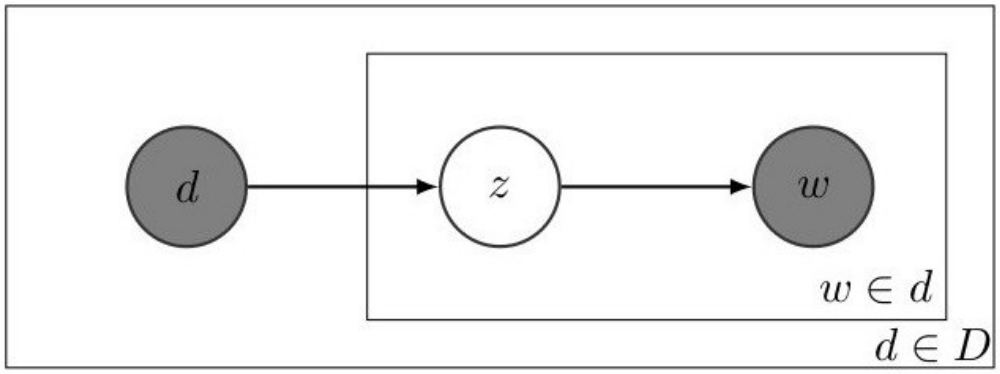

Рис. 6: Байесовская сеть модели pLSA

Модель pLSA можно представить в виде Байесовской сети, изображенной на рисунке 6. Байесовская сеть представляет собой ориентированный ацикличный граф, вершины которого соответствую случайным переменным, а ребра соответсвуют распределениям условной вероятности, где родительский узел сооветствует условной переменной, а дочерний — зависимой переменной. Темные вершины на графе соответстсуют наблюдаемым переменным (их значения известны), а белые вершины соответствуют латентным переменным, значение которых нужно найти. В модели pLSA d и w являются наблюдаемыми переменными, а z — латентной переменной.

Прямоугольник, включающий в себя некоторый подграф $\mathbf{G}$, обозначает набор из нескольких экземпляров подграфа $\mathbf{G}$. Число экземпляров определяется надписью в правом нижнем углу прямоугольника. [6].

В pLSA для оценивания параметров по коллекции документов D используется принцип максимума правдоподобия, который приводит к задаче максимизации следующего функционала (логарифма правдоподобия):

$$\begin{gathered}\mathcal{L}(\Phi,\Theta)=\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{z\in Z}\varphi_{wz}\vartheta_{zd}\to\max_{\Phi,\Theta};\\ \sum_{w\in W}\varphi_{wz}=1;\quad\sum_{z\in Z}\vartheta_{zd}=1.\end{gathered}\tag{3}$$

где n_{dw} — это число вхождений термина w в документ d.

Обычно для решения задачи (3) используется EM-алгоритм.

Основные недостатки модели pLSA:

- Число параметров линейно зависит от числа документов в коллекции, что ведет к переобучению модели;
- Невозможно вычислить $p(t|d)$ для документа d, если он добавлен в коллекцию после построения модели [8] .

2.2.2 Латентное размещение Дирихле

Как и в pLSA в основе LDA лежит вероятностная модель (2), но теперь делаются дополнительные предположения о том, что векторы документов $\vartheta_d=(\vartheta_{dz})\in\mathbb{R}^{|T|}$ и векторы тематик $\varphi_z=(\varphi_{wz})\in\mathbb{R}^{|W|}$ порождаются распределениями Дирихле с параметрами $\alpha\in\mathbb{R}^{|T|}$ и $\beta\in\mathbb{R}^{|W|}$ соответственно:

$$\mathrm{Dir}(\vartheta_d;\alpha)=\frac{\Gamma(\alpha_0)}{\prod_z\Gamma(\alpha_z)}\prod_z\vartheta_{zd}^{\alpha_z-1},\alpha_z>0,\alpha_0=\sum_z\alpha_z,\vartheta_{zd}>0,\sum_z\vartheta_{zd}=1;$$

$$\mathrm{Dir}(\varphi_z;\beta)=\frac{\Gamma(\beta_0)}{\prod_w\Gamma(\beta_w)}\prod_w\varphi_{wz}^{\beta_w-1},\beta_w>0,\beta_0=\sum_w\beta_w,\varphi_{wz}>0,\sum_w\varphi_{wz}=1;$$

где $\Gamma(z)$ — гамма-функция.

Учитывая данные предположения рассмотрим Байесовскую сеть мо-

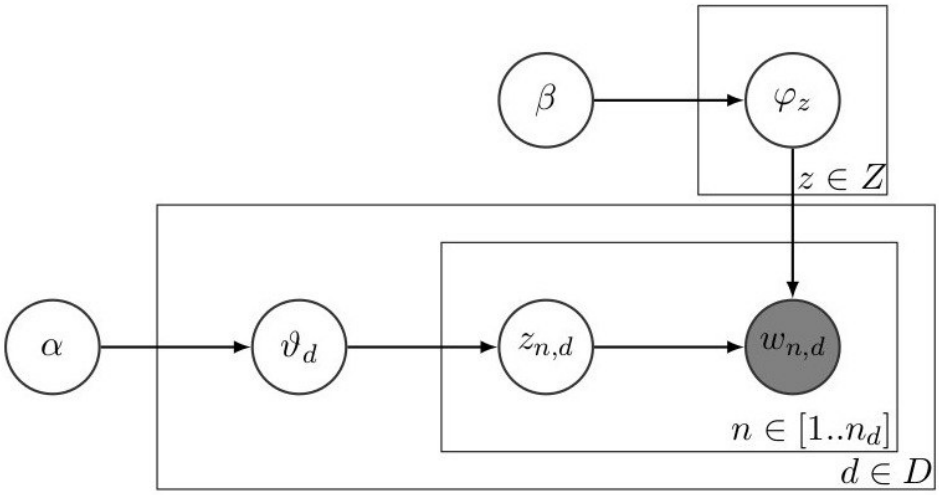

Рис. 7: Байесовская сеть модели LDA

дели LDA, изображенную на рисунке 7. Параметры α и β являются гиперпараметрами модели и параметрами распределения Дирихле, и, как правило, задаются до начало обучения модели. Переменная $w_{n,d}$ является наблюдаемой и представляет собой термин, стоящий на n-ой позиции в документе d. Все остальные переменные являются латентными (скрытыми).

Для оценки параметров модели LDA по коллекции документов D применяются вариационный Байесовский вывод, метод сэмплирования Гиббса или метод Expectation-Propagation.

Основной недостаток модели LDA заключается в том, что априорные распределения Дирихле не моделирует никаких особенностей языка и имеют слабые лингвистические обоснования. Они используются для того, чтобы облегчить Байесовский вывод для модели [8].

2.2.3 Вывод

Учитывая количество публикаций полученных из социальных сетей, достоинства и недостатки рассмотренных моделей разумным будет выбрать тематическую модель LDA. Выбор LDA облегчит работу с обучающей и тестовой выборками публикаций, так как для проверки работы модели на тестовых данных не придется выполнять построение модели заново. Также стоит обратить внимание на то, что модель pLSA больше подвержена переобучению, чем модель LDA.

В качестве метода для оценки параметров модели приоритет был отдан сэмплированию Гиббса, так как этот метод является относительно простым и эффективным алгоритмом для решения задач статистического оценивания. Псевдокод, реализующий алгоритм сэмплирования Гиббса приведен в приложении А.

Глава 3. Качество тематической модели

Одной из основных проблем тематического моделирования является оценка качества тематических моделей. Это вызвано тем, что в при оценивании модели нельзя ввести четкого понятия «ошибки». Все методы оценки качества тематических моделей подразделяются на два типа: внутренние и внешние. Внутренние методы оценки дают характеристику построенной тематической модели с точки зрения исходных данных, для которых она была построена. Внешние методы оценки оторажают уровень полезности тематической модели с точки зрения конечного пользователя [8].

3.1 Перплексия

Вычисление перплексии (perplexity) является одним из самых популярных методов внутрненней оценки качества тематической модели. Перплексия отражает меру несоответствия вероятностной модели $p(w|d)$ относительно терминов $w \in d$ и определяется следующим образом:

$$\mathcal{P}(D) = \exp\left(-\frac{1}{n}\mathcal{L}(\Phi, \Theta)\right) \tag{4}$$

где $\mathcal{L}(\Phi, \Theta)$ — логарифм правдободобия, описанный в задаче (3), значение $n = \sum_{d \in D} n_d$ выражет общее число терминов в коллекции.

Маленькие значения перплекскии говорят о хорошей предсказательной способности вероятностной модели $p(w|d)$ для терминов w в документе $d \in D$ [6].

Несмотря на свою популярность, перплексия имеет существенный недостаток, связанный с плохой интерпретируемостью числовых значений. Также важно, что перплексия зависит не только от качества построенной модели, но и от других параметров: мощность словаря, длина документов.

В тематическом моделировании различают перплексию обучающей и тестовой выборки. Значения перплексии обучающей выборки (4) являются оптимистично заниженными оценками качества модели. Для того чтобы оценить обобщающую способность модели, вычисляется перплексия на тестовой выборке, минус которой выражется в чувствительности к новым словам [8].

3.2 Экспертная оценка

Экспертная оценка относится к методам внешней оценки качества тематической модели. Для каждой тематики составляется список терминов, которые наиболее популряных с ней. Далее эти списки передаются экспертам или асессорам и для каждой тематики они определяют ее интерпретируемость и осмысленность. Задача экспертов заключается в классификации предоставленных им тематик на два класса: тематики, которым можно дать осмысленное название и которым его дать нельзя.

3.3 Когерентность

Так как экспертная оценка является затратной операцией, были разработаны методы автоматической оценки когерентности (согласованности) тематик.

Тематика называется когерентной, если наиболее встречающиеся в ней термины неслучайно часто встречаются рядом в документах коллекции.

Оценка когерентности модели выполняется с помощью поточечной взаимной информацией (pointwise mutual information, PMI) [9]:

$$PMI(z) = \sum_{i=1}^{k-1} \sum_{j=i}^{k} \log \frac{E(w_i, w_j)}{E(w_i)E(w_j)}, \tag{5}$$

где w_i — i-ый термин в порядке убывания в φ_{wz}, $E(w)$ — число документов, содержащих слово w, $E(w_i, w_j)$ указывает на количество документов, в которых хотя бы раз слова w_i и w_j встречаются рядом. Число k указывает на количество рассматриваемых терминов в φ_{wz} и обычно устанавливается равным 10 [8]

Среднее значение когерентности для всех тематик дает хорошую оценку интерпретируемости модели [9]. Чем выше среднее значение когерентности, тем лучше согласованы темы. Преимуществом когерентности над другими методами внутреннего оценивания тематических моделей является высокая корреляция с оценками экспертов.

3.4 Характеристики ядер тем

Ядром $\mathcal{J}_z$ тематики z называется множество терминов, имеющих высокую условную вероятность $p(z|w)$ для данной тематики:

$$\mathcal{J}_z = \{w \in W | p(z|w) > 0.3\}.$$

Ядро используется для получения следующих мер интерпретируемости тематики z:

- Чистота тематики: $purity(z) = \sum_{w \in \mathcal{J}_z} p(w|z)$
- Контрастность тематики: $contrast(z) = \frac{1}{|\mathcal{J}_z|} \sum_{w \in \mathcal{J}_z} p(z|w)$

Чемы больше показатели чистоты и контрастности, тем лучше интерпретируема тема.

Глава 4. Эксперименты

Для проведения экспериментов с моделью LDA реализован программный модуль на языке программирования C++. Задача данного модуля по заданной обучающей выборке и выбранным параметрам построить тематическую модель LDA. Для оценки параметров модели используется алгоритм сэмплирования Гиббса, псевдокод которого приведен в приложении А. Перед построением модели указывается следующий набор параметров:

- количество тематик;
- гиперпараметр α;
- гиперпараметр β;
- число итераций для сэмплирования Гиббса.

По окончании построения модели доступны матрица тематик Φ, матрица документов Θ, информация о тематике для каждого термина каждого документа. Также для каждой тематики доступно множество слов, наиболее характеризующих ее.

4.1 Обучающая и тестовая выборки

Рассмотрим тексты публикаций сообществ, полученных в результате предварительной обработки, рассмотренной в разделе 1.3. Если выражаться в терминах тематического моделирования, каждая такая публикация является документом. Объединим все публикации всех имеющихся сообществ в одно множество; назовем это множество коллекцей документов.

Имеющуюся коллекцию документов необходимо разбить на две части: обучающую и тестовую выборки. Обучающая выборка используется для оценки параметров Φ и Θ во время обучения тематической модели. Тестовая выборка необходима для оценки обобщающей способности построенной тематической модели. Не существует определенного правила для выбора пропорций разбиения коллекции документов. В работе [8] наиболее эффективно разбиение в отношении 9:1, где наибольшей частью является обучающая выборка.

В данной работе рассматривается разбиение коллекции документов на обучающую и тестовую выборки тремя различными способами в отношениях 4:1, 9:1 и 14:1. Для удобства последующего изложения обучающую и тестовую выборки в соотношении 4:1 обозначим D_A и D'_A, в соотношении 9:1 — D_B и D'_B, в соотношении 14:1 — D_C и D'_C. Получение обучающей и тестовой выборок является результатом случайного разбиения коллекции документов в соответствующих пропорциях.

Рассматриваемая коллекция состоит из 706007 документов. Размеры обучающих и тестовых выборок, полученных в результате разбиений, описанных выше, представлены в таблице **??**

Тип разбиения	**Размер обучающей выборки**	**Размер тестовой выборки**
4:1	564836	141241
9:1	635409	70688
14:1	658896	47181

Таблица 3: Размер обучающей и тестовой выборки для каждого типа разбиения

4.1 Пример обучения тематической модели

В данном разделе построим тематическую модель для обучающей выборки документов D_B, и рассмотрим примеры тематик, выделенных данной моделью. Для обучения модели выбраны следующие параметры: число тематик — 10, $\alpha = 0.1$, $\beta = 0.1$, число итераций — 2000. Далее для обучения всех моделей используется 2000 итераций. Время обучения данной модели составило около 14 часов.

В таблице 4 представлены 10 самых характерных терминов для каждой тематики. Из таблицы видно, что многие тематики можно попытаться интерпретировать, например, тематика 1 — «Дети», 3 — «Спорт», 7 — «Спорт», 8 — «Власть» или «Политика», 9 — «Культура». В тематиках 2, 4, 5, 6 термины также связаны между собой, но дать определенное название тематике несколько сложнее. Тематика 0 характеризуется терминами, которые не позволяют однозначно интерпретировать тематику.

Номер тематики	Термины
0	свой самый друг знать ребенок говорить становиться новый слово хороший
1	девочка ребенок месяц давать делать малыш врач подсказывать спать ночь
2	подсказывать добрый заказ здравствовать сайт размер заказывать икеа цвет товар
3	матч цска команда клуб игра россия сборная футбол чемпионат тренер
4	фильм конкурс свой новый самый получать друг хороший победитель становиться
5	новый компания свой получать рубль работа приложение сайт сеть телефон
6	город москва видео место россия область самолет дом читать житель
7	россия матч чемпионат мир спорт игра сборная команда олимпийский российский
8	россия читать российский далее страна украина президент сша заявлять путин
9	концерт группа билет новый песня фестиваль самый петербург свой театр

Таблица 4: Тематики

4.2 Перплексия

В рамках данного раздела были проведены эксперименты для выявления зависимости перплексии от количество тематик и значения гиперпараметра модели α.

Рассмотрим как влияет увеличение числа тематик модели на значение перплексии. Для этого обучим несколько моделей, гиперпараметры которых фиксированы $\alpha = 0.1$, $\beta = 0.1$. На рисунке 8 изображены графики зависимости значений перплексии обучающей выборки от количества тематик.

На графике видно, что увеличение количества тематик уменьшает значение перплексии, независимо от того, в каком соотношении выполнялось разбиение на обучающую и тестовые выборки.

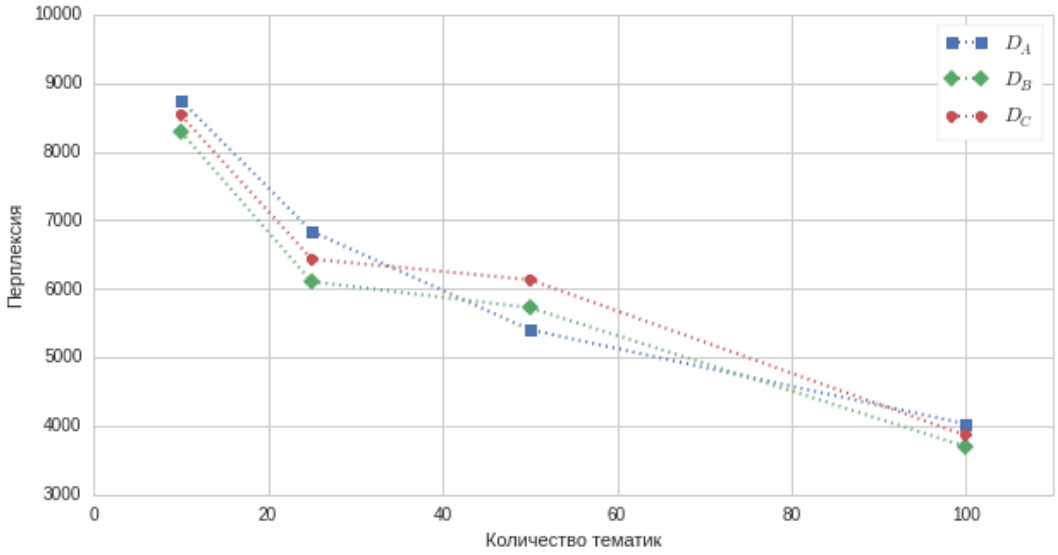

Рис. 8: Зависимость перплексии обучающей выборки от количества тематик

На рисунке 9 изображены аналогичные графики, но только для перплексии тестовой выборки. Опять же значения перплексии уменьшаются при увеличении количества тематик. Стоит обратить внимание на то, что значения перплексии тестовой выборки больше значений перплексии обучающей выборки для одного и того же количества тем. Данных эффект связан с переобучением тематической модели.

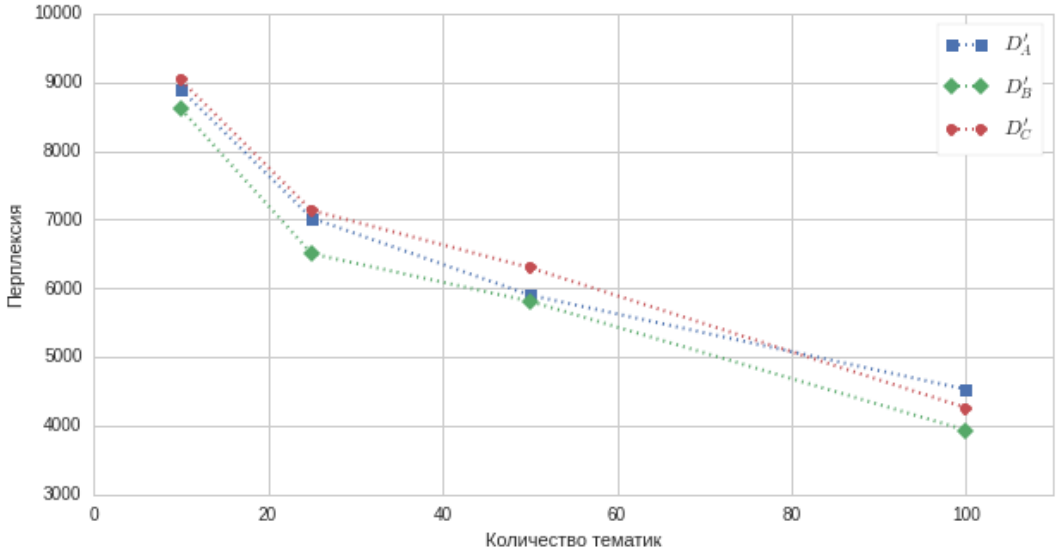

Рис. 9: Зависимость перплексии тестовой выборки от количества тематик

Теперь рассмотрим влияние гиперпараметра α на значение перплексии. Обучим тематические модели с различными параметрами α и с фиксированными числом тематик равным 100 и гиперпараметром $\beta = 0.1$.

На рисунке 10 представлены графики зависимости перплексиии обучающей выборки от значения гиперпараметра α, из которых видно, что

уменьшение гиперпараметра α влечет уменьшение значения перплексиии обучающей выборки.

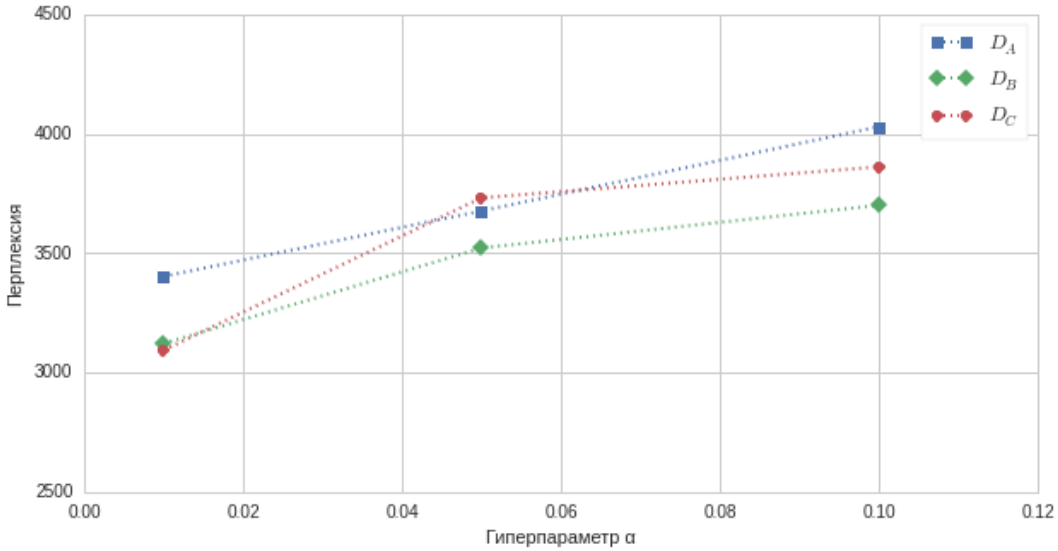

Рис. 10: Зависимость перплексии обучающей выборки от гиперпараметра α

Зависимость перплексии тестовой выборки от гиперпараметра α, изображенная на рисунке 11 прослеживается аналогично предыдущему примеру. Опять же заметим, что перплексия тестовой выборки превышает по показателям перплексию обучающей выборки.

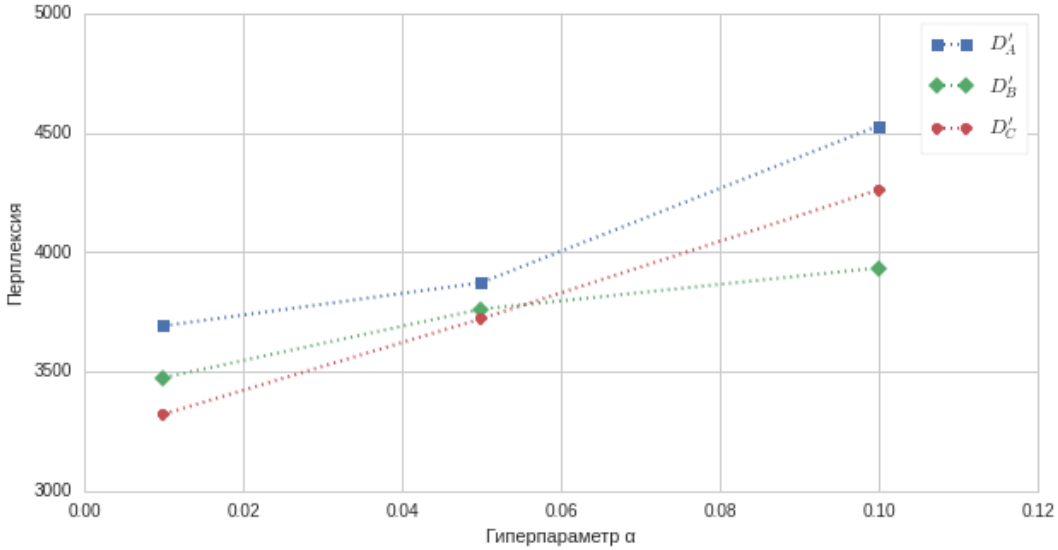

Рис. 11: Зависимость перплексии тестовой выборки от гиперпараметра α

Проанализировав предложенные графики зависимости перплексии от параметров, можно заметить, что наихудшие значения перплексии достигаются на тематической модели, для которой разбиение на обучающую и тестовую выборки выполнялось в отношении 4:1.

4.3 Когерентность

В данном разделе рассмотрим зависимость среднего значения когерентности для всех тематик модели от количества тематик. Из раздела 3.3 известно, что чем выше среднее значение когерентности, тем лучше согласованы тематики.

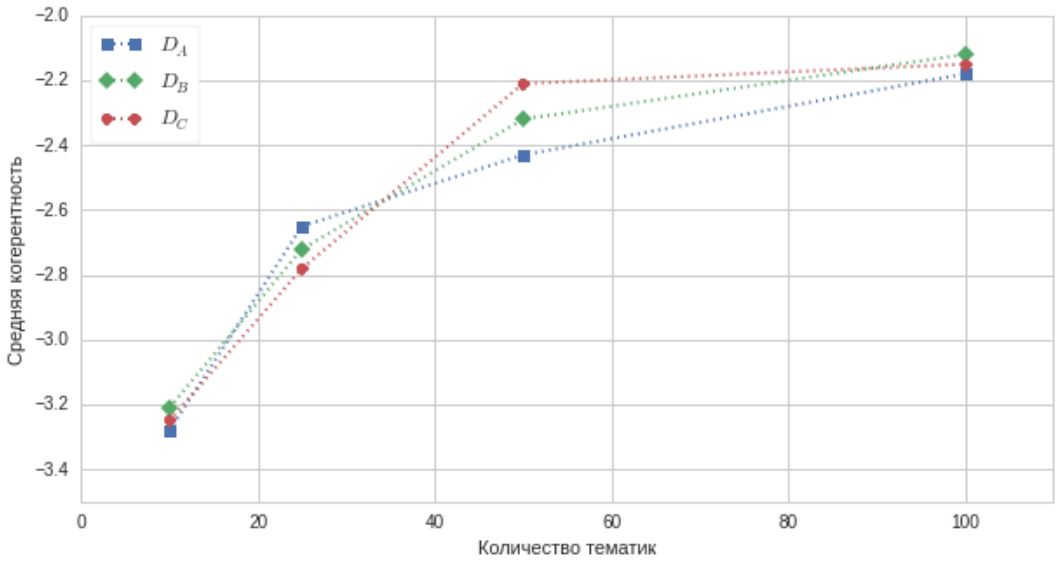

Рис. 12: Зависимость средней когерентности от количества тематик

На рисунке представлены графики зависимости средней когерентности от количества тематик. Результаты представленные на рисунке говорят о том, что увеличение числа тематик влечет лучшую их согласованность.

4.4 Результаты

В разделе 4.1 был рассмотрен пример тематической модели, которая уже неплохо определяла тематики для обучающей коллекции документов. Далее был проведен ряд эксперментов для выяснения влияния параметров тематической модели на ее перплексию и среднюю когерентность. В результате проведенных экспериментов выяснилось, что модели построенные по выборкам D_B и D_C имеют лучшее качество по сравнению с моделями, обученными по выборке D_A.

Из все построенных в ходе экспериментов тематических моделей наилучшее качество имела модель построенная на обучающей выборке D_C со следующими параметрами: число тематик — 100, $\alpha = 0.01$, $\beta = 0.1$.

Анализ результатов

В результате проделанной работы удалось выполнить все поставленные задачи. В качестве источника данных для контекстной обработки рассмотривалось множество социальных сетей, среди которых была выбрана социальная сеть Vkontakte. В качестве данных для последующей обработки были выбраны публикации в сообществах социальной сети. Далее были описаны методы загрузки и предварительной обработки данных. Для того чтобы реализовать контестную обработку данных были рассмотрены основные идеи тематического моделирования, а также две тематические модели pLSA и LDA. Для дальнейшей обработки данных была выбрана модель LDA с оцениванием параметров методом сэмплирования Гиббса.

В ходе проведенных экспериментов сравнивались различные стратегии разбиения исходной коллекции на обучающую и тестовую выборк, в ходе которых выяснилось, что разбиения в отношении 9:1 и 14:1 эффективнее разбиения в соотношении 4:1

Также в ходе проведенных экспериментов была построена тематическа модель LDA, которая неплохо определила основные тематики. Заметим, что изначально публикации были выбраны из сообществ следующих категорий «Новости», «Спорт», «Музыка», «Развлечения», «Бренды». Из тематик, которые были определены в ходе обучения модели, явно выделяются тематики о спорте, культуре и политике. Заметная связь построенных тематик и категорий сообществ.

Следующим шагом выполнялось сравнения построенной модели с другими моделями на основе оценок качества, таких как перплексия и когерентность. В ходе этих экспериментов выяснилась зависимость качества модели от количества тематик и от значений гиперпараметра α. В результате проведенных экспериментов среди всех построенных тематических моделей была выбрана модель с наилучшей оценкой качества. В таблице 5 приведены наборы темринов для некторых тематик, выделенных этой моделью. Среди тематик представленных в таблице выделяются тематики о власти, детях, спорте, культуре и другие.

Основывая на результатах работы построенной тематической модели можно сделать вывод, что метода тематического моделирования хорошо проявляют себя при контекстной обработке данных.

Номер тематики	**Термины**
5	россия украина страна президент сша путин москва глава владимир власть
21	девочка ребенок врач малыш помогать мама сын говорить добрый ребеночек
25	заказ заказывать размер приходить почта получать отправлять посылка возврат доставка
46	матч цска команда клуб игра россия сборная футбол чемпионат тренер
53	фильм фильм город история парк театр петербург kudago место музей выставка
61	новый хороший получать конкурс победитель ждать становиться участник приз участие поздравлять
63	город друг ребенок любить рука женщина слово помогать давать понимать жить
78	россия матч цска команда клуб игра россия сборная чемпионат футбол тренер

Таблица 5: Тематики

Заключение

В рамках данной работы было рассмотрено применение вероятностных тематических моделей для контекстной обработки данных, полученных из сообществ социальной сети Vkontakte в виде текстов публикаций. Для достижения этой цели были реализованы алгоритмы для загрузки данных из социальной сети, предварительной обработки данных и построения тематической модели LDA. В ходе экспериметов была проведена оценка качества нескольких тематических моделей, среди которых была выбрана модель с лучшими оценками качества. Также эксперименты показали зависимость качества модели от количества тематик и ее гиперпараметров.

В результате экспериментов выяснилось, что модель LDA хорошо справляется с задачей выделения ключевых тематик и концепций в коллекции документов.

Для дальнейших исследований имеет смысл рассмотреть робастные тематические модели и методы аддитивной регуляризации тематических моделей. Робастные модели основываются на том, что на появление отдельных терминов в документе влияет не только тематика документа, но также фон и шум. К фону, как правило, относятся стоп-слова, которые не удалось отбросить на стадии предварительной предобработки. К шуму относятся термины характерные для конкретного документа. Методы аддитивной регуляризации позволяют создавать большое число различных тематических моделей, не прибегая к различным вероятностным допущениям, как это было сделано в модели LDA.

Список литературы

[1] Arturas Kaklauskas. Biometric and Intelligent Decision Making Support // Springer, 2015, P. 220.

[2] Paul, MJ. and M. Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. // In Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM), 2011.

[3] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008. 506 P.

[4] Thomas Hofmann. Probabilistic latent semantic indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999.

[5] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation // Journal of Machine Learning Research 3, 2003. P. 993 – 1022.

[6] Gregor Heinrich. Parameter estimation for text analysis. Technical report, Fraunhofer IGD, Darmstadt, Germany, 2005.

[7] David Blei. Introduction to Probabilistic Topic Models // Communications of the ACM, 2012. P. 77–84.

[8] Воронцов К.В. Вероятностное тематическое моделирование // www.machinelearning.ru : web. — 2013.

[9] Newman D., Lau J. H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — HLT '10. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. — Pp. 100–108.

[10] Хеш-функция. https://ru.wikipedia.org/wiki/Хеширование

[11] http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[12] http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[13] API Vkontakte: https://vk.com/dev/apiusage

[14] Описание методов API Vkontakte: https://vk.com/dev/methods

[15] https://ru.wikipedia.org/wiki/Эмотикон

[16] Документация для морфологического анализатора pymorphy2: https://pymorphy2.readthedocs.io/en/latest/

[17] Документация для pymystem3: https://pypi.python.org/pypi/pymystem3/0.1.1

Приложение А

Псевдокод алгоритма сэмплирования Гиббса

Вход: D; число тем |Z|; параметры α и β

Выход: Распределения Φ и Θ

$n_{wz}, n_{zd}, n_z, n_d := 0$ для всех $d \in D, w \in W, z \in Z$;

для всех $i := 1, .., i_{\max}$

- **для всех** $d \in D$
  - **для всех** $w_1, ..., w_{n_d} \in d$
    - **если** $i > 1$ **тогда**
      - z $:=$ z_{dw}; $n_{wz} := n_{wz} - 1$; $n_{zd} := n_{zd} - 1$;
      - $n_z := n_z - 1$; $n_d := n_d - 1$;
    - **конец**
    - $p(z|d, w) = \text{norm}_{z \in Z} \left(\frac{n_{wz} + \beta_w}{n_z + \beta_0} \cdot \frac{n_{zd} + \alpha_z}{n_d + \alpha_0} \right)$ для всех $z \in Z$
    - сэмплировать одну тему z из распределения $p(z|d, w)$
    - $z_{dw} := z$; $n_{wz} := n_{wz} + 1$,; $n_{zd} := n_{zd} + 1$; $n_z := n_z + 1$;
    - $n_d := n_d + 1$;

$\varphi_{wz} := \frac{n_{wz}}{n_z}$ для всех $w \in W, z \in Z$

$\vartheta_{zd} := \frac{n_{zd}}{n_d}$ для всех $d \in D, z \in Z$