Санкт-Петербургский государственный университет
Кафедра технологий программирования

Башарин Егор Валерьевич

Выпускная квалификационная работа бакалавра

Контекстная обработка данных социальных сетей

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 [**?**].

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 при принципы и идеи тематического моделирования. Стоит обратить внимание, что под социальной сетью в данной работе 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
   - 1.1 Обзор социальных сетей
   - 1.2 Выбор социальной сети
2. Подготовка данных
   - 2.1 Загрузка данных с веб-страниц социальной сети
   - 2.2 Предобработка данных
   - 2.3 Разбиение данных на обучающую и тестовую части
3. Выбор тематической модели
   - 3.1 Анализ тематических моделей
   - 3.2 Выбор подходящей тематической модели
4. Построение тематической модели
   - 4.1 Анализ методов построения тематической модели
   - 4.2 Реализация программного модуля, реализующего выбранную тематическую модель
   - 4.3 Обучение тематической модели
5. Оценка качества модели
   - 5.1 Обзор и анализ оценок качества тематических моделей
   - 5.2 Оценка качества построенной модели
6. Анализ полученных результатов

Обзор литературы

Тема данной работы тесно пересекается с информационным поиском, основы которого подробно рассмотрены в книге Кристофера Майнинга "Introduction to Information Retrieval" [?]. Особое внимание стоит уделить главам 2 и 18. В главе 2 описываются методы подготовки и предобработки 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 [?]. Затем они были развиты Дэвидом Блеем в его статье 2003 года [?], 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 [?] 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" [?] 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 [?] 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. Работа завершается рассмотрением способов оценки качества тематических моделей.

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 [?]. В соответсвии с этим можно сделать вывод о том, что социальные сети прочно укрепляются в жизни совеременного человека, а их изучение становится актуальной проблемой.

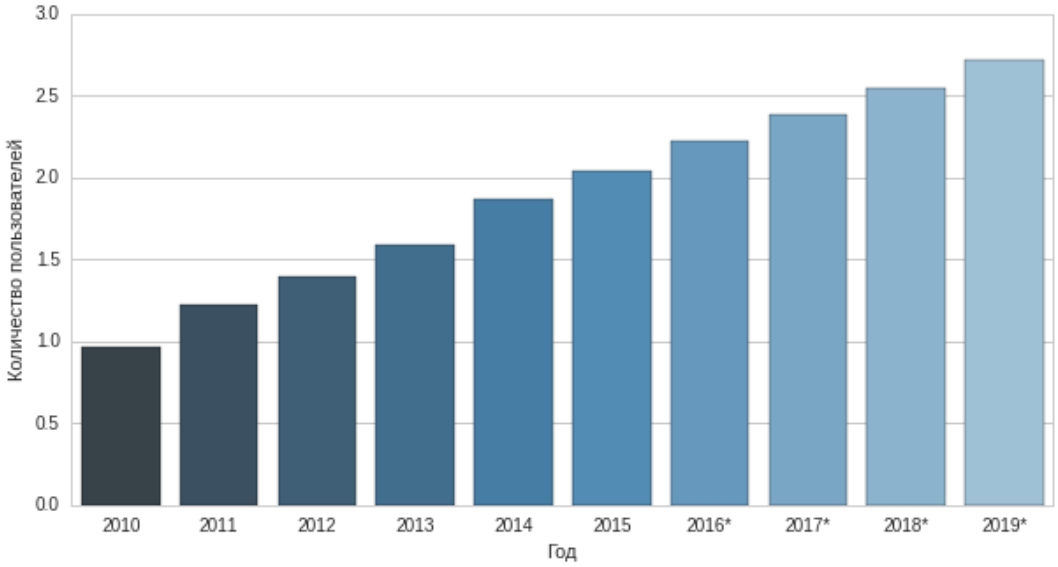

Рис. 1: Число пользователей социальных сетей по годам

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 [?]. 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.

Социальные сети Facebook и Vkontakte предоставляют похожие возможности своим пользователям: создание профиля с фотографией и ин-

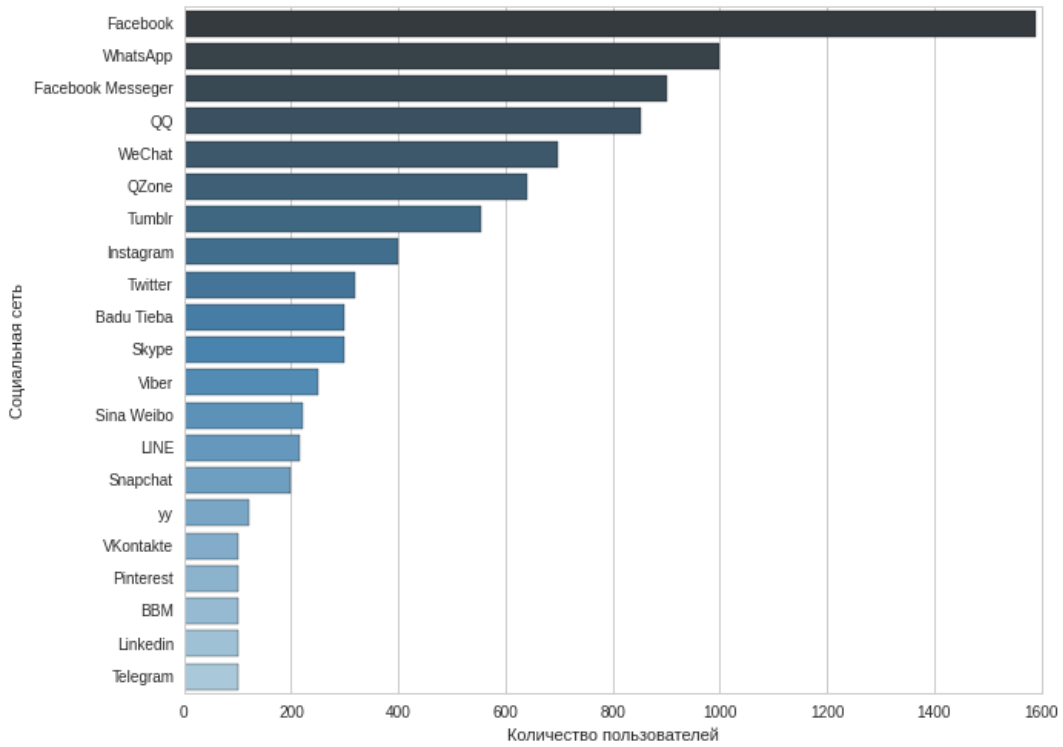

Рис. 2: Рейтинг самых популярных социальных сетей на апрель 2016 года

формацией о себе, обмен сообщениями с другими пользователями, публикация сообщений на страницах других пользователей или сообществ, создание 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 в основном 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.

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 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 [**?**], которая предоставляет методы для работы с данными социальной сети [**?**]. 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.

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»
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

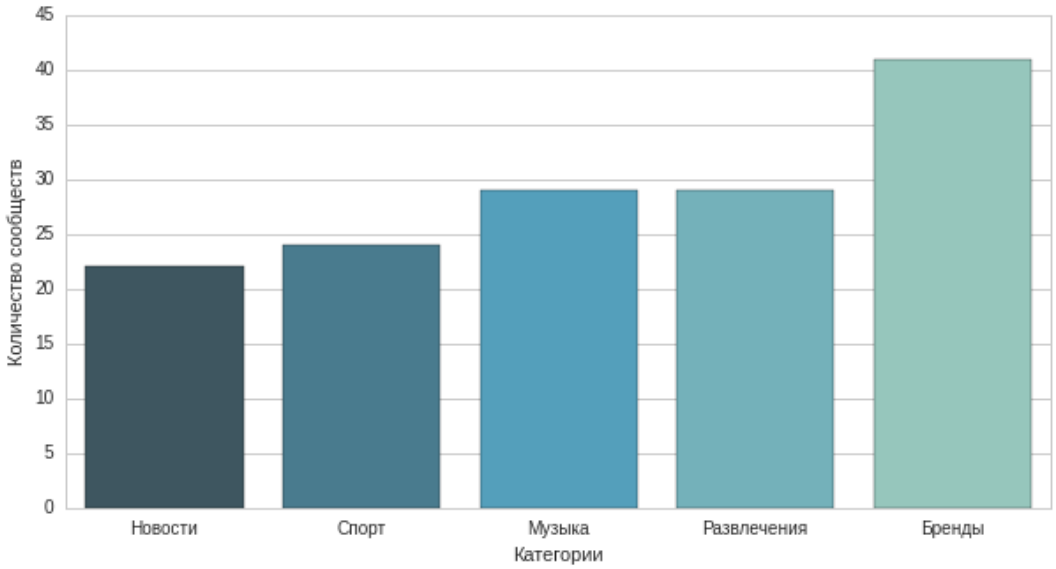

Рис. 3: Количество сообществ в категориях

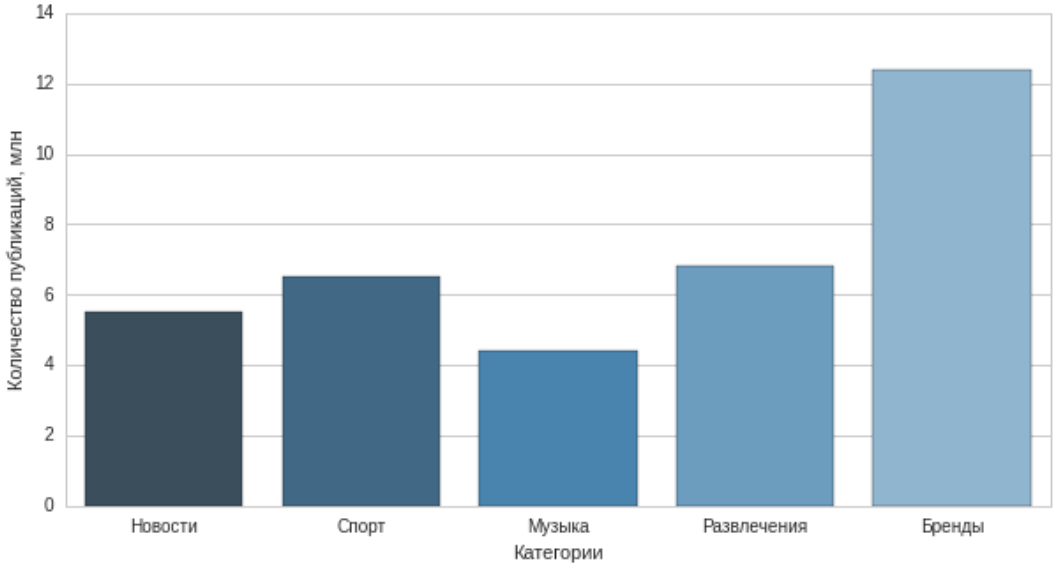

Рис. 4: Количество публикаций в категориях

из выбранных категорий был получен список популярных сообществ. Количество сообществ в каждой из категорий отображено на графике, представленном на рисунке 3. В результате была получена информация о 145 различных сообществах.

Последним этапом работы программного модуля является получение текстов всех публикаций из выбранных сообществ. На рисунке 4 изображен график, показывающий число скачанных публикаций в каждой из категорий. Общий размер скачанных данных составляет около 13 ГБ. График изображенный на рисунке **??** отражает какое количество памяти занимает каждая из категорий.

В результате для каждого сообщества был создан файл, на первой строке которого записаны идентификатор и название сообщества, а на сле-

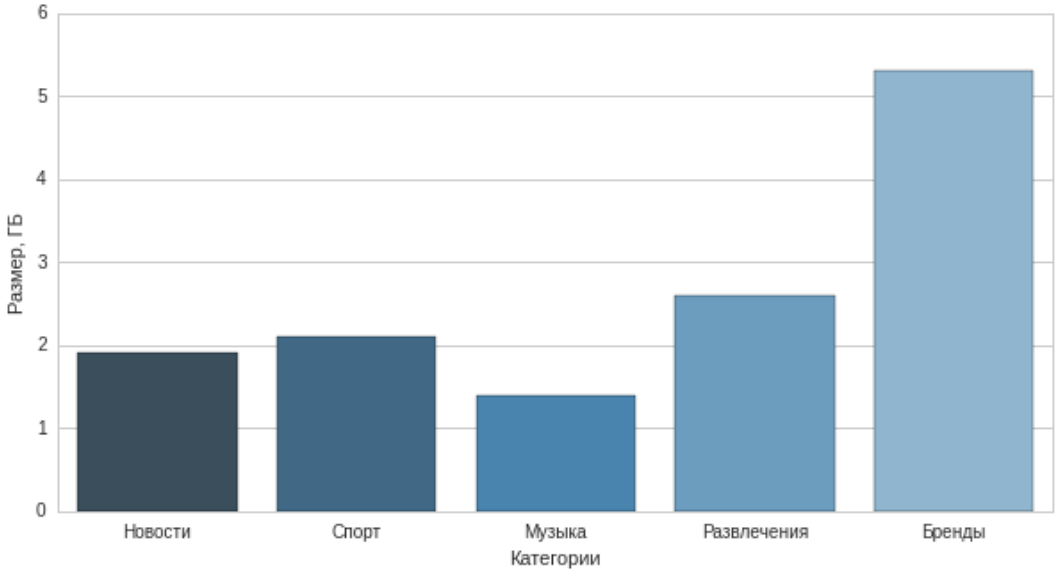

Рис. 5: Общий размер публикаций для каждой категории

дующих строках публикации этого сообщества (на одной строке одна публикация).

1.3 Предобработка данных

Перед тем как приступить к построению тематической модели необходимо провести предобработку данных. Ее необходимо сделать для того, чтобы избавиться от информации, которая не несет никакой смысловой нагрузки, а следовательно не оказывает заметного влияния на искомые тематики и концепции. Также предобработка включает в себя уменьшение числа форм слов в тексте, в результате чего каждому слову ставится в соответствие некий термин, который является результатом стемминга или лемматизации исходного слова.

В данной работе полагается, что к информации, которая не несет смысловой нагрузки относятся знаки препинания, эмотиконы [**?**], гиперссылки и другие символы, не являющиеся цифрами или элементами русского или английского алфавитов. Также к такой информации можно отнести часто используемые слова (стоп-слова): предлоги, местоимения, союзы и частицы [**?**].

Стоит отметить, что многие популярные группы связаны друг с другом и имеют одинаковые публикации. А значит возникает проблема дубликатов. В данной работе эта проблема решается с помощью хеш-функций, вычисляемых для текста каждой публикации. **Нужно ли писать о хеш функциях?**

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 лемма-

тизации более сложен и необходим для поиска нормальной формы слова **нужно ли давать определение нормальной формы?**. Как правило, для предобработки текста выбирается один из этих алгоритмов: для русских текстов наиболее эффективна лемматизация, для английских текстов — стемминг [**?**]. В виду того, что в данной работе рассматриваются публикации сообществ русскоязычной социальная сети, предпочтение отдается алгоритмам лемматизации.

Для предобработки данных был реализован программный модуль на языке Python 2.7. Среди средств для лемматизации были рассмотрены 2 морфологических анализатора из программных пакетов pymorphy2 [**?**] и pymystem3 [**?**]. В результате экспериментов выяснилось, что морфологический анализатор из пакета pymystem3 более эффективен, так как он учитывает контекст, к тому же у морфологического анализатора из пакета pymorphy2 возникали проблемы с обработкой имен и фамилий. Поэтому в реализации данного программного модуля предпочтение было отдано морфологическому анализатору из пакета pymystem3.

Работа программного модуля заключается в просмотре всех файлов полученных в пункте 1.2. В каждом файле последовательно считываются строки (текст публикации). Для каждой строки вычисляется хеш-функция, значение которой сравнивается со значениями хеш-функций уже просмотренных строк. Если такое значение уже было, то строка удаляется, иначе сохраняется значение хеш-функции. (**стоит ли писать про асимптотическую сложность, и про хранение значений хешей?**) Далее строка обрабатывается морфологическим анализатором. Результатом обработки является строка, в которой все слова приведены к нормальной форме. Полученная строка разбивается на подстроки по пробельному символу, после чего каждая подстрока проходит проверку и обработку, в результате чего она либо удаляется, либо заменяется на результат обработки. Из оставшихся подстрок формируется новая строка, которая и будет результатом предобработки.

Алгоритм обработки и проверки подстроки принимает на вход подстроку. **тут тоже алгоритм, нужно ли описывать его?**

стоит ли привести пример работы программного модуля?

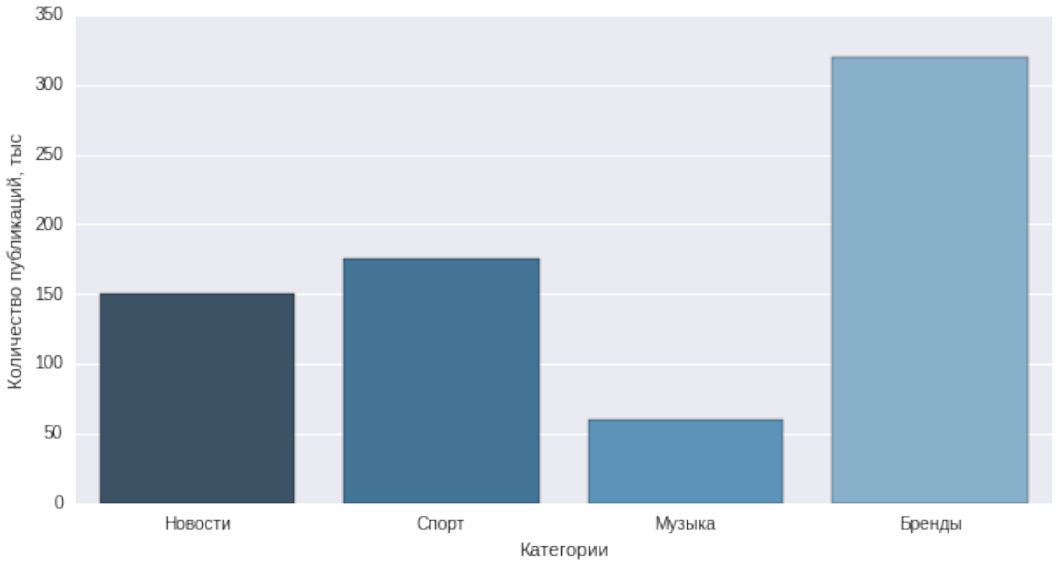

Рис. 6: Количество публикаций в категориях

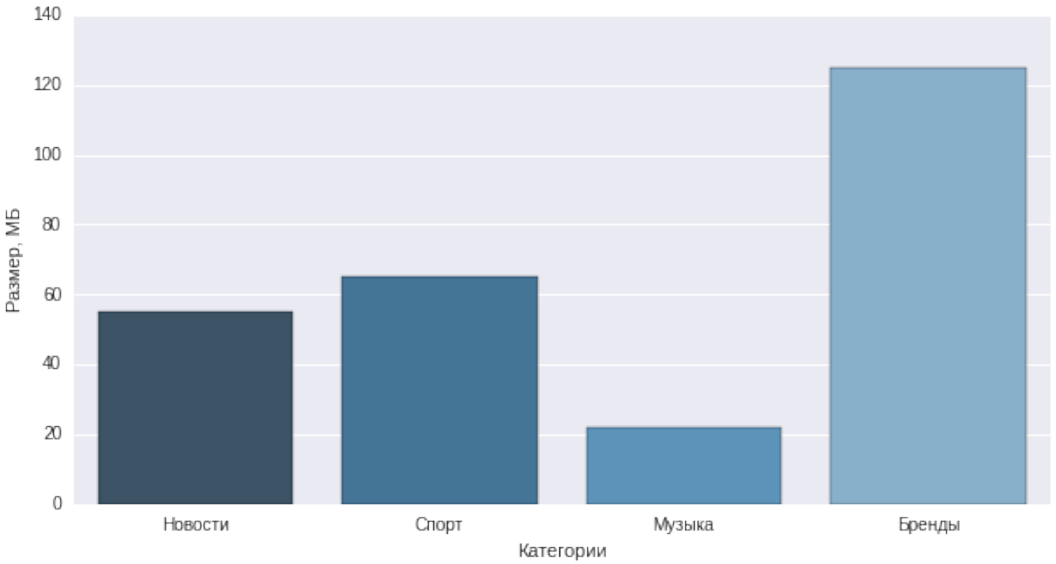

Рис. 7: Общий размер публикаций для каждой категории

1.4 Результаты

В результате предобработки данных число публикаций значительно уменьшилось. Также пропала одна из категорий сообществ «Развлечения». Это связано с тем, что публикации в данной категории сообществ точно такие же, как и в сообществах других категорий. На рисунке **??** изображен график количества публикаций после предобработки. Общее число публикаций уменьшилось с 34 миллионов до 700 тысяч. Также предобработка данных повлияла объем необходимой памяти для хранения публикаций. На рисунке **??** изображен график, на котором указан объем занимаемой памяти каждой из категорий. Общий объем памяти уменьшился с 13 ГБ до 250 МБ.

Глава 2. Выбор и построение тематической модели

2.1 Тематическое моделирование

2.2 Выбор тематической модели

2.3 Описание тематической модели LDA

2.4 Реализация тематической модели

2.5 Эксперименты

Глава 3. Оценка качества модели

3.1 Обзор оценок качества тематических моделей

3.2 Оценка качества построенной тематической модели

Анализ результатов

Заключение

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

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

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

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

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

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

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

[16] ссылка на machine learning

[17] можно всякие документации добавить