Федеральное государственное автономное образовательное
учреждение высшего образования
Санкт-Петербургский национальный исследовательский университет
информационных технологий механики и оптики

на правах рукописи

Карпович Сергей Николаевич

Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов

Специальность 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель
д.т.н, профессор
Смирнов Александр Викторович

Санкт-Петербург – 2017

Оглавление

Введение

В связи с развитием цифровых технологий, постоянным ростом интернета, увеличением количества новостей, электронных писем, постов в блогах, растет потребность в алгоритмах для автоматической обработки текстов. Алгоритмы вероятностного тематического моделирования являются одним из перспективных направлений дистрибутивного анализа коллекций и потоков текстовых документов на естественном языке.

Вероятностное тематическое моделирование – это способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов. Вероятностные тематические модели (далее ВТМ) задают мягкую кластеризацию слов и документов по кластерам-темам, означающую, что слово или документ могут быть отнесены сразу к нескольким темам с различными вероятностями. ВТМ описывает каждую тему дискретным распределением на множестве терминов, каждый документ – дискретным распределением на множестве тем. В результате синонимы с большой вероятностью будут отнесены к одной теме, а омонимы попадут в разные. ВТМ, как правило, основаны на гипотезе «мешка слов» и «мешка документов», т.е. порядок слов в документе и порядок документов в коллекции не имеют значения [51, 71, 100].

ВТМ применяются для анализа потоков текстов. В данной работе под текстовым потоком понимается последовательность текстовых документов с определенным для каждого описанного события временем происшествия. Под обработкой потока текстовых документов понимается комплексная задача кластеризации поступающих документов и анализа эволюции тем этих документов. Для изучения особенностей алгоритмов вероятностного тематического моделирования при работе с русским языком необходимы специальные русскоязычные текстовые корпуса.

Таким образом, подготовка данных для проведения исследований ВТМ, исследование свойств ВТМ и разработка методов вероятностного тематического

моделирования для решения задач интеллектуального анализа текстов на естественном языке являются актуальными и востребованными задачами.

Степень разработанности темы. Появлению методологических основ ВТМ способствовала работа Рагавана, Пападимитриу, Томаки и Вемполы опубликованная в 1998 году [89]. Развитие вероятностного тематического моделирования отражено в работах зарубежных ученых Томаса Хофмана [71], Дэвида Блея [48, 49, 51, 52, 70, 96, 105], Эндрю Ына, Майкла Джордана и др. Вклад в развитие ВТМ внесли российские ученые Воронцов К.В. [6], Потапенко А.А. [4], Лукашевич Н.В. [32], Нокель М.А. [33], Коршунов А.В. [21], Гомзин А.Г. Разработаны программные библиотеки для тематического моделирования, такие как Mallet [112], Gensim [113] и BigArtm [114], позволяющие создавать ВТМ.

ВТМ успешно применяются в задачах информационного поиска [106], рекомендательных сервисах [109], методах разрешения морфологической неоднозначности [55]. Существуют несколько классов вероятностных тематических моделей, направленных на решение конкретных практических задач. Автор-тематические [91] модели позволяют определять автора документа и находить документы одного автора. Темпоральные тематические модели [111] позволяют прослеживать изменение популярности тем во времени, делать визуализацию эволюции тем во времени. Динамические тематические модели [49] позволяют обрабатывать потоки текстовых документов.

Однако остаются слабоизученные свойства и возможности ВТМ. Одна из них – это применение тематического моделирования для многозначной классификации документов. В большинстве прикладных задач необходимо обрабатывать поток текстовых документов, поэтому улучшение темпоральных и динамических ВТМ является важным направлением изучения вероятностного тематического моделирования. Для исследования особенностей алгоритмов вероятностного тематического моделирования при работе с русским языком необходимы

русскоязычные текстовые корпуса, распространяемые по свободной лицензии, включающие востребованную при построении ВТМ метаинформацию.

Целью диссертационной работы является разработка математического и программного обеспечения вероятностного тематического моделирования потока текстовых документов, позволяющего повысить доступность применения ВТМ за счет использования открытого программного обеспечения при решении прикладных задач информационного поиска, создании сервисов рекомендаций, анализе коллекции и потока текстовых документов.

Для достижения цели в работе поставлены следующие задачи:

1. Провести анализ современных методов вероятностного тематического моделирования для оценки ситуации в проблемной области и выявления путей повышения эффективности обработки текстовых данных.
2. Подготовить русскоязычный корпус текстов для тестирования алгоритмов вероятностного тематического моделирования, включающий помимо основного текста документа метатекстовую разметку о темах, к которым относится документ, его авторе и дате описанных событий, позволяющий эмулировать поток текстовых документов, исследовать динамические и темпоральные ВТМ.
3. Для анализа потока текстовых документов и отслеживания эволюции тем разработать алгоритм многозначной классификации текстовых документов с помощью вероятностного тематического моделирования.
4. Для пополнения словаря динамической ВТМ предложить метод определения тематик для «новых слов», отсутствующих в ВТМ на момент ее построения.
5. Апробировать предложенные метод и алгоритм путем создания прототипа программного комплекса для вероятностного тематического моделирования.

Методы исследования. При решении поставленных задач использовались методы системного анализа, математического и компьютерного моделирования,

автоматической обработки естественного языка, теории вероятностей, математической статистики, прогнозирования временных рядов, теории машинного обучения и теории алгоритмов, разработки информационных систем и программирования.

Положения, выносимые на защиту:

1. Разработанный специальный русскоязычный корпус текстовых документов SCTM-ru позволяет исследовать алгоритмы вероятностного тематического моделирования.
2. Разработанный новый метод расчета матриц BTM на основе обучения с учителем (авторами документов) с учетом заданных связей между документами и темами упрощает построение BTM.
3. Разработанный оригинальный алгоритм классификации текстовых документов на базе BTM позволяет выполнять их многозначную классификацию.
4. Разработанный метод определения кластеров-тем для слова с использованием произведения Адамара позволяет определить темы «нового слова» в потоке текстовых документов.
5. Комплекс программных средств, разработанный на основе микросервисной архитектуры для вероятностного тематического моделирования, обеспечивает создание персонифицированных приложений для интеллектуального анализа коллекций и потоков текстовых документов**.**

Научная новизна работы состоит в следующем:

1. Создан русскоязычный корпус текстов SCTM-ru, позволяющий исследовать алгоритмы вероятностного тематического моделирования и отличающийся от других корпусов наличием оригинального текста документа и метатекстовой разметки: автор, время описанных событий, тема. Текст и метатекстовая разметка необходимы для построения различных видов BTM. Источником данных корпуса является сайт «Русские Викиновости».

2. Предложен метод расчета матриц ВТМ на основе обучения с учителем (авторами документов), учитывающий заданные связи между документами и темами, что позволяет упростить построение ВТМ за счет отсутствия итераций.
3. Предложен алгоритм многозначной классификации текстовых документов ml-PLSI, основанный на вероятностном тематическом моделировании, заключающийся в использовании матрицы «слово-тема» ВТМ для классификации документов, что позволяет определять темы «новых документов» при анализе потока текстовых документов в динамической тематической модели.
4. Предложен метод определения тем «нового слова», основанный на использовании произведения Адамара тематических векторов документов, содержащих это слово, позволяющий определять вектора тем для «новых слов» в потоке текстовых документов при построении динамической тематической модели с эффективностью, превосходящей существующие аналоги.
5. Разработан прототип комплекса программных средств для анализа потока текстовых документов с использованием вероятностного тематического моделирования, отличающийся использованием микросервисной архитектуры и позволяющий предоставить вариативность выбора подходящих способов решения конкретных практических задач, а также возможность визуализации промежуточных и конечных результатов вероятностного тематического моделирования.

Обоснованность и достоверность научных положений, основных выводов и результатов диссертационной работы обеспечиваются анализом состояния исследований в проблемной области, корректным использованием методов исследования, подтверждена результатами вычислительных экспериментов и эффективностью алгоритмов (сложность, трудоемкость) и программного обеспечения

(надежность) при внедрении, а также апробацией основных теоретических положений диссертации в печатных трудах и на конференциях.

Практическая ценность работы. Результаты диссертационной работы могут найти применение в задачах анализа текстов на естественном языке, информационном поиске и в сервисах рекомендаций. Разработанная система позволяет анализировать коллекции и потоки текстовых документов, строить BTM, анализировать изменение популярности тем во времени с помощью темпоральных BTM.

Реализация результатов работы. Исследования, отраженные в диссертации, проведены в рамках НИР № 714630 «Разработка теоретических и технологических основ социо-киберфизических систем», проводимой в Университете ИТМО (государственная программа поддержки ведущих университетов РФ, субсидия 074-U01). Результаты, полученные в ходе исследования, применяются в системе анализа новостного потока принятой к использованию в ООО «Олимп» (Правительство Москвы) и в сервисе многозначной классификации поисковых запросов пользователей, принятом к использованию в ООО «Rambler&Co», а также в учебном процессе по курсу «Управление знаниями» кафедры информационных систем Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики.

Апробация результатов работы. Результаты диссертационного исследования представлялись на международных научно-методических конференциях «Современное образование: содержание, технологии, качество» (Санкт-Петербург, 2013, 2015), международной конференции «Региональная информатика» (Санкт-Петербург, 2014), межрегиональной конференции «Информационная безопасность регионов России» (Санкт-Петербург, 2015), международной научной конференции научной «Корпусная лингвистика» (Санкт-Петербург, 2015, 2017), международной конференции ассоциации открытых инноваций FRUCT: FRUCT 20 (Санкт-Петербург, 2017). По разработанной системе было получено свидетельство о регистрации

программы для ЭВМ «Система для анализа текстовых документов с использованием вероятностного тематического моделирования // Карпович С.Н.» №2017615118 от 3 мая 2017.

Публикации. По теме диссертационной работы опубликовано 13 печатных работ, включая 3 работы в журналах из списка ВАК («Труды СПИИРАН», «Информационно-управляющие системы») и 1 работа в международном издании, индексирующимся в реферативных базах Web of Science и Scopus.

Структура и объем работы. Диссертация объемом 153 машинописных страниц, содержит введение, четыре главы и заключение, список литературы (117 наименований), 14 таблиц, 51 рисунок, одно приложение с копиями актов внедрения.

Краткое содержание глав

Первая глава посвящена методам вероятностного тематического моделирования как одному из перспективных направлений обработки текстов на естественном языке. Рассмотрены основные виды ВТМ и методы оценки качества вероятностного тематического моделирования. Глава содержит обзор существующих исследований в области вероятностного тематического моделирования, на основе которого выявлены существующие проблемы ВТМ. Уделено внимание применению вероятностного тематического моделирования в практических задачах.

Во второй главе определены требования к разрабатываемому комплексу программных средств вероятностного тематического моделирования потока текстовых документов. Предложена концептуальная схема программного комплекса. Определены требования к корпусу текстов для построения ВТМ. Предложен технологический процесс создания корпуса.

В третьей главе проведен обзор существующих методов обучения ВТМ и алгоритмов многозначной классификации. Предложен алгоритм многозначной классификации текстовых документов с использованием вероятностного тематического моделирования ml-PLSI. Выполнен обзор существующих подходов к

расширению словаря ВТМ. Реализован метод определения тем «нового слова», в котором тематический вектор «нового слова» рассчитывается через произведение Адамара тематических векторов документов, где это слово встретилось. Разработан алгоритм, позволяющий расширять словарь ВТМ.

Четвертая глава посвящена проектированию комплекса программных средств вероятностного тематического моделирования для анализа текстовых документов и рассмотрению методов применения системы для решения прикладных задач интеллектуального анализа текстов. Разработана архитектура комплекса, схемы отдельных частей, предложен сценарий использования ВТМ в задачах информационного поиска и в рекомендательных сервисах.

1. Анализ существующих подходов к построению вероятностных тематических моделей

Глава посвящена методам вероятностного тематического моделирования как одному из перспективных направлений обработки текстов на естественном языке. Рассмотрены основные классы ВТМ и методы оценки качества вероятностного тематического моделирования. Глава содержит обзор существующих исследований в области вероятностного тематического моделирования, на основе которого выявлены существующие проблемы ВТМ. Уделено внимание применению вероятностного тематического моделирования в практических задачах и сформулированы требования к системам анализа текстовых документов.

1.1 Введение в вероятностное тематическое моделирование

Вероятностная тематическая модель (ВТМ) корпуса текстов определяет, к каким темам относится каждый документ и какие термины образуют каждую тему. На рисунке 1 представлена концептуальная модель построения ВТМ.

ВТМ и сокращение размерности от пространства терминов в пространство тем помогает разрешить полисемию и синонимию терминов, а также находят свое применение в задачах информационного поиска, классификации, суммаризации, в алгоритмах машинного обучения и обработки естественного языка. ВТМ используют для анализа коллекций и потоков текстовых документов. Интуитивно понимая, что документ относится к одной или нескольким темам, документы одной темы имеют схожий словарный состав. Например, слова «Правительство» и «Госдума» встречаются в политических новостях, а «футбол» и «хоккей» в спортивных, предлоги в равной доле встречаются в обоих темах. Новость обычно относится к нескольким темам, в разных пропорциях. ВТМ определяет математическую структуру тем документа на основе частотных характеристик слов этого документа.

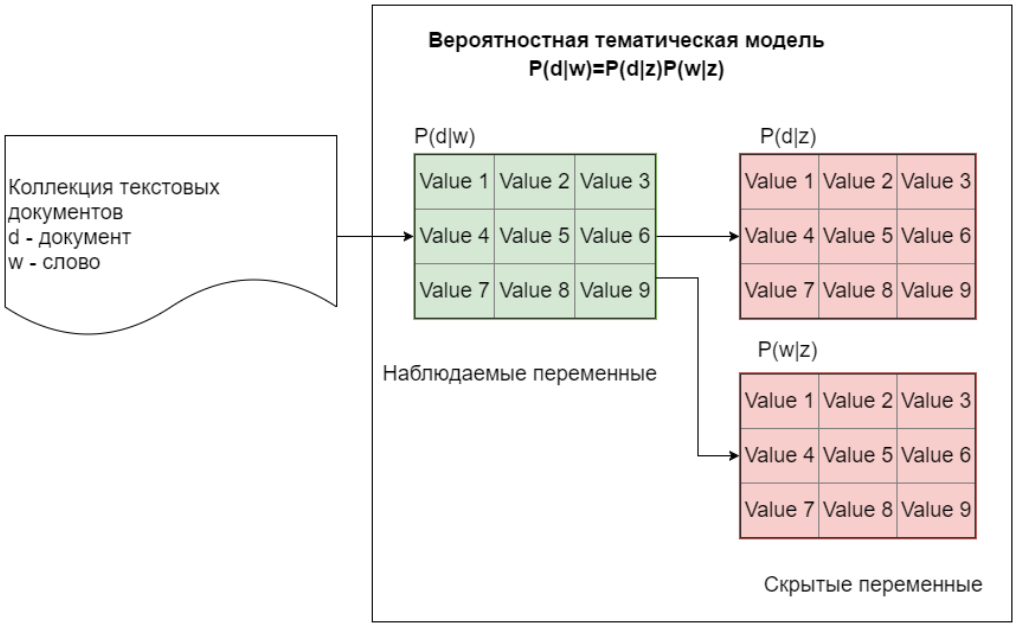


Рисунок 1 – Концептуальная модель вероятностного тематического моделирования. Где P(w|z) – матрица искомых условных распределений слов по темам; P(d|z) матрица искомых условных распределений тем по документам; d – документ; w – слово; d, w – наблюдаемые переменные; z – тема (скрытая переменная)

Первое упоминание тематического моделирования появилось в работе Рагавана, Пападимитриу, Томаки и Вемполы 1998 году [89]. Томас Хофманн в 1999 [71] году предложил вероятностное скрытое семантическое индексирование (PLSI). Вероятностный латентно-семантический анализ (PLSA), в отличии от классических методов кластеризации, основанных на функции расстояния, использует принцип максимума правдоподобия. В 2002 году была предложена одна из самых распространенных тематических моделей — это латентное размещение Дирихле (LDA) [51], которая является обобщением вероятностного семантического индексирования и разработана Дэвидом Блеем, Эндрю Ыном и Майклом Джорданом. Другие тематические модели, как правило, являются расширением LDA. В 2014 году К. Воронцов предложил Аддитивную регуляризацию для тематических моделей [100].

Тематические исследования

На основе предложенных алгоритмов вероятностного тематического моделирования было проведено множество исследований, коллекций и архивов текстовых документов. Тэмплтон [97] сгруппировал работы по тематическому моделированию в гуманитарных науках по синхронному и диахроническому принципу. Синхронные ВТМ определяют темы в некоторый момент времени, например, Джокерс исследовал, о чём писали блогеры в День Цифровых Гуманитарных наук в 2010 году.

Диахронические ВТМ рассматривают историческое развитие языка: Блок и Ньюман о временной динамике тем в Пенсильванской газете 1728—1800 года [84]; Грифитс и Стейверс анализ изменения популярности тем в журнале PNAS с 1991 по 2001 год [67]; Блевин ВТМ дневника Марты Балладс [53]; Мимно анализ 24 журналов по классической филологии и археологии за 150 лет [81].

Методы тематического моделирования

Наиболее популярный метод построения ВТМ – латентное размещение Дирихле, рассмотрен в работе Дэвида Блея «Введение в тематическое моделирование» [52]. На практике используется одна из эвристик метода максимального правдоподобия, методы сингулярного разложения (SVD), метод моментов, алгоритм, основанный на неотрицательной матрице факторизации (NMF) [107], вероятностные тематические модели, вероятностный латентно-семантический анализ, латентное размещение Дирихле. В работе [5] рассмотрены методы построения ВТМ: робастные, динамические, иерархические, многомодальные, многоязычные тематические модели и модели текста как последовательности слов.

Пусть D – множество текстовых документов (корпус текстов), W – множество слов, из которых состоят документы (словарь). Каждый документ $d \in D$ представляет собой последовательность n_d слов $(w_1, w_2, \dots, w_{n_d})$ из словаря W. Предполагается,

что существует конечное множество тем Z, и каждое употребление слова w в каждом документе d связано с некоторой темой $z \in Z$, которая неизвестна.

Предположения, на которых основаны BTM [5, 6, 21]:

- порядок документов не имеет значения, гипотеза «мешка документов»;
- порядок слов в документе не имеет значения, гипотеза «мешка слов»;
- слова, часто встречающиеся в большинстве документов, не важны для определения тематики;
- коллекция документов рассматривается как множество троек (d, w, z), документ, слово, тема, $d \in D, w \in W, z \in Z$;
- каждая тема $z \in Z$ описывается неизвестным распределением $p(w|z)$ на множестве слов $w \in W$;
- каждый документ $d \in D$ описывается неизвестным распределением $p(z|d)$ на множестве тем $z \in Z$;
- гипотеза условной независимости. Появление слов в документе описывается общим распределением $p(w|z)$ и не зависит от документа $p(w|z,d) = p(w|z); p(d|w,z) = p(d|z); p(d,w|z) = p(d|z)p(w|z)$.

Построить тематическую модель – значит найти множество темZ, распределения $\Phi = \{p(w|z)\}$ для всех тем и распределения $\Theta = \{p(z|d)\}$ для всех документов коллекции D. Развитие алгоритмов BTM направлено на замену этих предположений более реалистичными.

Вероятностный латентно-семантический анализ (PLSA)

Вероятностный латентно-семантический анализ (PLSA) был предложен в [5, 32, 33, 71]. Три эквивалентных способа записи вероятностной модели появления пары «документ-слово»:

$P(d,w) = \sum_{z \in Z} P(z)P(w|z)P(d|z) = \sum_{z \in Z} P(d)P(w|z)P(z|d) = \sum_{z \in Z} P(w)P(z|w)P(d|z),$

где Z – множество тем, $p(z)$ – неизвестное априорное распределение тем во всей коллекции, $p(d)$ – известное априорное распределение на множестве документов,

эмпирическая оценка $p(d) = \frac{n_d}{n}$, где $n = \sum_d n_d$ – суммарная длина всех документов, $p(w)$ – известное априорное распределение на множестве слов, эмпирическая оценка $p(w) = \frac{n_w}{n}$, где n_w – число вхождений слова w во все документы. На рисунке 2 представлена графическая модель вероятностного латентно-семантического анализа.

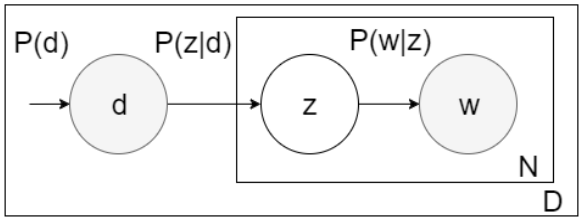


Рисунок 2 – Графическая модель вероятностного латентно-семантического анализа (PLSA). Где d – документ; w – слово; d, w – наблюдаемые переменные; z – тема (скрытая переменная); $p(d)$ – априорное распределение на множестве документов; $p(w|z), p(z|d)$ – искомые условные распределения; D – коллекция документов; N – длина документа в словах

Восстановление скрытых распределений $P(w|z), P(z|d)$ тематической модели осуществляется с помощью принципа максимума правдоподобия:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} n_{dw} log P(d, w) \rightarrow \max_{\Phi, \Theta}, \quad (1)$$

где n_{dw} – частотность слова w в документе d, $\Phi = P(w|z)$ – матрица скрытых распределений $P(w|z)$, $\Theta = P(z|d)$ – матрица скрытых распределений $P(z|d)$, при ограничениях нормировки:

$$\sum_w p(w|z) = 1, \sum_z p(z|d) = 1, \quad (2)$$

для решения задачи применяется EM-алгоритм [60]. Это итеративный, двухшаговый алгоритм:

– На E-шаге (Ожидание) применяется формула Байеса для расчёта условных вероятностей $P(z|d, w)$ всех тем z, документов d и слов w по текущим значениям параметров $P(w|z), P(z|d)$:

$$P(z|d,w) = \frac{P(w,z|d)}{P(w|d)} = \frac{P(w|z)P(z|d)}{P(w|d)}. \quad (3)$$

– На М-шаге (Максимизация) вычисляются новые оценки условных вероятностей параметров $P(w|z), P(z|d)$ по условным вероятностям тем $P(z|d,w)$:

$$\Phi = \frac{n_{wz}}{n_z} = \frac{\sum_{d\in D} n_{dwz}}{\sum_{d\in D}\sum_{w\in d} n_{dwz}};\ \Theta = \frac{n_{zd}}{n_d} = \frac{\sum_{w\in d} n_{dwz}}{\sum_{w\in W}\sum_{z\in Z} n_{dwz}}. \quad (4)$$

ЕМ шаги повторяются до сходимости. В работе [61] представлено теоретическое обоснование эквивалентности метода PLSA и неотрицательной матричной факторизации (NMF) [107], который минимизирует расстояние Кульбака-Лейблера.

Недостатки BTM построенных методом PLSA:

- Переобучение модели за счет линейного роста числа параметров по числу документов в исследуемой коллекции.
- Необходимость перестраивать модель с каждым добавлением нового документа d для расчёта распределения $p(t|d)$.

Латентное Размещение Дирихле (LDA)

Метод Латентного Размещения Дирихле (LDA) [51] основан на той же вероятностной модели что и PLSA:

$$p(d,w) = \sum_{z\in Z} p(d)p(w|z)p(z|d). \quad (5)$$

При дополнительных предположениях:

- вектора документов $\theta_d = (p(z|d): z \in Z)$ порождаются одним и тем же вероятностным распределением на нормированных $|Z|$ мерных векторах, это распределение удобно взять из параметрического семейства распределений Дирихле $Dir(\theta,\alpha), \alpha \in R^{|Z|}$;
- вектора тем $\varphi_z = (p(w|z): w \in W)$ порождаются одним и тем же вероятностным распределением на нормированных векторах размерности

$|W|$, это распределение удобно взять из параметрического семейства распределений Дирихле $Dir(\varphi, \beta), \beta \in R^{|W|}$.

Для предотвращения переобучение используется байесовская регуляризация, основанная на априорном распределении Дирихле.

$$Dir(\theta_d; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_w(\alpha_z)} \prod_z \theta_{zd}^{\alpha_z - 1}, \alpha_z > 0, \alpha_0 = \sum_z \alpha_z, \theta_{zd} > 0, \sum_z \theta_{zd} = 1, \quad (6)$$

$$Dir(\varphi_z; \beta) = \frac{\Gamma(\beta_0)}{\prod_w(\beta_w)} \prod_w \varphi_{wz}^{\beta_w - 1}, \beta_0 = \sum_w \beta_w, \varphi_{wz} > 0, \sum_w \varphi_{wz} = 1, \quad (7)$$

где $\Gamma(x)$ – гамма функция. Графическая модель LDA представлена на рисунке 3.

Для определения параметров модели LDA по коллекции документов используется сэмплирование Гиббса, вариационный байесовский вывод или метод распространения ожидания и EM-алгоритм.

Преимущества распределения Дирихле как байесовского регуляризатора вероятностных тематических моделей. Распределения Дирихле являются параметрическим семейством распределений на единичном симплексе, которое описывает как разреженные, так и сконцентрированные дискретные распределения. Модель LDA хорошо подходит для описания кластерных структур. Чем меньше значения гиперпараметров α, β, тем сильнее разрежено распределение Дирихле, и тем дальше стоят друг от друга порождаемые векторы. Чем меньше α, тем сильнее различаются документы θ_d. Чем меньше β, тем сильнее различаются темы φ_z. Векторы $\varphi_z = p(w|z)$ в пространстве терминов $R^{|W|}$ представляют центры тематических кластеров. Элементами кластеров являются векторы документов с эмпирическими распределениями $p'(w|d, z)$. Чем меньше гиперпараментры β, тем больше межкластерные расстояния по сравнению с внутрикластерными. βпозволяют моделировать тематические кластера различной степени выраженности. Также распределение Дирихле является сопряженным к мультиномиальному, что упрощает вывод апостериорных оценок вероятностей $\theta_{zd}, \varphi_{wz}$.

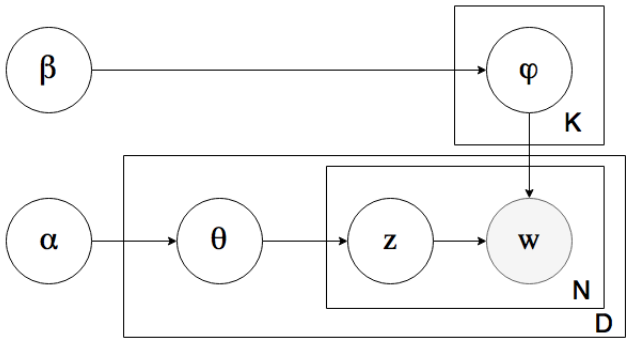


Рисунок 3 – Графическая модель Латентного размещения Дирихле LDA. Где w – слово (наблюдаемая переменная); z – тема (скрытая переменная); D – коллекция документов; N – длина документа в словах; K – количество тем в коллекции; θ – распределение тем в документе; φ – распределение слов в теме

Однако метод LDA имеет и недостатки, отмеченные в работах [5, 6]. Априорные распределения Дирихле и их обобщения – процессы Дирихле и Питмана-Йора – имеют слабые лингвистические обоснования и не моделируют явления естественного языка. Также параметры $\theta_{zd}, \varphi_{wz}$ не могут обращаться в нуль, что противоречит гипотезе разреженности.

Аддитивная регуляризация тематических моделей ARTM

В работе [6] был предложен подход к регуляризации вероятностного тематического моделирования под названием Аддитивная регуляризация тематических моделей ARTM. В отличие от ранее описанных методов регуляризации BTM, ARTM предлагает обобщенный подход к тематическому моделированию как к задаче многокритериальной оптимизации.

В ARTM наряду с правдоподобием:

$$L(\Phi,\Theta) = ln\prod_{d\in D}\prod_{w\in d} p(w|d)^{n_{dw}} = \sum_{d\in D}\sum_{w\in d} n_{dw} ln \sum_{z\in Z} \varphi_{wz}\theta_{zd} \to \max_{\Phi,\Theta}; \quad (8)$$

$$\sum_{w\in W} \varphi_{wz} = 1, \varphi_{wz} \geq 0; \sum_{z\in Z} \theta_{zd} = 1, \theta_{zd} \geq 0; \quad (9)$$

требуется максимизировать еще r критериев $R_i(\Phi,\Theta), i = 1,\dots,r$, называемых регуляризаторами. Для многокритериальной оптимизации максимизируют линейную

комбинацию критериев $L(\Phi,\Theta), R_i(\Phi,\Theta)$ с неотрицательными коэффициентами регуляризации τ_i:

$$R(\Phi,\Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi,\Theta), L(\Phi,\Theta) + R(\Phi,\Theta) \to \max_{\Phi,\Theta}. \quad (10)$$

Решение задачи находится с помощью EM-алгоритма, в котором модифицирован M-шаг:

$$\varphi_{wz} \propto \left(n_{wz} + \varphi_{wz}\frac{\partial R}{\partial \varphi_{wz}}\right)_+ ; \theta_{zd} \propto \left(n_{dz} + \theta_{zd}\frac{\partial R}{\partial \theta_{zd}}\right)_+ ; \quad (11)$$

где $(z)_+ = \max\{x, 0\}$. Функция $R(\Phi,\Theta)$ должна быть непрерывно дифференцируема, она может быть суммой нескольких регуляризаторов. ARTM может быть получен из PLSA или LDA заменой формулы на M-шаге. В работе рассмотрены сглаживающий, разреживающий, декоррелирующий, ковариационный и другие регуляризаторы, позволяющие улучшить качество BTM.

1.2. Виды вероятностных тематических моделей

Классические вероятностные тематические модели позволяют решать следующие практические задачи:

- кластеризацию текстовых документов;
- поиск похожих по теме документов.

При этом существуют другие виды вероятностных тематических моделей, направленных на решение конкретных практических задач. В работах [5, 59] был сделан обзор нескольких направлений вероятностного тематического моделирования.

Автор-тематическая модель

В 2004 году была предложена модификация LDA для совместного анализа тем и авторов документов. Модель получила название Автор-тематическая модель (The author-topic model) [91]. Основная идея заключается в поиске зависимостей между авторами, темами и документами с помощью скрытого тематического слоя, для

выявления интересов авторов. Ориентируясь на предположение, что каждый автор обладает определенным набором знаний, личным словарным запасом и преимущественно пишет на интересующие его темы, вероятностная тематическая модель может выявить закономерности между интересами авторов и документами.

Графическое представление модели представлено на рисунке 4. Мультиномиальное распределение φ в пространстве слов определяет каждую тему. Мультиномиальное распределение θ в пространстве тем определяет авторов. Для распределений φ, θ определены симметричные априорные распределения Дирихле с параметрами α, β. На первом шаге для каждого слова в документе выбирается автор, затем тема z, соответствующая автору и после выбирается слово w из тематического распределения, соответствующего теме z:

$$p(w|a, d, \varphi, \theta) = \sum_{z=1}^{z} p(w|z, \varphi_z) p(z|a, \theta_a). \quad (12)$$

Для обучения модели используется сэмплирование Гиббса.

В автор-тематической модели два набора скрытых переменных z, x. Каждая пара (z_i, x_i) переменных обуславливается всеми другими переменными:

$$P(z_i = j, x_i = k|w_i = m, z_{-i}, x_{-i}, w_{-i}, a_d) \propto \frac{C_{mj}^{WZ} + \beta}{\sum_{m'} C_{m'j}^{WZ} + V\beta} \frac{C_{kj}^{AZ} + \alpha}{\sum_{j'} C_{kj'}^{AZ} + Z\alpha}, \quad (13)$$

где $z_i = j$ и $x_i = k$ представляют отношение i слова в документе к теме j и автору k соответственно. $w_i = m$ представляет собой наблюдение, что i слово есть m слово в словаре тематической модели V, и z_{-i}, x_{-i} представляет все темы и авторов за исключением i слова. C_{kj}^{AZ} количество раз, когда автор отнесен к теме, не включает текущий экземпляр. Случайные величины φ, θ оцениваются по следующим формулам:

$$\varphi_{mj} = \frac{C_{mj}^{WZ} + \beta}{\sum_{m'} C_{m'j}^{WZ} + V\beta}, \quad (14)$$

$$\theta_{kj} = \frac{C_{kj}^{AZ} + \alpha}{\sum_{j'} C_{kj'}^{AZ} + Z\alpha}. \quad (15)$$

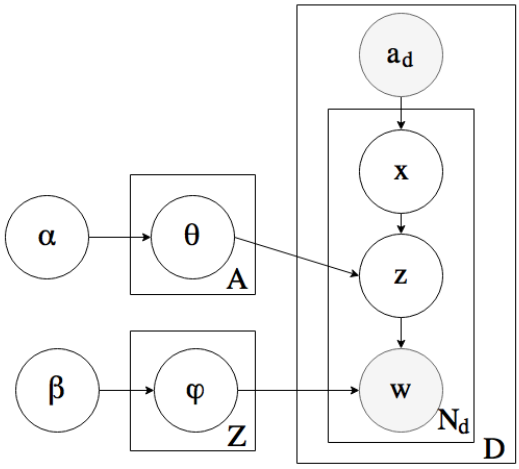


Рисунок 4 – Графическая модель автор-тематической модели

Автор-тематическая модель во времени (AToT)

Развитием автор-тематической модели стала автор-тематическая модель во времени (Author -Topic over Time, AToT), предложенная в работе [108]. Графическая модель представлена на рисунке 5. Обозначения, используемые в модели описаны в таблице 1.

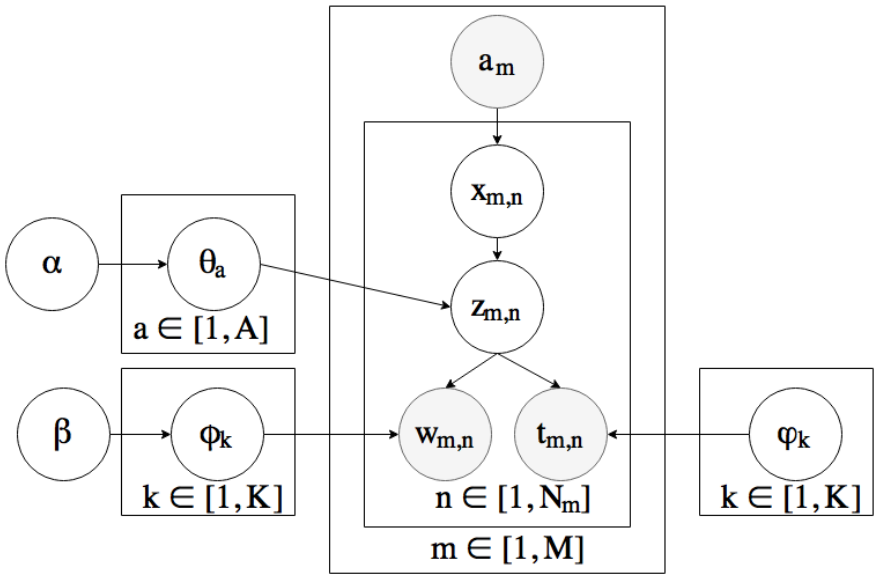


Рисунок 5 – Графическая модель AToT

Таблица 1 – Обозначения, используемые в модели ATоT

Символ	Описание
K	Количество тем
M	Количество документов
V	Количество уникальных слов
A	Количество уникальных авторов
N_m	Количество слов в документе m
A_m	Количество авторов в документе m
a_m	Авторы документа m
θ_a	Мультиномиальное распределение тем, специфичных для автора a. $\Theta = \{\theta_a\}_{a=1}^{A}$
φ_k	Мультиномиальное распределение слов, специфичных для темы k. $\Phi = \{\varphi_k\}_{k=1}^{K}$
ϕ_k	Бета распределение во времени, специфичное для темы k. $\Psi = \{\phi_k\}_{k=1}^{K}$
$Z_{m,n}$	Темы, связанные с n словом в документе m
$w_{m,n}$	n слово в документе m
$x_{m,n}$	Автор, связанный со словом $w_{m,n}$
$t_{m,n}$	Время, связанное с n словом в документе m
α	Гиперпараметр мультиномиального распределения θ
β	Гиперпараметр мультиномиального распределения φ

Формула условной вероятности для распределения P:

$$P\left(z_{m,n} = k, x_{m,n} = a \middle| w, z_{\neg(m,n)}, x_{\neg(m,n)}, t, a, \alpha, \beta, \Psi\right) \propto \quad (16)$$

$$\frac{n_k^{w_{m,n}} + \beta_{w_{m,n}} - 1}{\sum_{v=1}^{V}(n_k^v + \beta_v) - 1} \times \frac{n_a^k + \alpha_k - 1}{\sum_{k=1}^{K}\left(n_a^k + \alpha_k\right) - 1} \times Beta\left(\phi_{z_{m,n},1}, \varphi_{z_{m,n},2}\right),$$

где n_k^v количество раз, когда слово w было отнесено к теме k, и n_a^k – количество раз, когда автор a был отнесен к теме k.

ATоT является динамической автор-тематической моделью, которая позволяет отслеживать изменения интересов авторов во времени, что может найти применение в практических задачах по отслеживанию активностей блогеров.

В работе [77] была предложена модель автор-получатель, специально созданная для анализа взаимосвязей пользователей в социальных сетях. Распределение тем в документе модели зависит от автора и получателя сообщения. Что позволяет учитывать связи между авторами и получателями при построении BTM.

Вероятностные тематические модели, обучаемые с учителем

Вероятностное тематическое моделирование выполняет мягкую кластеризацию слов и документов по одному пространству тем. Кластеризация относится к методам самообучения или обучения без учителя. Однако BTM находят свое применение в задачах классификации, которые относятся к методам обучения с учителем.

В тематической модели, обучаемой с учителем (supervised Latent Dirichlet Allocation, sLDA), каждому документу соответствует метка. Метки могут быть вещественными, порядковыми или номинальными. В работе [92] метка — это рейтинг, поставленный фильму пользователем.

На рисунке 6 представлена графическая модель sLDA. Главное отличие модели от базовой LDA в том, что появляется скрытая переменная метки Y_d с математическим ожиданием η и дисперсией δ. В порождающем процессе sLDA сначала генерируется документ d, затем по нему генерируется значение y метки Y. Таким образом, метка y зависит от частоты тем, которые появлялись в данном документе. Параметры $\alpha, \beta_k, \eta, \delta$ оцениваются по обучающему корпусу текстовых документов. Для приближенной максимизации правдоподобия используется вариационный EM-алгоритм. Предложенный алгоритм подходит также для задач частичного обучения (semi-supervised learning) тематических моделей.

В работе [90] описан алгоритм тематической модели классификации под названием Labeled LDA. В основе работы алгоритма лежит базовый алгоритм LDA, векторы документов и тем порождаются одним и тем же распределением Дирихле. Модель определяет взаимное и однозначное соответствие между темами BTM и пользовательскими тэгами-метками. Метод основан на двух сильных ограничениях, темы отождествляются с тэгами-метками, предполагается, что для каждого документа точно известно множество всех тэгов-меток, к которым он относится. Графическая модель представлена на рисунке 7.

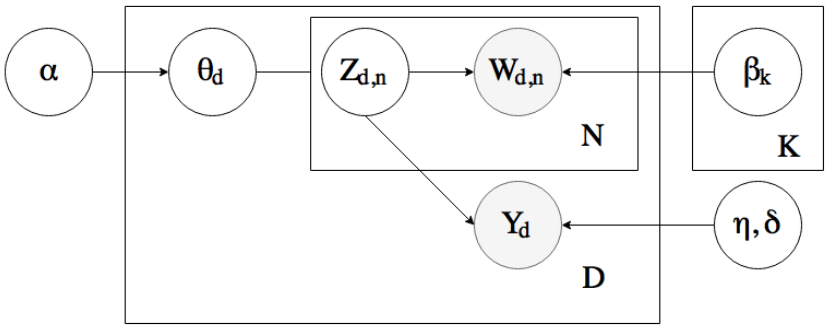


Рисунок 6 – Графическая модель sLDA. Узлы являются случайными величинами, ребра указывают на зависимость, затененный узел – наблюдаемая переменная, простой узел – скрытая переменная

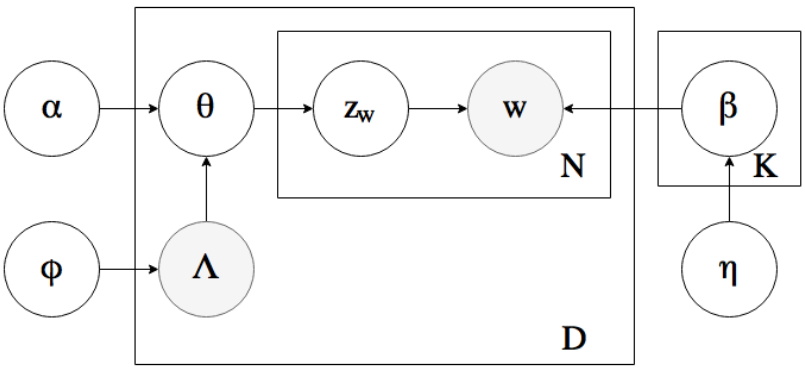


Рисунок 7 – Графическая модель Labeled LDA

Основная особенность модели Labeled LDA в том, что для каждого документа появляется наблюдаемая переменная Λ, которая соответствует тэгам пользователей. Λ – это бинарный вектор размерности K, где K – количество тегов на всем наборе данных. При этом i -я компонента Λ равна 1, если документ помечен

соответствующим i-м тэгом, 0 – в противном случае. Распределение тем в документе θ зависит не только от априорного распределения, но и от наблюдаемых меток: в каждом документе вычисляется распределение только по тем темам, которые соответствуют тэгам с $\Lambda_i = 1$, для остальных тем вероятность равна 0. Для оценивания скрытых параметров в работе [90] используется сэмплирование Гиббса, в котором каждая переменная z_w выбирается только из соответствующего набора тем.

Аналогичные ограничения касаются алгоритма Flat LDA, описанного в работе [92]. Для задачи классификации несбалансированных классов в этой работе был предложен алгоритм Prior-LDA, использующий частотную регуляризацию. По утверждению авторов работы [92], Flat-LDA, Prior-LDA являются частными случаями более общего алгоритма Dependency LDA, в котором предлагается моделировать классы документов через распределение тем документов и вводится новая неизвестная матрица класс-тема. В работе [88] предложен подход к многозначной классификации методом LDA с использованием знаний толпы под названием ML-PA-LDA-C. Используется информация не только о присутствующем классе, но и об отсутствующем, применяется для построения модели по зашумленным размеченным данным. В работе устранено одно ограничение Label-LDA, предполагается, что точно не известно множество всех классов, к которым принадлежит документ.

Вероятностные тематические модели, учитывающие зависимости между словами документа

Гипотеза «мешка слов», используемая в классических и большинстве других вероятностных тематических моделей, оправдана с точки зрения вычислительной эффективности, но далека от реальности. Учет совместно встречаемых слов, учет фраз может существенно улучшить интерпретируемость BTM. Существует два основных подхода к учету словосочетаний в BTM: создание унифицированной вероятностной тематической модели и предварительное извлечение словосочетаний [32, 33].

В работе [50] была предложена модель, которая описывает зависимости между словами документа с помощью скрытых марковских моделей (HMM), получившая название Aspect Hidden Markov Model (AHMM). Сильное ограничение модели заключается в том, что каждый документ может относиться только к одной теме. Это ограничение было впоследствии снято в работе [68], где был предложен метод HMM-LDA. Каждое предложение модели HMM-LDA разбивается на значимые слова, за генерацию которых отвечает скрытая марковская модель, и на простые слова, генерируемые BTM LDA. Ограничение модели состоит в том, что для определения темы не используется дополнительная информация, заключенная в локальной структуре текста. В работе [101] была предложена биграммная тематическая модель (bigram topic model), в которой используются скрытый тематический слов и предшествующие слова (марковская структура слов) для порождения слова. Биграммная тематическая модель (БТМ) и HMM-LDA описывают локальные синтаксические и глобальные семантические зависимости между словами в документах.

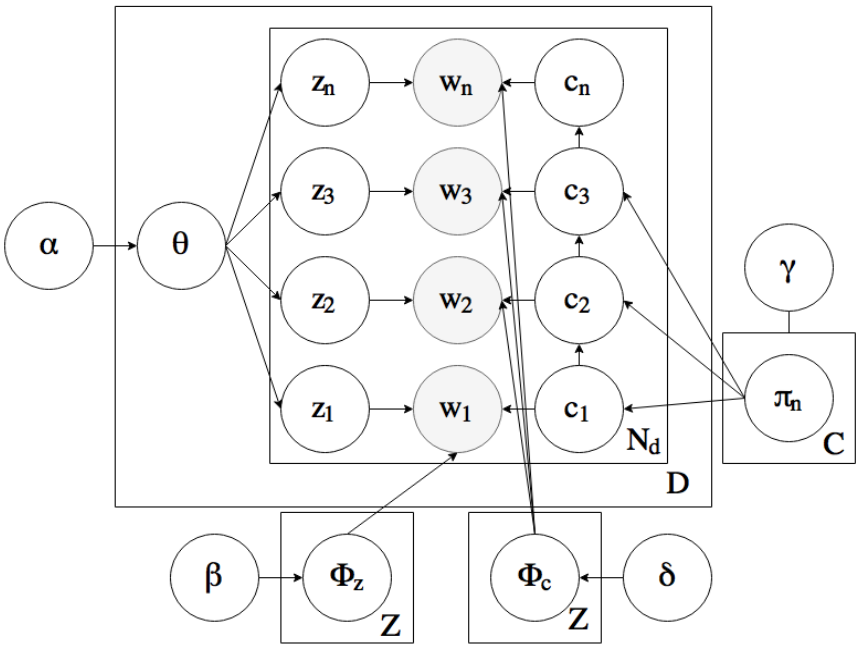


Рисунок 8 – Графическая модель HMM-LDA

Графическая модель для HMM-LDA представлена на рисунке 8. Скрытая марковская модель (HMM) описывает закономерности между соседними словами, модель LDA задает тематическое описание документа. Модель состоит из слов w, тем z и последовательности бинарных классификаций $c = (c_1, \dots, c_n)$. Если $c_i = 1$, то соответствующее слово анализируется в семантическом аспекте, т.е. на основании тематического распределения Φ_z. Если же $c_i \neq 1$, то слово порождается распределением Φ_c и не несет смысловой нагрузки. Каждому документу соответствует распределение в пространстве тем θ_d и матрица вероятностей переходов для описания переходов между классами c_{i-1}, c_i в марковской цепи.

Еще одна реализация LDA с использованием скрытых марковских моделей предложена в работе [69] под названием Скрытые тематические марковские модели (Hidden Topic Markov Models, HTMM), основанная на предположении, что последовательность тем в документе является марковской цепью. Важную роль в модели играет порядок и близость слов. Графическая модель HTMM представлена на рисунке 9.

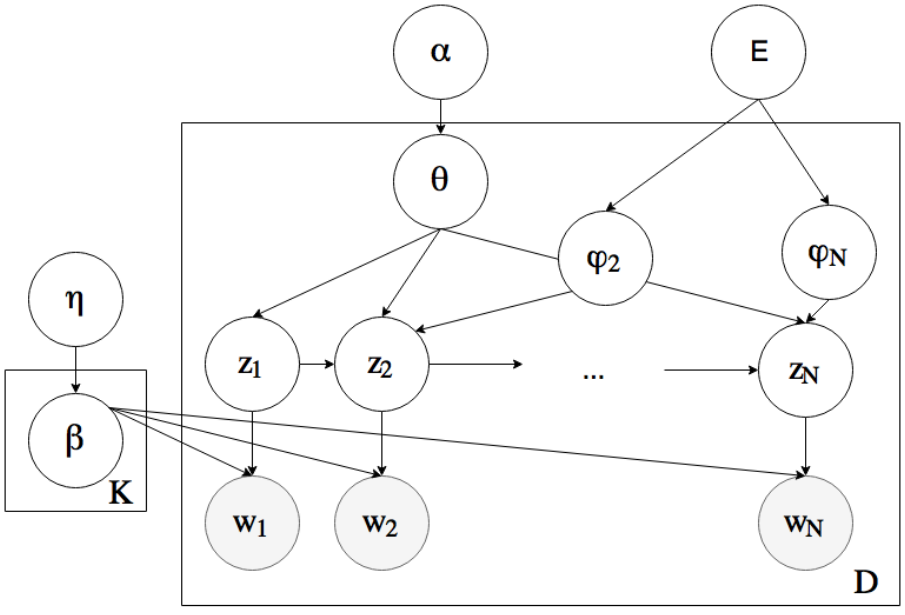


Рисунок 9 – Графическая модель HTMM. Где K – число тем, N – длина документа d, D – коллекция документов

Темы в документе образуют цепь Маркова с вероятностью перехода зависящей от θ и переменной перехода φ_n. Когда $\varphi_n = 1$, новая тема берется из θ. Когда $\varphi_n = 0$ тема n-ого слова идентична предыдущей. Предполагается, что переходы между темами могут происходить только между предложениями, так что φ_n может быть отличным от нуля только для первого слова в предложении. В результате каждое предложение целиком относится к одной теме и соседние предложения, скорее всего, относятся к этой же теме. Это позволяет точнее относить слова к темам с учетом полисемии и приводит к более интерпретируемым темам, в сравнении с LDA.

В работе [103] представлена N-граммная тематическая модель, которая к предыдущим моделям добавляет возможность формирования словосочетаний в текстах в зависимости от контекста. Графическая модель представлена на рисунке 10.

Процесс порождения текстовой коллекции следующий:

1. Для каждой темы z из распределения Дирихле β построить распределение Φ_z.
2. Для каждой темы z и каждого слова w из Бета-распределения γ построить распределение Ψ_{zw}.
3. Для каждой темы z и каждого слова w из распределения Дирихле δ построить распределение σ_{zw}.
4. Для каждого документа d из распределения Дирихле α построить распределение θ_d. Затем для каждого слова w_i в документе d:
    a. выбрать x_i из распределения $\Psi_{t_{i-1},w_{i-1}}$;
    b. выбрать тему z_i из распределения θ_d;
    c. если $x_i = 1$, то выбрать слово w_i из распределения $\sigma_{z_i w_{i-1}}$. Иначе выбрать w_i из распределения Ψ_{z_i}.

У всех описанных выше моделей, учитывающих зависимости между словами документа, большое количество параметров для настройки, что ограничивает их использование в практических задачах. Число параметров в Биграммной

тематической модели равно W^2Z, у N-граммной тематической модели – W^NZ. Число параметров в классических BTM LDA равно WZ, PLSA – $WZ + DZ$, где W – размер словаря тематической модели, D – число документов, T – число тем, N – размер N-грамм.

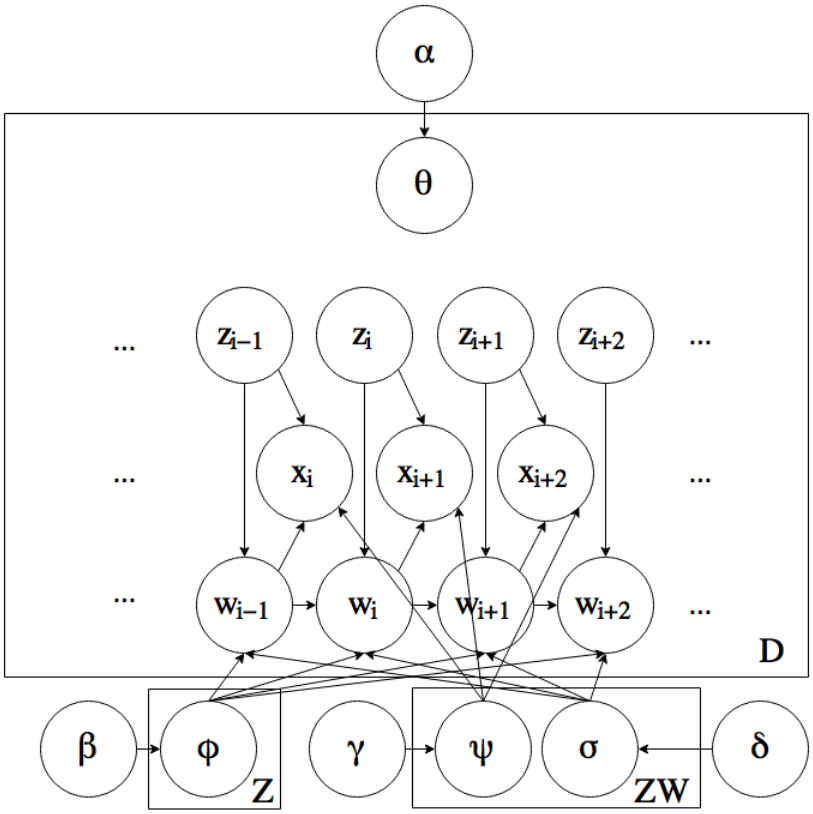


Рисунок 10 – Графическая модель N-граммной тематической модели

В работе [32] предложен алгоритм PLSA-SIM, являющийся модификацией PLSA, позволяющий добавлять биграммы и учитывать сходство между биграммными и униграммными компонентами. Модификация коснулась EM-алгоритма, в котором в BTM добавляются предварительно вычисленные множества похожих слов и биграмм. Данная модификация не увеличивает число параметров модели PLSA, остается $WZ + DZ$. Отличительной особенностью алгоритма является учет внутренней структуры биграмм и между биграммами и униграммами. Если похожие униграммы и биграммы встречаются вместе в одном документе, то PLSA-SIM старается отнести их к одним и тем же темам, предполагая, что они обладают семантической и тематической

близостью. Если униграммы и биграммы вместе в одном документе не встречаются, то исходный алгоритм не модифицируется, предполагая, что они обладают семантическими различиями.

Тематические модели, учитывающие время

Большинство модификаций BTM рассчитаны для работы со статичными данными, однако на практике чаще формулируются задачи по отношению к потоковым данным, задачи, в которых необходимо учитывать время создания документа или время описанных в тексте документа событий. Поэтому темпоральные онлайн и динамические вероятностные тематические модели имеют повышенную значимость для практического применения.

Пример модели, использующей дату под названием «Тематики во времени» (Topic over Time - TOT) представлен в работе [104]. При построении временной модели наряду со стандартными распределениями слов по темам и тем по документам оцениваются распределения каждой темы по времени, что позволяет проследить и отобразить динамику изменения тем во времени. Графическая модель представлена на рисунке 11. Где: Z – количество тем, D – количество документов, N_d – количество слов в документе d, θ_d – мультиномиальное распределение тем, специфичных для документа d, φ_z – мультиномиальное распределение слов, специфичных для темы z, Ψ_z – бета распределение по времени специфичное для темы z, z_{di} – тема, связанная с i-ым словом документа d, w_{di} – i-ое слово документа d, t_{di} – время, связанное с i-ым словом в документе d. Вместо моделирования последовательности изменения состояний с предположениями Маркова о динамике, TOT-модели (нормированные) абсолютные значения времени. Это позволяет отслеживать изменения тем в большие временные промежутки, прогнозировать абсолютные значения времени и прогнозировать распределение тем с учетом времени. Избегая дискретизации, в TOT каждая тема связана с непрерывным распределением по времени. Для построения модели используется сэмплирование Гиббса.

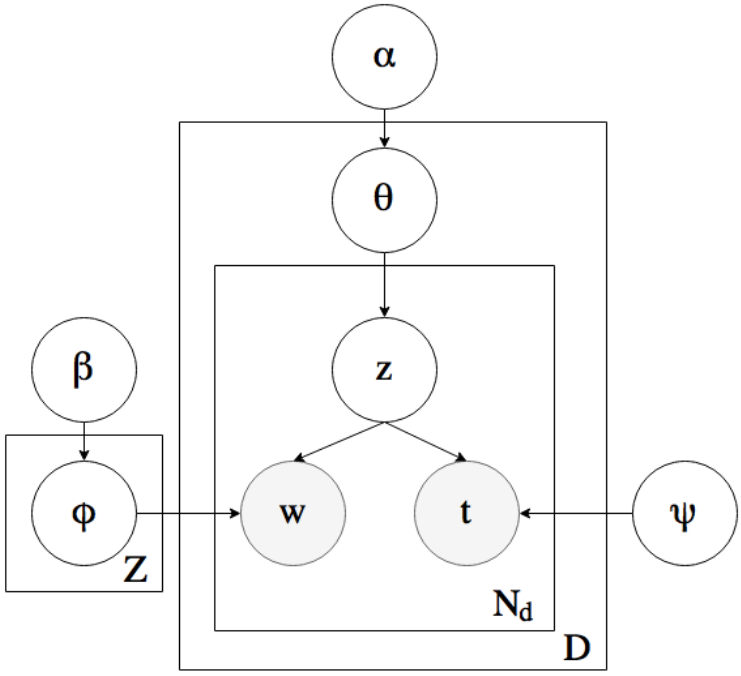


Рисунок 11 – Графическая модель TOT

Развитием алгоритма TOT стал алгоритм PAMTOT, предложенный в работе [75]. В модели используется четырехуровневое распределение Пачинко. Для расчета параметров модели используется сэмплирование Гиббса. Графическая модель представлена на рисунке 12.

В работе [49] предложена «Динамическая тематическая модель» (Dynamic Topic Model), позволяющая отслеживать эволюцию тем в последовательно организованном корпусе документов. Время предполагается дискретным, в котором отслеживаются повышения и понижения популярности тем. Для каждого интервала времени имеются свои распределения тем в документе θ и распределения слов в темах β, эти распределения зависят от соответствующих распределений в предыдущем временном отрезке. Графическая модель «Динамической тематической модели» изображен на рисунке 13.

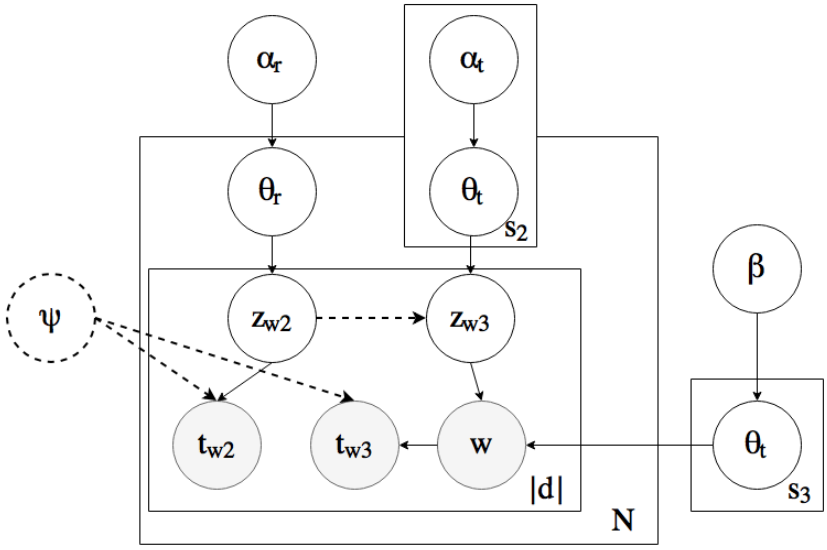


Рисунок 12 – Графическая модель для четырех уровневой PAMTOT.
Пунктирные линии показывают, как выбираются временные метки из тем.
Сплошные линии соответствуют PAM без времени

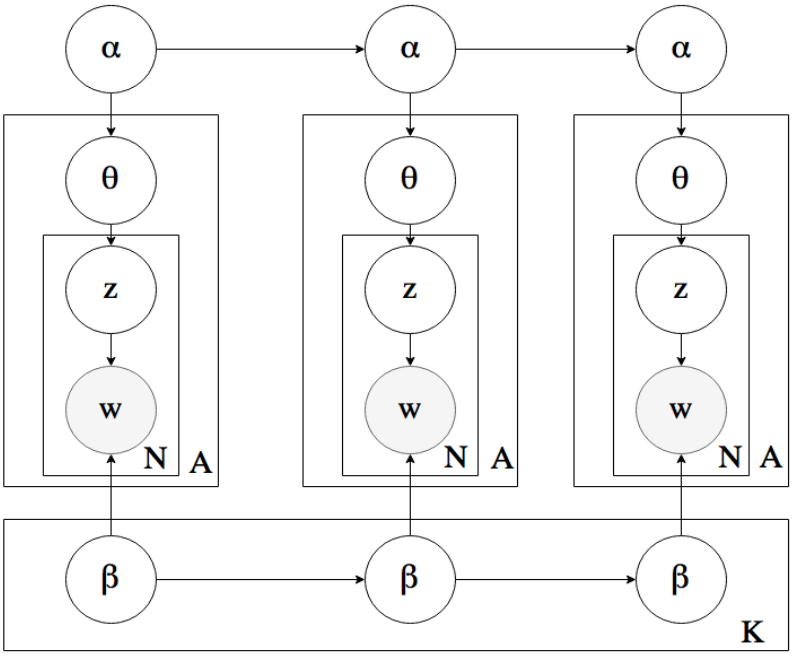


Рисунок 13 – Графическая модель для динамической тематической модели DTM

Альтернативный подход предложен в работе «Многомасштабная томографическая модель» (Multiscale Topic Tomography) [82], где используются сопряженные распределения одновременно для нескольких масштабов времени. В работе [105] предложена «динамическая тематическая модель с непрерывным временем» (Continuous Time Dynamic Topic Models), основанная на использовании броуновского движения для моделирования эволюции тем в непрерывном времени, что позволяет снять некоторые ограничения DTM. Очередным расширением LDA стала «онлайн LDA» (Online Learning for Latent Dirichlet Allocation), предложенная в работе [70]. В выше описанных работах представлены подходы к построению динамических тематических моделей с фиксированным словарем. Эти ТМ позволяют проследить изменение тематики во времени, но не позволяют оценить изменение словарного состава модели.

В работе Online Latent Dirichlet Allocation with Infinite Vocabulary [110] описан алгоритм создания онлайн ТМ с бесконечным словарем. Основное отличие модели заключается в том, что мультиномиальное распределение берется из бесконечного процесса Дирихле по всем возможным словам вместо конечного распределения Дирихле. В словарь добавляются новые слова с каждой итерацией добавления новых документов в ТМ, но количество слов в словаре ограничено заданным значением, поэтому словарь является усеченным упорядоченным множеством вероятностного распределения слов. Таким образом, словарный состав ТМ не меняется в размере, но изменяется по своему составу по мере добавления новых документов.

В работе On-line Trend Analysis with Topic Models: #twitter trends detection topic model online [74] рассмотрен алгоритм построения онлайн ТМ с изменяемым словарем. Время в модели дискретно, документы, поступающие в ТМ, разбиты на временные отрезки, временной отрезок может быть равен часу, дню, месяцу или году. Вводится понятие «окна», которому соответствует набор документов из нескольких временных отрезков. В модели документы обрабатываются итеративно в рамках

одного «окна». Одним из ключевых отличий этого подхода от OLDA, представленного в работе [70], является передача параметров от ранее построенной модели в новую. Второе отличие — это динамически изменяющийся словарь ТМ, где тематические векторы слов, которые присутствовали в предыдущем «окне», рассчитываются по данным старой модели, а тематические векторы новых слов берутся из равномерного распределения Дирихле. Новые слова, встретившиеся более 10 раз, добавляются в модель, а старые слова, встретившиеся менее 10 раз в «окне», удаляются из ТМ. Отсутствует ограничение на размер словаря ТМ, поэтому он может изменяться по размеру и составу в ходе добавления новых документов. Улучшение модели для анализа сообщений в социальной сети Twitter было предложено в работе [93] под названием TwitterTTM, которая может фиксировать динамику интересов пользователей и способна к интерактивному взаимодействию.

Некоторые виды вероятностных тематических моделей

Многоязычная тематическая модель предложена в работе [12], в которой на базе ARTM строится вероятностная тематическая модель одновременно учитывающая двуязычный словарь и связи между документами параллельной или сравнимой коллекции. Модель является развитием направления многоязычных моделей: ML–LDA (MultiLingual LDA) [87], PLTM (PolyLingual Topic Model) [79] и BiLDA (Bilingual LDA) [94]. Многоязычные тематические модели решают важную практическую задачу кросс-язычного поиска, когда запросом является документ на одном языке, а поиск производится среди документов других языков.

Непараметрические модели. При решении практических задач зачастую невозможно предсказать количество тем, для решения таких задач предназначены непараметрические вероятностные тематические модели. В работах [96, 102] предложен Иерархический процесс Дирихле (Hierarchical Dirichlet Process, HDP), который является Байесовской непараметрической моделью, позволяющей моделировать бесконечное число тем. Иерархический процесс Дирихле можно

интерпретировать как процесс Дирихле над процессом Дирихле. При генерации случайного элемента из иерархического процесса Дирихле сначала выбирается, из какого процесса верхнего уровня следует генерировать элемент, после чего к выбранному процессу Дирихле применяется обычная схема генерации элемента. В работе [102] предложен метод онлайн оценки параметров процесса Дирихле.

Визуализация вероятностных тематических моделей

Важным направлением разработки программ для построения ВТМ является визуализация результатов вероятностного тематического моделирования. В работе [47] предложен метод визуализации ТМ, включающий программное обеспечение с открытым исходным кодом. Основная идея метода заключается в том, что визуализация модели обобщает и организует коллекцию документов. Цели метода: суммировать информацию о тематическом моделировании для пользователя, выявить зависимости между содержанием и результатом, найти связи по контенту. Предложенный навигатор имеет два типа страниц: один для отображения тем, другой для документов. Для связи страниц и навигации пользователя используются гиперссылки, объединяющие темы и документы. Каждая тема связывает несколько документов, каждый документ связывает несколько тем. В работе [58] представлены инструменты для визуального анализа и оценки качества тематического моделирования. Предлагаемый подход содержит два основных интерфейса: отображение матрицы отношений слово-тема и отображение документа. ТМ представлена в виде матрицы, где строки – это слова, а столбцы – это темы. Система обладает продвинутым интерактивным графическим интерфейсом, реализованным с помощью JavaScript библиотеки d3.js. В работе [65] представлена система LDAExplore для анализа ТМ и визуализации результатов тематического моделирования. Основная цель работы – обеспечить пользователя интерактивной средой визуализации для исследования ТМ. Интерфейс позволяет фильтровать документы и слова – модели.

1.3. Оценка качества вероятностного тематического моделирования

Оценка качества вероятностного тематического моделирования является нетривиальной задачей. В отличие от задач классификации нет четкого понятия «ошибки». Критерии оценки качества кластеризации, такие как среднее внутрикластерное расстояние или межкластерное расстояние не подходят для оценки «мягкой» кластеризации документов и терминов. Критерием оценки качества языковых моделей является Перплексия [59]. Перплексия является мерой несоответствия модели $P(w|d)$ словам w встречаемым в текстовых документах коллекции, и определяется через логарифм правдоподобия:

$$Perplexity(D) = \exp(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} logP(w|d)), \quad (17)$$

где n – количество слов в исследуемой коллекции документов, D – множество документов коллекции, n_{dw} – частотность слова w в документе d, $P(w|d)$ – вероятность появления слова w в документе d.

Полученное значение соответствует полезному размеру словаря ВТМ. Таким образом, чем меньше перплексия, тем лучше модель обобщает данные. В работе [51] отмечено, что оценка перплексии, вычисленной по той же самой коллекции, будет заниженной. Поэтому рекомендован к использованию метод расчета контрольной перплексии [44]. В работах [57, 85] отмечено, что перплексия не является хорошим критерием оценки качества ВТМ, тем не менее ее продолжают использовать для оценки вероятностного тематического моделирования.

Другим способом оценки качества вероятностного тематического моделирования является Когерентность. Тема называется когерентной, если слова, наиболее часто встречаемые в теме, неслучайно часто встречаются рядом в документах коллекции [85, 86]. Когерентность может оцениваться по сторонней [83] или по той же самой коллекции [80].

Предлагалось несколько оценок когерентности. Поточечная взаимная информация (pointwise mutual information, PMI):

$$PMI(z) = \sum_{i=1}^{k-1} \sum_{j=1}^{k} log \frac{N(w_i, w_j)}{N(w_i)N(w_j)}, \quad (18)$$

где w_i – i-ое слово в порядке убывания φ_{wz}, $N(w)$ – число документов, в которых встречается слово w , $N(w, w')$ – число документов, в которых слова w, w' встречаются рядом (в окне заданной ширины h, обычно h полагают равным 10). Число k обычно полагается равным 10.

В работе [80] продемонстрировано, что более адекватной мерой является логарифм условной вероятности, где оценивается вероятность менее частого слова при условии более частого:

$$LCP(z) = \sum_{i=1}^{k-1} \sum_{j=1}^{k} log \frac{N(w_i, w_j)}{N(w_i)}. \quad (19)$$

Существуют и другие методы оценки качества вероятностного тематического моделирования, например, экспертные оценки, однако они менее популярны.

Следует отметить, что для проведения исследований BTM и для оценки качества вероятностного тематического моделирования необходимы текстовые корпуса на том языке, на котором строится модель. Однако русскоязычных текстовых корпусов, распространяемых по свободной лицензии и пригодных для построения BTM, не создано.

Выводы к главе 1

Проведенный анализ позволил выявить следующие проблемы:

1. Отсутствие русскоязычных текстовых корпусов, пригодных для исследования алгоритмов вероятностного тематического моделирования и включающих помимо основного текста документа дату, необходимую для построения темпоральных BTM, информацию об авторе документа,

необходимую для построения автор-тематических моделей, информацию о том к каким темам относится каждый документ, необходимую для построения ВТМ методом обучения с учителем. Наличие русскоязычных текстовых корпусов, распространяемых по открытой лицензии, позволит специалистам, исследующим особенности ВТМ, сосредоточить свои усилия на исследованиях алгоритмов обработки текстов на естественном языке и сделает более доступным изучение особенностей русского языка, а именно синонимию, полисемию, омонимию.

2. Несоответствие существующих программных решений с использованием вероятностного тематического моделирования практическим потребностям, таким как, обучение ВТМ, визуализация результатов ВТМ, экспорт ВТМ, анализ потока текстовых документов. Большинство алгоритмов вероятностного тематического моделирования используют статичный набор данных, в то время как в реальных задачах востребованы методы постоянного пополнения ВТМ новыми документами, словами и темами. При использовании классических алгоритмов ВТМ для потока текстовых документов, необходимо постоянно переобучать ВТМ с получением новых документов, так как словарь модели не пополняется новыми словами, неперестроенная модель теряет свою актуальность со временем.
3. Для анализа потока текстовых документов с помощью ВТМ, необходимо построить ВТМ на начальном наборе данных, а затем решать задачу многозначной классификации для определения к каким темам относится новый документ. Поэтому существует потребность в методах применения вероятностного тематического моделирования для решения задач многозначной классификации, а также в методах обучения ВТМ с учителем.

4. Для анализа потока текстовых документов не решен вопрос определения тем «нового слова» в ВТМ. Под «новым словом» в данной работе подразумевается слово, отсутствующее в словаре ВТМ на момент ее построения.

Решение обозначенных проблем позволит повысить эффективность применения вероятностного тематического моделирования в задачах обработки текста на естественном языке.

2. Требования к программному комплексу для построения ВТМ

2.1. Системы анализа текстов и потоков текстовых документов

Традиционно исследованием и анализом данных занимается Data Mining, используя Машинное обучение. Изучением особенностей языка и обработкой естественного языка занимается Прикладная лингвистика.

«Data Mining – исследование и обнаружение «машиной» (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком". Автор определения – Григорий Пятецкий-Шапиро [2, с. 68]. «Методы Data Mining находятся на стыке разных направлений информационных технологий и технологий искусственного интеллекта: статистики, нейронных сетей, нечетких множеств, генетических алгоритмов и др. Основными задачами Data Mining являются: классификация, регрессия, поиск ассоциативных правил и кластеризация» [2, с. 69]. Задачи делят по назначению на описательные и предсказательные. По способам решения задачи разделяют на обучаемые с учителем и обучаемые без учителя.

Основные концептуальные подходы к пониманию сущности классификации [35, с. 26]:

a. Классификация как свойство любого живого организма.
b. Классификация как эмпирическое действие или обобщение, которое не является интеллектуальной деятельностью человека;
c. Классификация как исключительно интеллектуальная деятельность человека, необходимая для понимания реальной действительности.

Классификация объектов, процессов или явлений произвольной природы базируется на двух фундаментальных действиях [35, с. 50]:

a. Отождествления – как установления степени тождества классификационных объектов;
b. Различения – как установления степени различия классификационных объектов.

«В задачах классификации и регрессии требуется определить значение зависимой переменной объекта на основании значений других переменных, характеризующих данный объект» [2, с. 102]. «Решением задачи классификации является отнесение каждого из объектов данных одному (или нескольким) из заранее определенных классов и построение в конечном счете одним из методов классификации модели данных, определяющей разбиение множества объектов данных на классы» [2, с. 160]. Задача классификации документов: дан документ, требуется отнести его к одному или нескольким предопределенным классам [16, с. 258]. Выделяют следующие методы классификации [2]:

- наивный Байесовский;
- деревья решений;
- линейные и нелинейные методы;
- метод опорных векторов (Support Vector Machines, SVM);
- регуляризационные сети;
- дискретизация и редкие сетки.

«Задача кластеризации состоит в разделении исследуемого множества объектов на группы «похожих» объектов, называемых кластерами. В задаче кластеризации отнесение каждого из объектов данных осуществляется к одному (или нескольким) из заранее не определенных классов. Разбиение объектов данных по кластерам осуществляется при одновременном их формировании. Определение кластеров и разбиение по ним объектов данных выражается в итоговой модели данных, которая является решением задачи кластеризации» [2, с. 160]. Кластеризация документов – это

группировка похожих неразмеченных документов на основе некоторой меры схожести [16, с. 213]. Различают следующие алгоритмы кластеризации [2]:

- Иерархические:
  - агломеративные,
  - дивизимные.
- Неиерархические:
  - к-средних (k-means),
  - fuzzy c-means,
  - кластеризация по Гюстафсону-Кесселю.

Похожесть объектов кластеризации и классификации определяется по свойствам этих объектов. Основным свойством текстовых документов является наличие в тексте какого-либо слова. Для компьютерного представления свойств текстовых документов вводится понятие – модель векторного пространства. «Представление множества документов в виде векторов в общем векторном пространстве называется моделью векторного пространства (vector space model) и является фундаментальным для многих задач информационного поиска, включая ранжирование документов по запросу, классификацию и кластеризацию документов» [28, с. 137].

Основные этапы анализа данных [2, с. 91]:

1. Понимание и формулировка задачи анализа.
2. Подготовка данных для автоматизированного анализа (препроцессинг).
3. Применение методов Data Mining и построение моделей.
4. Проверка построенных моделей.
5. Интерпретация моделей человеком.

Выделяют визуальный анализ данных, обнаружение данных в текстовых документах, методы анализа данных в реальном времени, анализ Веб-документов и сайтов. **Визуальный анализ данных (Visual Mining)** – это представление данных в некоторой визуальной форме, позволяющей человеку погрузиться в данные, работать

с визуальным представлением, понять их суть, сделать выводы и напрямую взаимодействовать с данными [2, с. 192]. Методы анализа в неструктурированных текстах лежат на стыке нескольких областей: Data Mining, обработка естественных языков, поиск информации, извлечение информации и управление знаниями [2, с. 211]. **Обнаружение знаний в тексте (Text Mining)** – это нетривиальный процесс обнаружения действительно новых, потенциально полезных и понятных шаблонов в неструктурированных текстовых данных. **Методы Data Mining в реальном времени (Real-Time Analytics)** в основном относятся к задаче предсказания. В отличие от статических методов они обучаются динамически и основаны на обратной связи от прогноза, полученного с помощью предсказательной модели (постоянном переобучении) [2, с. 325]. **Web Mining** – технология, использующая методы Data Mining для исследования и извлечения информации из Web-документов и сервисов [2, с. 351].

Термин **Машинное обучение** используется для обозначения всех технологий Data Mining [2]. **Машинным обучением** называется систематическое обучение алгоритмов и систем, в результате которого их знания или качество работы возрастает по мере накопления опыта [41, с. 16]. Смысл машинного обучения состоит в использовании нужных признаков для построения моделей, подходящих для решения правильно поставленных задач. Признаки определяют «язык», на котором описываются объекты предметной области. Задача – это абстрактное представление проблемы с участием объектов предметной области, которую необходимо решить. Представление в виде отображения исходных данных на результаты называют моделью. Модели обеспечивают разнообразие предмета машинного обучения, тогда как задачи и признаки придают ему единство. Различают геометрические, вероятностные и логические модели.

В прикладной лингвистике традиционные методы анализа текстов основаны на глубинной обработке естественного языка, ориентированы на взаимодействие с

хранилищами документов, изменения в которых вносятся сравнительно редко: национальными корпусами текстов, электронными библиотеками, базами научных статей или архивами веб-сайтов, словарей, тезаурусов и онтологий. Прикладная лингвистика сосредоточена на построении логико-лингвистических моделей и соответствующих им программ и алгоритмам. Направление обработки естественного языка сосредоточено на моделировании всего того, что изучает лингвистика в целом. Первоначально оба направления стремились построить точные языковые модели, основанные на формальном синтаксическом и семантическом анализе.

В СССР термин **прикладная лингвистика** стал широко употребляться в 1950-е годы в связи с разработкой компьютерных технологий и появлением систем автоматической обработки информации. В русскоязычной литературе часто используются термины «компьютерная лингвистика», «вычислительная лингвистика», «автоматическая лингвистика», «инженерная лингвистика» в том же контексте, что и прикладная лингвистика. Прикладная лингвистика понимается как деятельность по приложению научных знаний об устройстве и функционировании языка в нелингвистических научных дисциплинах и в различных сферах практической деятельности человека, а также теоретическое осмысление такой деятельности [1, с. 12]. Сущность **контент-анализа** заключается в том, чтобы по внешним (количественным) характеристикам текста на уровне слов и словосочетаний сделать правдоподобные предположения о его плане содержания и, как следствие, сделать выводы об особенностях мышления и сознания автора текста: его намерениях, установках, желаниях, ценностных ориентациях и т.д. [1, с. 265].

«Корпус данных представляет собой сформированную по определенным правилам выборку данных из проблемной области». **Корпус текстов** – это вид корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты, включающие, например, какие-то полные фрагменты макроструктуры текстов данной проблемной области [1, с. 125]. «Требования к

корпусу текстов с точки зрения пользователя: репрезентативность, полнота, экономичность, структуризация материала, компьютерная поддержка».

Корпусная лингвистика – сложная лингвистическая дисциплина, которая сформировалась в последние десятилетия на базе электронной вычислительной техники. Она изучает построение лингвистических корпусов, способы обработки данных в них и собственно технологию их создания и использования. «Корпус – это информационно справочная система, основанная на собрании текстов на некотором языке в электронной форме», - такое определение текстового корпуса дано на сайте Национального Корпуса Русского Языка [115]. В научной работе [15] отмечено: «Корпусы, как правило, предназначены для неоднократного применения многими пользователями, поэтому их разметка и их лингвистическое обеспечение должны быть определенным образом унифицированы». Целесообразность создания и смысл использования корпуса определяется следующими предпосылками:

1. достаточно большой (репрезентативный) объем корпуса;
2. данные разного типа находятся в корпусе в своей естественной контекстной форме;
3. однажды созданный и подготовленный массив данных может использоваться многократно.

Корпусами первого порядка называют собрание текстов, объединенных общим признаком, например, источник, автор, место публикации. **Специальный корпус текстов** – это сбалансированный корпус, репрезентативный, как правило, небольшой по размеру, подчиненный определенной исследовательской задаче и предназначенный для использования преимущественно в целях, соответствующих замыслу составителя. Текстовый корпус (text corpora, corpus), большая коллекция документов (large collection of documents), набор данных (dataset), как отмечено в работе, являются синонимичными понятиями.

Направление прикладной лингвистики оперирует следующими понятиями. **Морфология** – это раздел лингвистики, который изучает структуру слова и его грамматическое значение. [31]. **Семантика** – изучение смысла слов и связей между ними, благодаря которым образуются более крупные смысловые единицы [16, с. 369]. **Дискурс** базируется на семантическом уровне: дискурс-анализ ставит целью установить связи между предложениями [16]. **Прагматика** изучает, какой вклад в смысл текста вносят контекст, знания о мире, языковые соглашения и прочие абстрактные свойства [16]. **Тезаурус** – это словарь, в котором слова и словосочетания с близкими значениями сгруппированы в единицы, называемые понятиями, концептами или дескрипторами, и в котором явно (в виде отношений, иерархии) указываются семантические отношения между этими понятиями (концептами, дескрипторами) [25, с. 20]. **Онтология** – артефакт, структура, описывающая значения элементов некоторой системы [25, с. 101]. **Онтология** – это точная спецификация концептуализации, где, концептуализация – это структура реальности, рассматриваемая независимо от словаря предметной области и конкретной ситуации [25]. «Онтология в философии создается для изучения свойств реальности. Онтология в инженерии создается для решения инженерной задачи» [24, с. 19].

Задачи интеллектуального анализа текстовых документов включают:

- кластеризацию – автоматическое выделение групп похожих документов;
- классификацию – распределение текстовых документов по категориям, группам, темам заранее известным;
- обнаружение тем – обнаружение новых тем в потоке текстовых документов;
- отслеживание эволюции тем – исследование динамики тем потока текстовых документов во времени.

Треугольник интеллекта: алгоритмы (обработка), контент (данные), ссылки (знание) [29].

В практических задачах применение традиционных методов анализа текстов связано с большими затратами, требующими привлечения специалистов в области компьютерной лингвистики, создания специализированных онтологий по предметным областям коллекций документов. Для многих практических задач удовлетворительные результаты могут быть получены более простыми и вычислительно эффективными способами на основе статистических данных о текстах. Традиционные методы анализа текстов не учитывают возрастающую потребность анализировать поток текстовых документов, с постоянно изменяющимися данными, в которых появляются новые слова-термины. Одним из перспективных направлений интеллектуального анализа текстовых документов на основе дистрибутивного анализа является вероятностное тематическое моделирование. Алгоритмы вероятностного тематического моделирования являются одним из перспективных направлений машинного обучения для анализа текстов на естественном языке.

Алгоритм – это любая корректно определенная вычислительная процедура, на вход которой подается некоторая величина или набор величин, и результатом выполнения которой является выходная величина или набор значений. Алгоритм представляет собой последовательность вычислительных шагов, преобразующих входные величины в выходные. Также алгоритм можно рассматривать как инструмент, предназначенный для решения корректно поставленной вычислительной задачи [20]. **Вероятностная модель** – это способ кодирования общих знаний о неопределенной ситуации [36, с. 117]. **Система** – это некоторое целое, составленное из частей [35, с. 54].

Алгоритмы и направления анализа потока данных (stream mining) [43]:

- кластеризация потока данных (Data Stream Clustering);
- классификация потока данных (Data Stream Classification);
- извлечение частотных образцов данных (Frequent Pattern Mining);

- отслеживание изменений в потоках данных (Change Detection in Data Streams);
- анализ многомерных данных в потоке (Stream Cube Analysis of Multi-dimensional Streams);
- загрузка данных в поток (Loadshedding in Data Streams);
- анализ потоков данных с использованием скользящего окна (Sliding Window Computations in Data Streams);
- аннотирование потоков данных (Synopsis Construction in Data Streams);
- объединения процессов по обработке потоков данных (Join Processing in Data Streams);
- индексация потоков данных (Indexing Data Streams);
- уменьшение размерности и прогнозирование в потоках данных (Dimensionality Reduction and Forecasting in Data Streams);
- распределенное извлечение и исследование потоков данных (Distributed Mining of Data Streams);
- анализ потока в сетях датчиков (Stream Mining in Sensor Networks).

В работе [23] подчеркнута значимость анализа информационных потоков в условиях постоянно растущего количества новостей и электронных сообщений. Выделены модели информационных потоков: линейная, экспоненциальная, логистическая. Также выделен подход к анализу потока как дискретных сигналов.

Современные задачи интеллектуального анализа текстовых потоков включают кластеризацию, классификацию, обнаружение тематик, отслеживание эволюции тем. В данной работе под текстовым потоком понимается последовательность текстовых документов с определенным для каждого описанного события временем происшествия. Под обработкой потока текстовых документов понимается комплексная задача кластеризации поступающих документов и анализа тем этих документов.

Существует несколько научных направлений, которые заинтересованы в интеллектуальном анализе текстов и потоков текстовых документов, поэтому разработка методов, алгоритмов и программных комплексов, выполняющих такой анализ, является актуальной востребованной задачей.

Системы и программные комплексы для анализа текстовых документов

Проект АОТ [1] (Автоматическая обработка текста) – инструмент обработки текста на естественном языке. Технологии базируются на многоуровневом представлении естественного языка. Компоненты, составляющие языковую модель, друг за другом обрабатывают входной текст. Включены следующие компоненты: графематический анализ, морфологический анализ, синтаксический анализ, семантический анализ.

GATE [2]– открытый программный комплекс для обработки текстов. Включает средства для семантических аннотаций, извлечения знаний, построения и работы с онтологиями, машинное обучение. Предназначен для работы с английским языком.

Apache OpenNLP [3]– интегрированный пакет инструментов обработки текстов, работающий на основе машинного обучения. Включает средства токенизации, разбиения на предложения, морфологическую разметку, извлечение именованных сущностей, синтаксический разбор предложений.

Natural Language Toolkit [4]– пакет библиотек, написанных на языке Python, предназначенный для символьной и статистической обработки естественного языка. Ориентирован для работы с английским языком.

Stanford NLP [5]– программное обеспечение для обработки естественного языка, разработанный в Стэндфордском университете. Ориентирован для работы с текстами

[1] http://aot.ru/
[2] https://gate.ac.uk/
[3] http://opennlp.apache.org/
[4] http://www.nltk.org/
[5] https://nlp.stanford.edu/software/index.shtml

на английском языке. Реализует методы синтаксического и морфологического анализа. Представляет байесовские и регрессионные средства классификации.

RapidMiner [6]– платформа для интеллектуального анализа данных. Средства для анализа текста включают: векторный анализ текстов, вычисление сходства, методы кластеризации и классификации. Включает методы анализа потоков данных на базе скользящего окна.

Apache Mahout [7] – система машинного обучения и анализа данных. Предоставляет инструмент для построения BTM методом LDA.

Mallet [8] – кросс-платформенный комплекс программных средств для вероятностного тематического моделирования, классификации документов и разметки последовательностей разработан в Массачусетском университете. Включает методы машинного обучения и методы численной оптимизации. Написан на языке Java.

BigARTM [9] – библиотека для вероятностного тематического моделирования реализует метод APTM. Позволяет выполнять мягкую кластеризацию больших объемов текстовых документов.

Gensim [10] – библиотека с открытым исходным кодом для автоматического анализа текста. Написана на языке Python. Включает методы вероятностного тематического моделирования, классификацию, кластеризацию, алгоритм word2vec.

Figaro [11] – язык и программная библиотека для вероятностного программирования, написанная на языке Scala. Поддерживает разработку большого количества вероятностных моделей [36].

---

[6] https://rapidminer.com/
[7] http://mahout.apache.org/
[8] http://mallet.cs.umass.edu/
[9] http://bigartm.org/
[10] http://radimrehurek.com/gensim/
[11] https://github.com/p2t2/figaro

Stan [12] – популярная в области статистики система вероятностного программирования. Ориентирована на непрерывные переменные и предлагает широкий спектр распределений. Поддерживает многочисленные методы статистического вывода.

Для решения практических задач анализа потока текстовых документов, исследования алгоритмов вероятностного тематического моделирования необходимо наличие открытых человеко-ориентированных программных комплексов, в которых специалист, решающий свою прикладную задачу, сможет самостоятельно выбрать и задействовать подходящий алгоритм или программную библиотеку. Суть человеко-ориентированного программного комплекса в том, что в процессе обработки данных программный комплекс должен тесно взаимодействовать с пользователем, который принимает решение о выборе подходящих алгоритмов для решения конкретной задачи, оценивает результаты, при необходимости вносит изменения в порядок работы программного комплекса. Большинство существующих программ, нацелены на решение одной конкретной задачи, не предоставляют возможности пользователю самостоятельно внести изменения в порядок работы системы, вплоть до того, что ориентированы на работу с данными в определенном формате.

Ключевыми особенностями анализа потока текстовых документов является скорость обработки новых документов и изменение содержания понятий с течением времени, эволюция потока. Скорость анализа должна быть выше скорости поступления новых документов. Если поток текстовых документов незначительный и есть возможность с поступлением каждого нового документа полностью перестраивать тематическую модель, то острой необходимости в динамической тематической модели не наблюдается. При этом с возрастанием скорости поступления новых документов возрастает потребность в быстрых алгоритмах классификации новых документов и алгоритмах определения темы новых слов. Рассмотренные выше

[12] http://mc-stan.org/

динамические и онлайн тематические модели с изменяемым словарем требуют выполнения нескольких вычислительно сложных действий при поступлении новых документов, поэтому разработка вычислительно эффективных методов для анализа потока текстовых документов с использованием вероятностного тематического моделирования – актуальная научная задача. Изменение содержания понятия, а именно расширение словаря вероятностной тематической модели, необходимо для следования за тенденциями изменения потока текстовых документов. Постоянно появляются новые термины, тематическую принадлежность которых необходимо учитывать при построении и пополнении вероятностной тематической модели.

Целью работы является разработка математического и программного обеспечения вероятностного тематического моделирования потока текстовых документов, позволяющего повысить доступность применения ВТМ в практических задачах за счет использования открытого программного обеспечения. Опираясь на общие требования к системам анализа текстовых документов, могут быть сформулированы специфические требования к разрабатываемому программному комплексу для вероятностного тематического моделирования. К общим требованиям относят: модульность, масштабируемость, гибкость, технологическую разнородность, повторное использование кода, интероперабельность, открытость данных. Уточнения базовых требований для разрабатываемого программного комплекса:

- модульность – создание таких модулей, в которых каждый созданный модуль может быть использован как отдельная программа;
- масштабируемость – свойство, которое должно быть реализовано не только в программном комплексе в целом, но и в отдельных его модулях;
- повторное использование кода – поддержка свойства не только через реализацию объектно-ориентированного подхода, но и через копирование кода;

- техническая разнородность – возможность использования различных библиотек и языков программирования для построения различных ВТМ;
- интероперабельность по отношению ко всем модулям программного комплекса, а не только системы в целом;
- открытость данных – открытость используемых источников данных и открытость исполняемого программного кода, возможность вносить изменения в исполняемый код программного комплекса.

Основные требования к разрабатываемому комплексу программных средств для вероятностного тематического моделирования:

1. Программный комплекс должен реализовать возможность управляемого препроцессинга входных данных. В зависимости от стоящих перед пользователем задач, выполнять очистку данных от посторонней разметки, выполнять морфологический разбор, извлекать необходимую метаинформацию.
2. Программный комплекс должен поддерживать возможности построения ВТМ различными методами, включая метод обучения с учителем на основе информации о тематической принадлежности документов в обучающем множестве.
3. В комплексе должен быть реализован алгоритм многозначной классификации текстовых документов с использованием вероятностного тематического моделирования.
4. Комплекс программных средств должен в равной степени реализовать возможность проведения анализа как статичной коллекции, так и потока текстовых документов.

Существующие системы и программные комплексы в полной мере не удовлетворяют перечисленным требованиям и принципам, имеют ограниченную поддержку построения ВТМ.

2.2. Сценарии использования программного комплекса

С функциональной точки зрения систему интеллектуального анализа текстов представляют в виде последовательности из 4 шагов [63]:

1. предварительная подготовка (препроцессинг), выполняющая конвертацию текстовых данных в форму пригодную для проведения анализа;
2. алгоритмы интеллектуального анализа, специфичные для решаемой задачи;
3. визуализация или представление результатов проведенного анализа;
4. конкретизация информации, т.е. отсеивание избыточных данных не представляющих интереса для решаемой задачи.

На рисунке 14 представлена контекстная диаграмма для программного комплекса построения ВТМ.

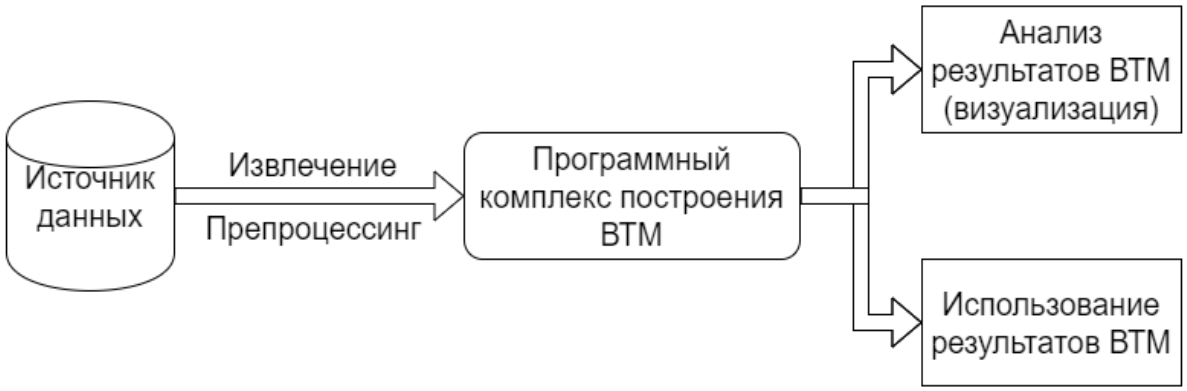


Рисунок 14 – Контекстная диаграмма для программного комплекса построения ВТМ

На входе осуществляется выбор источника данных. Прежде чем передать их модулю, осуществляющему вероятностное тематическое моделирование, необходимо выполнить их предварительную обработку (препроцессинг), подготовить под формат, воспринимаемый модулем ВТМ. Программный комплекс построения ВТМ, отвечает

за вероятностное тематическое моделирование. В зависимости от стоящей задачи и опираясь на имеющиеся данные, строится BTM. Результат работы модуля передается на выход программного комплекса либо для проведения анализа текстовых документов и визуализации BTM, либо для последующего использования в других программных комплексах и системах.

В рамках разрабатываемого программного комплекса должны быть реализованы следующие модули:

- выбор источника данных: должна быть реализована возможность настроить программный комплекс на получение данных в различных форматах, XML, CSV, …;
- препроцессинг: должна быть реализована предварительная обработка входных данных под формат, воспринимаемый программным комплексом;
- построение BTM: должна быть реализована поддержка нескольких методов построения BTM;
- анализ результатов BTM: должна быть реализована визуализация результатов вероятностного тематического моделирования;
- экспорт результатов BTM: должен быть реализован экспорт результатов вероятностного тематического моделирования в различных форматах, пригодных для последующего использования в сторонних системах и программных комплексах.

Опираясь на представленную схему, могут быть выделены основные группы заказчиков или потребителей результатов вероятностного тематического моделирования. В первую очередь это специалисты по анализу текстовых данных, лингвисты, веб-аналитики, информационные-архитекторы. Основная задача, решаемая этими специалистами с помощью BTM – это выявление схожих групп документов в наборе или потоке текстовых документов с последующей визуализацией результата. Как было отмечено ранее, темпоральная BTM позволяет проследить

изменение популярности тем во времени. Статичные коллекции текстовых документов позволяют выделить значимые для предметной области темы, например, для последующего построения информационной архитектуры. Вторая группа заказчиков – это специалисты по машинному обучению, использующие результаты вероятностного тематического моделирования в следующих системах: информационный и разведочный поиск, рекомендательные системы. Основная решаемая задача – это автоматическая кластеризация потока текстовых данных как документов, так и запросов пользователей в поисковую систему.

Информационная архитектура [37, с. 25]:

- структурное проектирование совместных сред информации;
- набор систем организации, предметизации, поиска и навигации в пределах веб-сайтов и интрасетей;
- искусство и наука формирования информационных продуктов и связанного с ними опыта взаимодействия для обеспечения требуемого уровня юзабилити и поисковой доступности;
- развивающаяся дисциплина и сообщество практиков, ставящие своей задачей распространение принципов проектирования и архитектуры на цифровых просторах.

В ходе изучения сценариев работы специалистов по анализу данных были выделены базовые сценарии, которые легли во основу прецедентов использования программного комплекса. Прецеденты использования программного комплекса для вероятностного тематического моделирования:

- подготовить корпус текстов (описание представлено в таблице 2);
- построить BTM (описание в таблице 3);
- многозначная классификация с использованием BTM (таблица 4);
- анализ коллекции и потока текстовых документов (таблица 5).

Таблица 2 – Описание прецедента подготовки корпуса текстов

Краткое описание	Этот прецедент позволяет действующему лицу, пользователю программного комплекса, извлекать данные из подходящего источника, выполнять препроцессинг и сохранение в корпусе текстов для построения BTM
Действующие лица	Пользователь программного комплекса, заказчик корпуса текстов.
Предусловие	Заказчик формулирует требования к корпусу текстов; формат данных в корпусе текстов заранее определен.
Основной поток	1. Прецедент использования начинается с выбора подходящего источника. 2. Пользователь программного комплекса оценивает соответствие выбранного источника требованиям заказчика. 3. Из подходящего источника извлекаются данные, необходимые для построения BTM. 4. Выполняется предварительная подготовка данных. В зависимости от поставленной задачи и набора данных в предварительную обработку могут входить следующие шаги: • Извлечение текстовых данных: ○ очистка текстов документов от элементов HTML или другой специализированной разметки; ○ приведение слов к нормальным словоформам; ○ морфологическая обработка.

Продолжение таблицы 2

	• Извлечение метаинформации: ○ извлечение информации о дате создания документа или времени описанных в документе событий; ○ извлечение информации об авторстве документа; ○ в зависимости от стоящих задач и входных данных на этом шаге может быть извлечена необходимая дополнительная метаинформация. 5. Обработанные данные сохраняются в корпус текстов для вероятностного тематического моделирования.
Альтернативные потоки	Заказчик может поставить задачу обновления существующего корпуса. В этом случае пользователь программного комплекса уже знает источник данных, повторно выполняет шаги извлечение, предварительная подготовка, сохранение корпуса.
Постусловия	Если прецедент использования завершился успешно, то корпус текстов сохранен и готов для использования в построении ВТМ. В противном случае корпус не создан.

Таблица 3 – Описание прецедента построения ВТМ

Краткое описание	Этот прецедент использования позволяет действующему лицу строить ВТМ на основе данных корпуса.
Действующие лица	Пользователь программного комплекса, заказчик результатов тематического моделирования.

Продолжение таблицы 3

Предусловие	Заказчик формулирует требования к построению BTM, указывает источник данных. Методы BTM заранее определены.
Основной поток	1. Прецедент использования начинается с выбора библиотек, реализующих метод построения BTM. 2. Пользователь программного комплекса использует выбранный корпус текстов. 3. Пользователь программного комплекса задает параметры вероятностного тематического моделирования, опираясь на предварительную оценку размера корпуса и требования заказчика к BTM. 4. Выполняется построение BTM, результаты которой сохраняются для последующего использования. 5. В зависимости от требований заказчика выполняется экспорт результата вероятностного тематического моделирования для последующего использования в других системах, либо производится анализ корпуса текстов с визуализацией.
Альтернативные потоки	Требования к BTM могут формулироваться недостаточно точно, в этом случае пользователь программного комплекса должен взять на себя роль аналитика данных и выбрать подходящий метод и параметры построения BTM для получения наиболее наглядного результата.
Постусловия	Если прецедент завершился успешно, то BTM построена, данные проанализированы или переданы другим сервисам. В противном случае BTM не построена.

Таблица 4 – Описание прецедента многозначной классификации

Краткое описание	Этот прецедент использования позволяет построить ВТМ методом обучения на размеченных данных, а затем выполнить многозначную классификацию текстовых документов.
Действующие лица	Пользователь программного комплекса, заказчик многозначной классификации.
Предусловие	Заказчик предоставляет данные для построения ВТМ (обучающее множество) и документы, которые необходимо классифицировать.
Основной поток	1. Прецедент использования начинается с построения ВТМ на основе метаинформации в размеченных данных, поданных на вход программному комплексу. 2. Определить тему нового документа. 3. Передать результат внешнему сервису или другому модулю программного комплекса.
Альтернативные потоки	Этот прецедент используется как составная часть прецедента анализа потока текстовых документов.
Постусловия	Если прецедент выполнился успешно, то документы классифицированы. В противном случае классификация не выполнена.

Распределение требований программного комплекса и прецедентов использования представлено в таблице 6. Диаграмма задач использования представлена на рисунке 15.

Таблица 5 – Описание прецедента анализа потока текстовых документов

Краткое описание	Этот прецедент использования позволяет анализировать поток текстовых документов с помощью темпоральных и динамических BTM.
Действующие лица	Пользователь программного комплекса, заказчик анализа.
Предусловие	Заказчик предоставляет для анализа статичную коллекцию документов, в которых каждый документ имеет информацию о дате описанных в документе событиях либо предоставляет поток текстовых документов с указанием сколько данных необходимо копить до первого построения BTM.
Основной поток	1. В случае динамической тематической модели прецедент начинается с накопления текстовых документов для построения стартовой BTM. 2. Следующим шагом выполняется препроцессинг накопленных данных. 3. Пользователь программного комплекса выполняет построение стартовой BTM. 4. Все следом поступающие документы проходят прецедент многозначной классификации. 5. Если в поступающих документах встречаются новые для BTM слова, то выполняется определение темы «нового слова». 6. Далее поступающие документы и новые слова пополняют BTM.

Продолжение таблицы 5

	7. Обработка потока текстовых документов может быть остановлена пользователем программного комплекса, либо поток прекратится самостоятельно. 8. В зависимости от требований заказчика, результат построения ВТМ может быть визуализирован либо передан внешним системам.
Альтернативные потоки	Для построения темпоральной ВТМ может быть взята статичная коллекция документов, в которой имеется информация о дате описанных в документе событиях либо есть дата создания документа. Поток данных может быть эмулирован из статичной коллекции документов на основе информации о дате создания документа или дате описанных событий.
Постусловия	Если прецедент выполняется успешно, то результат ВТМ визуализируется либо передается внешним сервисам. В противном случае анализ не выполнен.

Требования к данным. Вероятностное тематическое моделирование основано на модели «мешка слов», т.е. порядок слов в документе не имеет значения для построения классических ВТМ – PLSA, LDA, ARTM, достаточно коллекции документов в формате «мешка слов». Как было отмечено в первой главе, гипотеза «мешка слов» далека от реальности, существуют методы построения ВТМ, в которых учитывается последовательность слов. Для построения таких моделей необходим оригинальный текст документов. Наличие оригинального текста в коллекции позволит пользователю программного комплекса самостоятельно принимать решение, на основе какой текстовой информации строить модель. ВТМ могут быть построены на именах существительных, прилагательных, именных группах,

входящих в состав текстовых документов, и других выборках из исходных данных. При этом в модуль предварительной подготовки должна быть включена сегментация и токенизация текста, морфологическая обработка, метод разделения текста на предложения.

Таблица 6 – Распределение требований и прецедентов использования

Требование	Прецеденты использования
1. Программный комплекс должен реализовать возможность управляемого препроцессинга входных данных. В зависимости от стоящей перед пользователем задачи выполнять очистку данных от посторонней разметки, выполнять морфологический разбор, извлекать необходимую метаинформацию.	Подготовка корпуса текстов.
2. Программный комплекс должен поддерживать возможности построения ВТМ различными методами, включая метод обучения с учителем на основе информации о тематической принадлежности документов в обучающем множестве.	Построение ВТМ.
3. В комплексе должен быть реализован алгоритм многозначной классификации текстовых документов с использованием вероятностного тематического моделирования.	Многозначная классификация.
4. Комплекс программных средств должен в равной степени реализовать возможность проведения анализа как статичной коллекции, так и потока текстовых документов.	Анализ потока текстовых документов.

Главной задачей разрабатываемого программного комплекса является анализ потока текстовых документов. Для эмуляции потока текстовых документов необходима информация о дате создания документов или о дате описанных в документах событиях. Одной из задач программного комплекса является разработка алгоритма многозначной классификации. Для реализации этой задачи необходима информация о принадлежности документов коллекции к темам. Это могут быть теги либо категории. Для построения BTM с помощью обучения необходим корпус текстов, в котором для каждого документа определен набор тем. Для построения других специфичных BTM может потребоваться дополнительная информация о документах, такая как авторство этих документов для построения автор-тематических моделей.

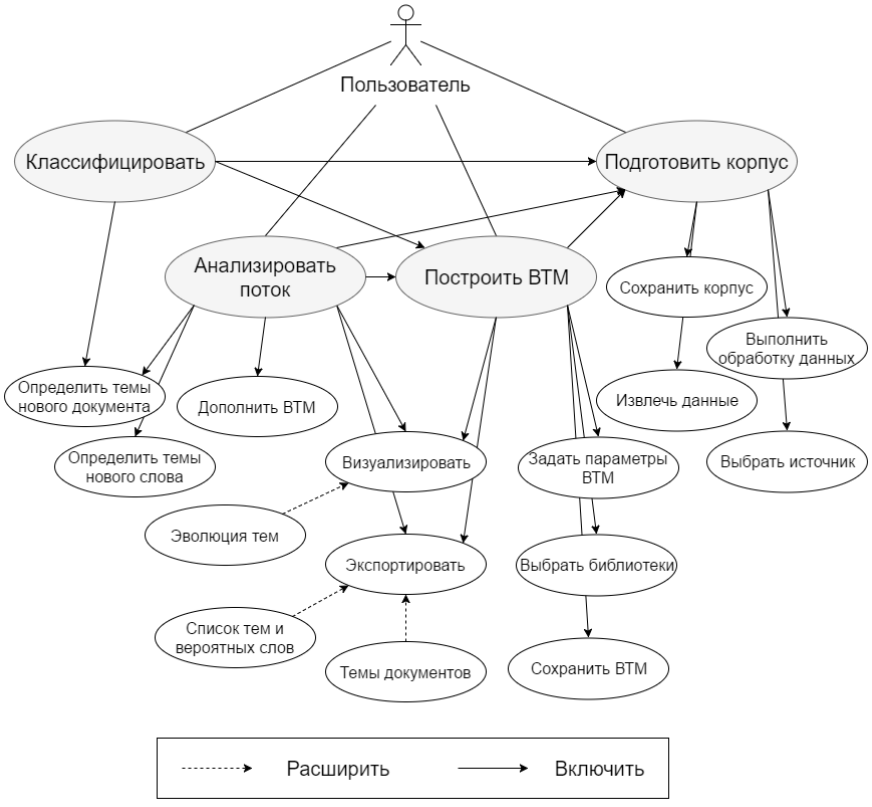


Рисунок 15 – Диаграмма задач использования программного комплекса

2.3. Концептуальная схема программного комплекса

На рисунке 16 представлена концептуальная схема программного комплекса. Стрелками обозначен поток данных внутри программного комплекса. В ходе выполнения намеченной задачи последовательно задействуются необходимые модули. Для построения классической BTM последовательно выполняются шаги 1-создание корпуса, 2-сохранение данных в корпус, 3-построение BTM, 4 – сохранение результатов вероятностного тематического моделирования, 5 – визуализация или экспорт BTM. Для многозначной классификации выполняются шаги 1, 2, 3, 4, 6 – поступление нового документа, 7 – определение тем нового документа с помощью BTM, 8 – вывод результата классификации нового документа. Для анализа потока текстовых документов выполняются шаги 1, 2, 3, 4, 6, 7, 9 – определение тем «нового слова», 11 – передача данных в модуль пополнения BTM, 12 – пополнение BTM, 5.

Принципы, лежащие в основе предлагаемого подхода к разработке программного комплекса для вероятностного тематического моделирования:

- человеко-ориентированный подход – автоматизированный режим работы, предоставление пользователю полного контроля над программным комплексом, выбора необходимых алгоритмов, возможности изменить программный код;
- событийно-ориентированный подход (Event-driven) – предоставление пользователю интерфейса, в котором реализована возможность определять очередность событий при использовании программного комплекса, возможность составлять и запускать свой сценарий выполнения программы;
- подход, определяемый данными (Data-Driven) – входные данные определяют возможности выполнения программы, выполнение сценариев работы модулей в зависимости от входных данных.

Рисунок 16 – Концептуальная схема программного комплекса

Введем **системные ограничения** разрабатываемого комплекса программных средств в соответствии с планом подготовки технического задания, описанного в [21]. Для обеспечения удобства использования программного комплекса следует реализовать хороший уровень представления для доступа пользователя к параметрам программного комплекса, промежуточным данным и результатам. Уровень представления должен включать:

- Средства изменения параметров работы программного комплекса и отдельных его модулей. Эти средства не требуют наличия дружелюбного графического интерфейса, но должны обеспечить максимальные возможности по настройке параметров.

- Низкоуровневый доступ к данным и исполняемому программному коду с возможностью их модификации.
- Административные средства для создания, модификации и управления иерархиями понятий и сущностями предметной области.
- Возможности визуализации промежуточных данных и результатов построения BTM, которые должны обеспечивать:
    - краткость – способность одновременного отображения большого количества разнотипных данных;
    - относительность и близость в отображении тем и слов BTM;
    - масштабируемость – возможность перемещаться от микро к макропредставлениями;
    - темпоральность – отображение потоковых данных во времени.

Опираясь на практическую потребность работать с данными различного объема, комплекс программных средств должен корректно работать на персональных компьютерах. В случае, если объем обрабатываемых данных позволяет это сделать, и также корректно работать на специализированных мощных серверах для обработки больших объемов данных. Программный комплекс будет распространяться по свободной лицензии для некоммерческого использования, поэтому программные библиотеки, используемые в модулях программного комплекса, а также коллекции документов также должны распространятся по свободным лицензиям.

Требования к корпусу для построения BTM

Создавая тематическую модель, необходимо учитывать языковые особенности текстов. Для тестирования разрабатываемого программного комплекса, а также для развития методов тематического моделирования, работающих с русским языком, необходим русскоязычный корпус текстов, распространяемый по свободной лицензии.

Основные требования к корпусу текстов для построения BTM:

- распространение корпуса по свободной лицензии;
- количество документов в корпусе должно быть достаточным для проведения исследований;
- корпус должен содержать:
    - оригинальный текст документов на естественном языке;
    - даты описанных событий или даты создания документов;
    - дополнительную метаинформацию для построения специализированных BTM;
    - темы или тематические категории, к которым относятся документы.

Предлагаемый технологический процесс создания корпуса состоит из следующих шагов: определение источника, предварительная обработка текстов документов, извлечение информации о датах описанных событий или датах создания документов, извлечение дополнительной метаинформации, извлечение тем или тематических категорий, к которым относится документ, токенизация и лемматизация текста, морфологическая обработка текста, создание корпуса с сохранением метатекстовой разметки параметров документов, обеспечение доступа к корпусу (рисунок 17).

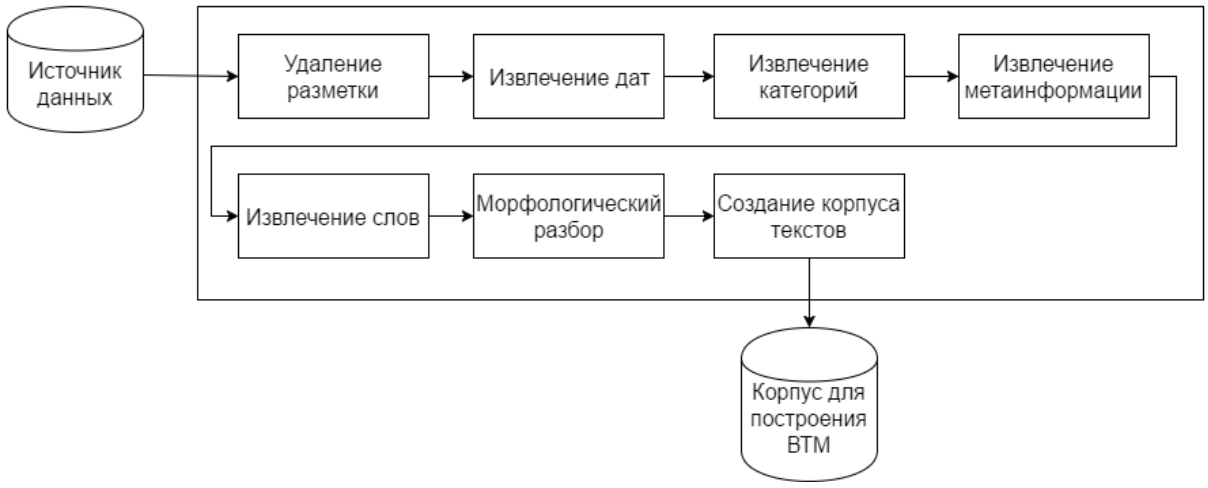


Рисунок 17 – Технологический процесс создания корпуса текстов для построения BTM

Выводы к главе 2

1. Рассмотрены основные задачи интеллектуального анализа текстовых документов, алгоритмы анализа потоков данных и современные задачи интеллектуального анализа потоков текстовых документов, что позволило подтвердить перспективность использования вероятностного тематического моделирования для интеллектуального анализа коллекций и потоков текстовых документов. Выявлена потребность в наличии открытых человеко-ориентированных программных комплексов, в которых специалист, решающий свою прикладную задачу, может самостоятельно выбрать подходящий алгоритм или программную библиотеку для построения ВТМ. Определено требование к анализу потока текстовых документов, состоящее в том, что скорость анализа должна быть выше скорости поступления новых документов.
2. Опираясь на общие требования к системам анализа текстовых документов, сформулированы специфические и основные требования разрабатываемому программному комплексу для вероятностного тематического моделирования. Существующие системы в полной мере не удовлетворяют перечисленным требованиям, поэтому разработка комплекса программных средств для построения ВТМ является важной практической задачей.
3. Рассмотрены сценарии использования программного комплекса. Создана контекстная диаграмма для программного комплекса построения ВТМ. Выделены основные действующие лица в работе программного комплекса. В ходе изучения сценариев работы специалистов по анализу данных были выделены базовые сценарии, которые легли в основу описания прецедентов использования программного комплекса.

4. Создана концептуальная схема программного комплекса. Сформулированы принципы, лежащие в основе предлагаемого подхода к разработке программного комплекса: человеко-ориентированный подход, событийно-ориентированный подход, подход, определяемый данными. Определены системные ограничения разрабатываемого комплекса программных средств.
5. Сформулированы основные требования к специальному корпусу текстовых документов, предназначенному для построения ВТМ. Разработан технологический процесс создания корпуса.

3. Вероятностное тематическое моделирования потока текстовых документов

В третьей главе проведен обзор существующих методов обучения ВТМ и алгоритмов многозначной классификации. Предложен метод обучения ВТМ, заключающийся в расчете матрицы «слово-тема» по известным, размеченным данным «слово-документ», «документ-тема». Предложен алгоритм многозначной классификации текстовых документов с использованием вероятностного тематического моделирования ml-PLSI. Выполнен обзор существующих подходов к расширению словаря ВТМ. Реализован метод определения тематики «нового слова», в котором тематический вектор «нового слова» рассчитывается через произведение Адамара тематических векторов документов, где это слово встретилось. Разработан алгоритм, позволяющий расширять словарь ВТМ.

3.1. Обзор алгоритмов многозначной классификации

Одним из направлений обработки текстов является категоризация. Категоризация выполняется различными алгоритмами классификации и кластеризации. Задачи классификации волновали исследователей с середины XIX века. В работе [13] дан обзор ряда алгоритмов классификации 1969-1980 гг. Под задачей кластеризации документов понимают задачу разбиения заданной выборки текстовых документов на непересекающиеся подмножества – кластеры так, чтобы каждый кластер состоял из похожих документов, а документы разных кластеров существенно отличались. Под задачей классификации документов понимают задачу отнесения документа к одной из нескольких категорий (классов, тем) на основании содержания документа. Классификация относится к задачам обучения с учителем, когда алгоритм классификации сначала обучается на размеченных документах, а затем классифицирует новые документы.

Большое количество исследований алгоритмов кластеризации и классификации связано с определением одной категории, к которой может принадлежать документ. В реальном мире чаще бывает, что один и тот же документ может быть отнесен к нескольким категориям. Например, статья или новость про футбольный матч может быть отнесена к категориям: спорт, футбол, спортивные соревнования, городские мероприятия. Поэтому особенно актуальны методы и алгоритмы многозначной (нечеткой) классификации (multi-label classification) и мягкой кластеризации.

Multi-label classification не имеет устоявшегося русскоязычного термина. В литературе встречаются многозначная классификация и нечеткая классификация. В этой работе мы используем термин многозначная классификация. В машинном обучении многозначная классификация представляет собой вариант задач классификации, в которой к каждому документу (классифицируемому объекту) должны быть определены несколько меток. Не следует путать многозначную классификацию с многоклассовой классификацией, цель которой определить один класс из более чем двух классов кандидатов. В работе [99] представлен обзор алгоритмов многозначной классификации. Существует два основных метода для решения задач многозначной классификации: метод преобразования проблемы и метод адаптации. Метод преобразования проблемы трансформирует задачу в набор двоичных классификационных задач. Методы адаптации выполняют классификацию множества меток классов, решают задачу в ее полном виде.

Для решения задачи многозначной классификации используют адаптированные версии алгоритмов классификации, такие как Boosting (AdaBoost), k-ближайших соседей, деревья решений, ядерные методы, SVM, нейронные сети. Метрики оценки качества многозначной классификации отличаются от обычной классификации в силу особенности задачи.

В работе [90] описан алгоритм тематической модели классификации под названием Label-LDA. В основе работы алгоритма лежит базовый алгоритм LDA,

векторы документов и тем порождаются распределением Дирихле. Основан он на двух сильных ограничениях, темы отождествляются с классами, предполагается, что для каждого документа точно известно множество всех классов, к которым он относится. Аналогичные ограничения касаются алгоритма Flat LDA описан в работе [92]. Для задачи классификации несбалансированных классов в этой работе был предложен алгоритм Prior-LDA, использующий частотную регуляризацию. По утверждению авторов работы [92] Flat-LDA, Prior-LDA являются частными случаями более общего алгоритма Dependency LDA, в котором предлагается моделировать классы документов через распределение тем документов и вводится новая неизвестная матрица класс-тема. В работе [88] предложен подход к многозначной классификации методом LDA с использованием знаний толпы под названием ML-PA-LDA-C. Используется информация не только о присутствующем классе, но и об отсутствующем, применяется для построения модели по зашумленным размеченным данным. В работе устранено одно ограничение Label-LDA, предполагается, что точно неизвестно множество всех классов, к которым принадлежит документ.

3.2. Метод построения BTM на основе обучения с учителем

Модели, разработанные на основе латентного размещения Дирихле (LDA), как указано в работе [4], не имеют сильных лингвистических обоснований. При этом классическая модель вероятностного латентно-семантического анализа PLSA [71] не связана с какими-либо параметрическими априорными распределениями.

Пусть D – множество текстовых документов, W – словарь терминов. Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, \dots, w_n)$ из словаря W.

С учетом гипотезы условной независимости $p(w|d,z) = p(w|z)$ по формуле полной вероятности получаем вероятностную модель порождения документа d:

$$p(w|d) = \sum_{z \in Z} p(w|d,z)p(z|d) = \sum_{z \in Z} p(w|z)p(z|d) = \sum_{z \in Z} \varphi_{wz}\theta_{zd}. \quad (20)$$

Для вычисления φ_{wz} и θ_{zd} используется EM-алгоритм.

Согласно вероятностному тематическому моделированию, ВТМ появления пары «документ-слово» может быть записана тремя эквивалентными способами:

$$p(d,w) = \sum_{z \in Z} p(z)p(w|z)p(d|z) = \sum_{z \in Z} p(d)p(w|z)p(z|d) = \sum_{z \in Z} p(w)p(z|w)p(d|z), \quad (21)$$

где:

- Z – множество тем;
- $p(z)$ – неизвестное распределение тем в коллекции;
- $p(d)$ – распределение на множестве документов, эмпирическая оценка $p(d) = \frac{n_d}{n}$, где $n = \sum_d n_d$ – суммарная длина всех документов, а n_d – длина документов в словах;
- $p(w)$– распределение на множестве слов, эмпирическая оценка $p(w) = \frac{n_w}{n}$, где n_w – число вхождений слова w во все документы (рисунок 18).

Если отождествить понятие темы тематической модели и категории документа, учесть, что задача построения ВТМ имеет бесконечно много решений [3], то можно построить один из вариантов тематической модели, обучившись на размеченном корпусе. Построенная на данных предположениях тематическая модель зависит от качества выбранной для обучения коллекции. Например, категории в корпусе SCTM-ru проставлены авторами новостей. Перед авторами не стояла задача указать все категории, которые только возможны для каждой новости, поэтому часть документов не получила полный набор категорий, даже если они этого заслуживали. При этом объем корпуса позволяет предположить, что в своем большинстве авторы использовали наиболее характерные категории. Основываясь на знании толпы, мы можем рассчитывать вероятностную оценку отнесения слова к категории:

$$p(w|d) = \sum_{z \in Z} p(w|z)p(z|d) = \sum_{z \in Z} p(w|c)p(c|d) = \sum_{z \in Z} \varphi_{wz}\theta_{zd}\,. \qquad (22)$$

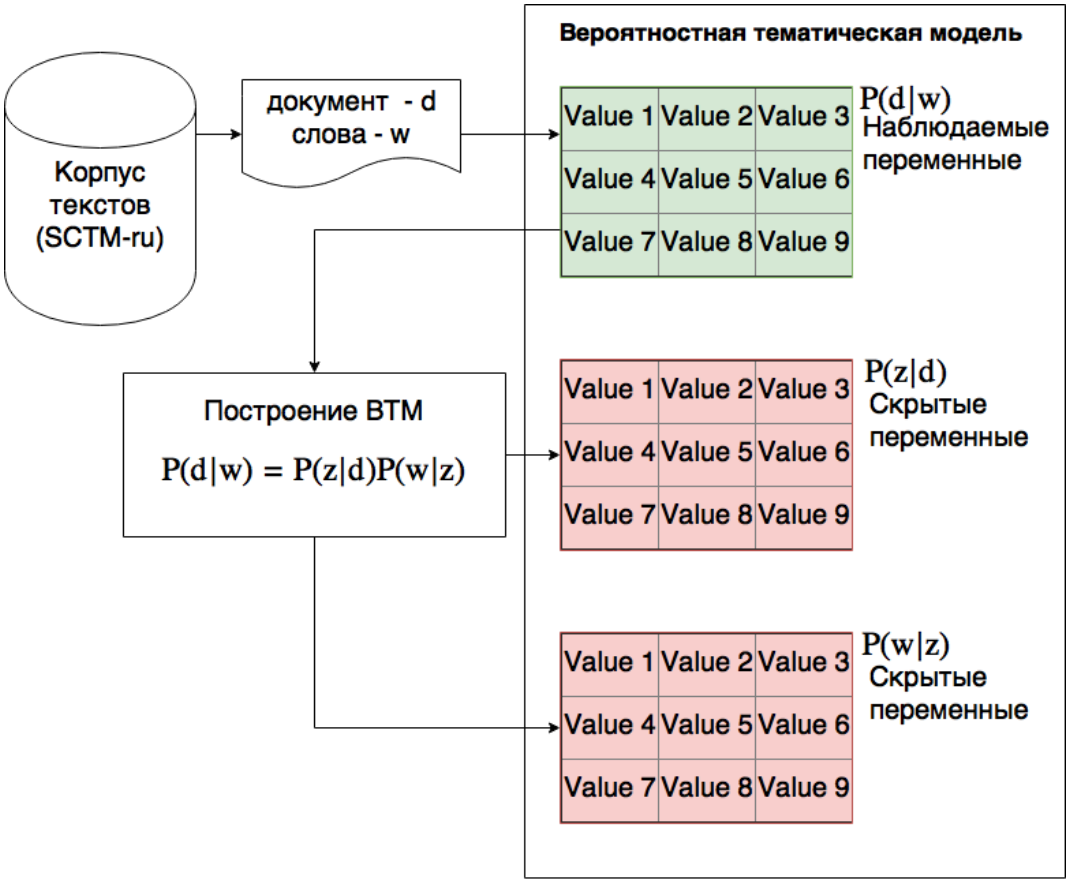


Рисунок 18 – Схема построения вероятностной тематической модели по корпусу SCTM-ru

Построенная таким образом тематическая модель может быть одним из множества решений задачи тематического моделирования. В работе [48] рассмотрен алгоритм создания тематической модели методом обучения с учителем (Листинг 1).

Для уменьшения размерности векторного пространства рекомендуется все слова в корпусе привести к нормальной словоформе. Если документов в корпусе немного и алгоритм будет выполнен за конечное время, то этого делать необязательно, т.к. словоформа также может стать важной информацией для определения категории документа. Слова в конкретной словоформе чаще могут встречаться в документах, принадлежащих одной категории. На основании семиологии [18] слово в знаковой

системе наделено смыслом, является частью языка. Поэтому в данной работе мы не приводим слова в корпусе к нормальной словоформе.

Листинг 1: Алгоритм построения вероятностной тематической модели методом обучения с учителем

Вход: коллекция документов D с указанием тематических категорий C .

Выход: распределения $P(w|c), P(d|c), P(w|d)$.

1. Для всех $d \in D, w \in d$, рассчитать $p(w|d) = \frac{n_{dw}}{n}$,
2. Для всех $c \in C, d \in D$, рассчитать $p(d|c) = \frac{n_c}{n}$,
3. Для всех $w \in W, c \in C$, рассчитать $p(w|c) = \varphi_{wz} = \frac{n_{dwc}}{n_{dw}}$.

Обучить тематическую модель по размеченным данным – это значит рассчитать матрицы «слово-документ», «документ-категория» и «слово-категория» для каждого слова из коллекции документов (схема обучения BTM представлена на рисунке 19). На первом шаге рассчитываем матрицу «слово-документ». Значения матрицы – это количество повторений слова в документе. На втором шаге рассчитываем матрицу «документ-категория». Для этого для каждого документа в корпусе получаем список категорий, к каждой категории документ может быть отнесен не более одного раза, поэтому значения матрицы «документ-категория» – это единицы в том случае, если документ связан с категорией и нуль, если такой связи нет. На третьем завершающем шаге рассчитываем матрицу «слово-категория». Значения матрицы – это вероятность встретить слово в этой категории. Как ранее мы отмечали, разметка документов категориями скорее всего содержит ошибки. Эти ошибки можно разделить на два вида: слово, которое имеет отношение к определенной категории и слово, которое редко встречается и недополучило связь с категорией. Документ мог быть отмечен какой-либо категорией по ошибке. В данной работе для уменьшения влияния ошибок

на результат предлагаем использовать регуляризацию. Возможны и другие варианты, например, учитывать когерентность категории для совместно встречаемых слов.

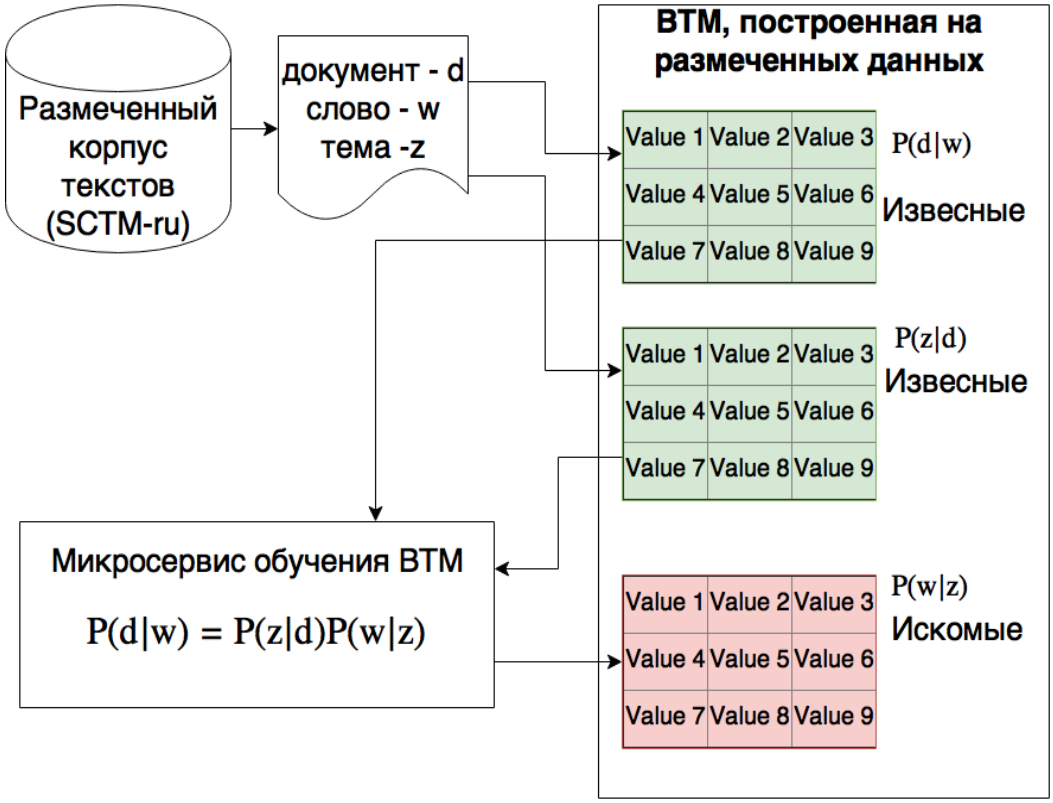


Рисунок 19 – Схема обучения BTM по размеченным данным корпуса SCTM-ru

В результате обучения тематической модели получено три матрицы, по которым можно узнать, с какой вероятностью то или иное слово относится к теме, из каких слов формируется тема и к каким темам относится каждый документ. По матрице «слово-документ» можно восстановить каждый документ в формате «мешка слов». По матрице «слово-категория» можно оценить, с какой вероятностью то или иное слово относится к категории, можно рассчитать вероятность отнесения нескольких слов к категориям, для этого достаточно просуммировать вероятность отнесения каждого слова к категории $\sum_{w \in d} p(w|c)$. По матрице «документ-категория» можно получить список всех документов, которые связаны с категорией.

3.3. Алгоритм многозначной классификации ml-PLSI

Для реализации алгоритма многозначной классификации текстовых документов необходима ВТМ, обученная на размеченных данных. Для предсказания категорий документа выполняется Листинг 2.

Листинг 2: Многозначная классификация на базе вероятностного тематического моделирования

Вход: Вероятностная тематическая модель, новый документ d_{new} .

Выход: Список предсказанных категорий.

1. Для всех $w \in d_{new}$, $\sum_{w \in d} p(w|c)$,
2. Вернуть список категорий по убыванию суммы вероятностных оценок.

Следует учитывать, что слова, которые есть в новом документе и отсутствуют в корпусе, на котором обучалась тематическая модель, не будут учтены при предсказании категорий. На рисунке 20 представлена схема многозначной классификации документа с использованием вероятностного тематического моделирования.

Для того чтобы отобрать только наиболее релевантные категории, используем регуляризацию, а именно:

- Регуляризация по документам. Если слово встречается в большом количестве документов, то оно перестает быть информативным для предсказания категорий. Вероятность отнесения документа к теме по этому слову может быть обнулена.
- Регуляризация по темам. Если слово встречается в большом количестве категорий, то оно перестает быть информативным для предсказания категорий. Вероятность отнесения слова к категории по этому слову может быть обнулена.

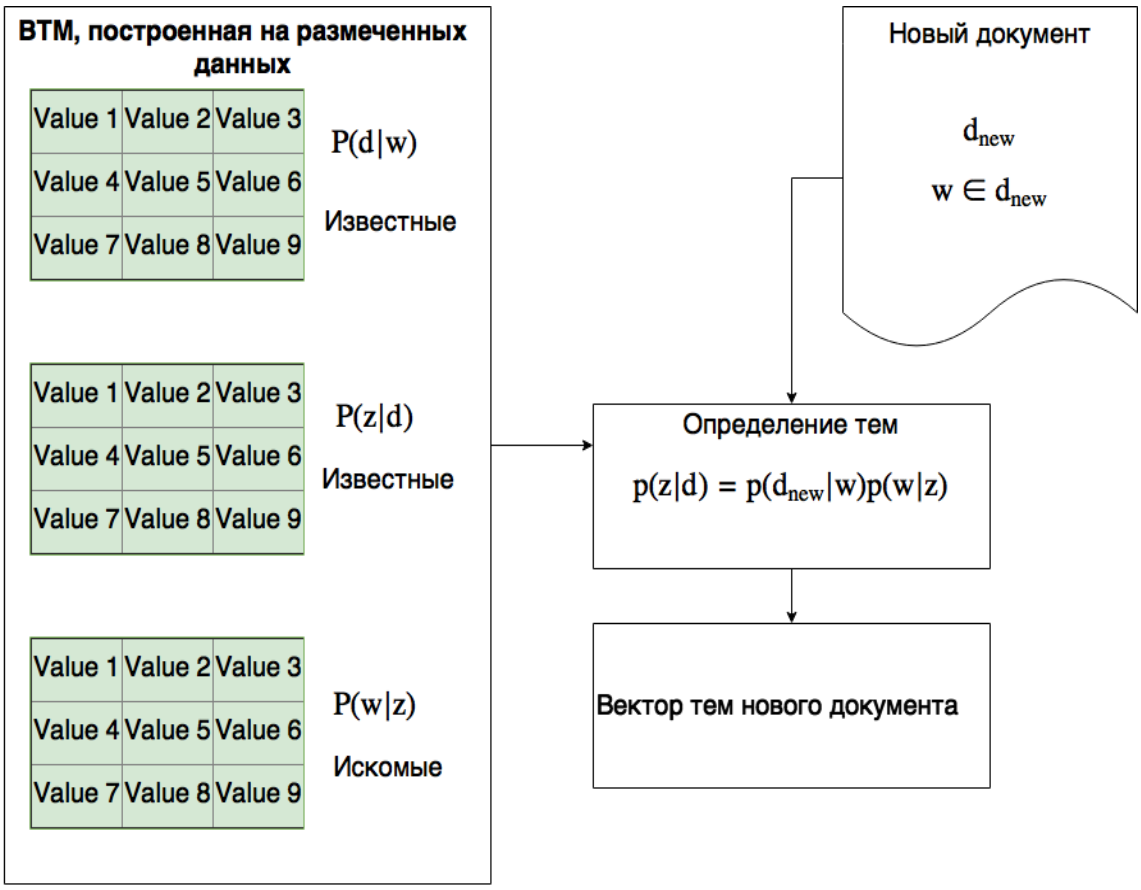


Рисунок 20 – Схема многозначной классификации текстовых документов с использованием BTM

Экспериментальная оценка качества кластеризации. В качестве данных для исследования используем корпус SCTM-ru [17], созданный специально для тестирования задач тематического моделирования. Источником данных корпуса является международный новостной сайт «Русские Викиновости». Корпус SCTM-ru состоит из 7000 документов, 185 авторов, почти 12000 уникальных категорий. События, описанные в документах, распределены по более чем 2000 уникальным датам, с ноября 2005 года по июнь 2014 года. В корпусе SCTM-ru 2400000 словоупотреблений, состоящих только из русских букв. Словарный состав корпуса – 150600 уникальных словоформ, 59000 уникальных лемм.

Каждая новость содержит указанные автором категории. У автора новости не стояла задача перечислить все категории, к которым новость может иметь отношение,

тем не менее указанные категории дают весомые основания полагать, что новость сильно связана с темой этих категорий. Обычно новость определена автором к нескольким категориям, поэтому мы рассматривали алгоритмы многозначной классификации. Для каждого документа мы не знаем точного множества всех категорий.

В корпусе SCTM-ru часть категорий авторы используют редко, 19284 категориями размечено менее 50 новостей из корпуса. Так как эти категории не представляют большой ценности для нашей тематической модели и могут создавать дополнительные трудности при прогнозировании категорий новых документов, мы не будем учитывать их при построении тематической модели. Категории, которыми размечены более 50 новостей, 230 штук, далее их называем рабочие категории или просто категории, их мы используем для обучения тематической модели. Всего документов, размеченных рабочими категориями, 6428. Для обучения модели используем 5000 новостей. Чтобы отсеять неинформативные слова, междометия и предлоги, создаем словарь стоп слов, в который вошло 151 слово.

Для предсказания тематических категорий используются не затронутые в обучении новости. Корпус текстов делится на тестовое и обучающее множество, (схема разделения представлена на рисунке 21).

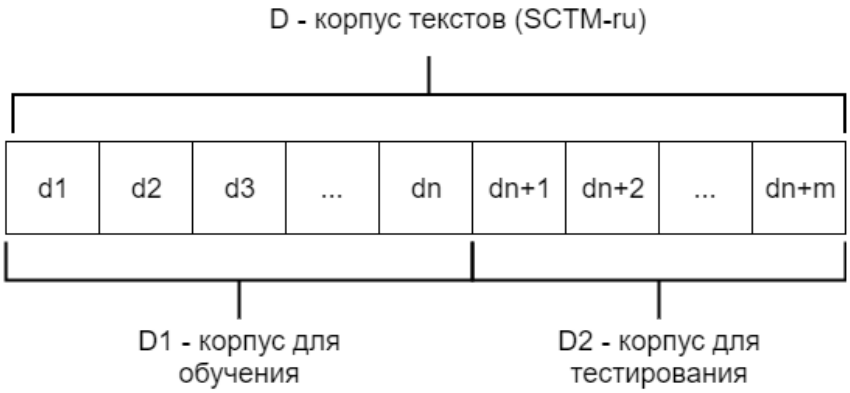


Рисунок 21 – Разделение корпуса текстов на обучающее и тестовое множество

Используем следующие метрики оценки качества многозначной классификации:

1. Функция потерь Хэмминга для ошибочных предсказаний.
2. Количество документов, в которых первая предсказанная категория совпала с любой из указанных автором новости.
3. Процент верных предсказаний из первых 50 предсказанных категорий.

Результаты оценки качества приведены в таблице 8.

Таблица 8 – Оценка качества многозначной классификации

Метрика качества	Результат (100 новостей)	Результат (500 новостей)
Функция потерь Хэмминга	0,18	0,17
Первая предсказанная (точность)	82%	84%
Процент верных предсказаний из 50 первых категорий (полнота)	71%	72%

Построенная тематическая модель позволяет определить наиболее вероятные категории не только для текстового документа, но и для отдельной фразы или слова. Например, для фразы «выборы президента России» десять первых категорий отображены на рисунке 22, для слова «футбол» на рисунке 23 десять наиболее вероятных категорий.

Проведенные эксперименты на корпусе SCTM-ru демонстрируют перспективность использования тематического моделирования в задачах многозначной классификации. Вероятностная тематическая модель может быть использована в задачах ассоциативной классификации [9, 10] в комбинации с другими алгоритмами классификации, а также может быть решателем в алгоритмах коллективного распознавания, описанных в работе [8].

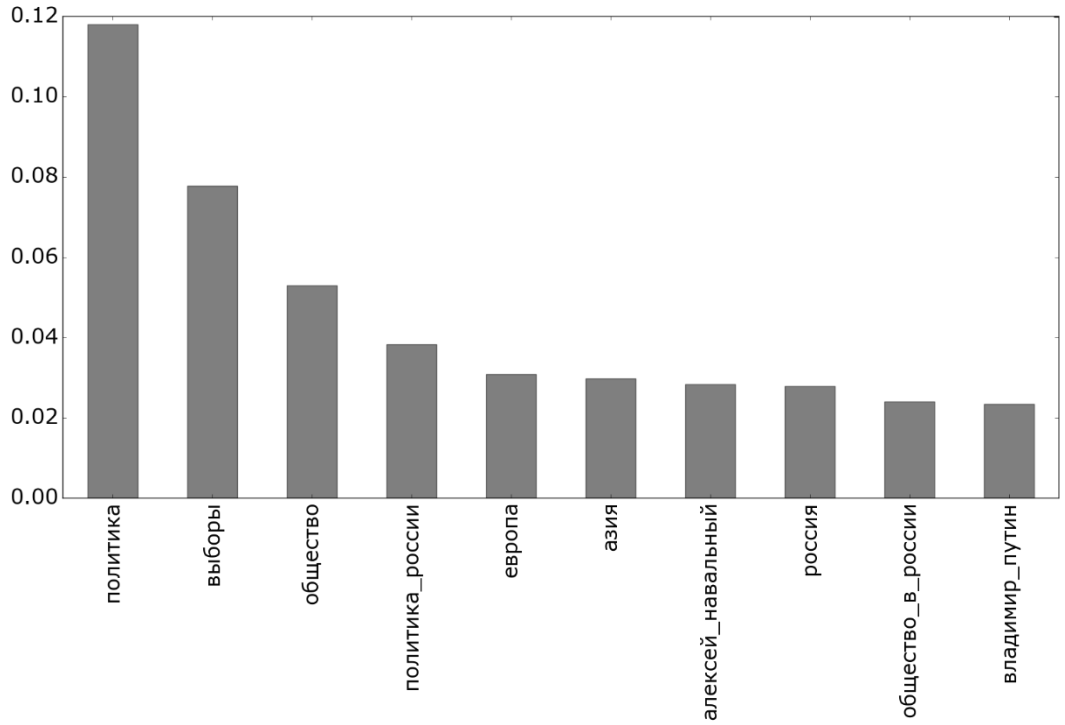


Рисунок 22 – Наиболее вероятные категории для фразы «выборы президента России»

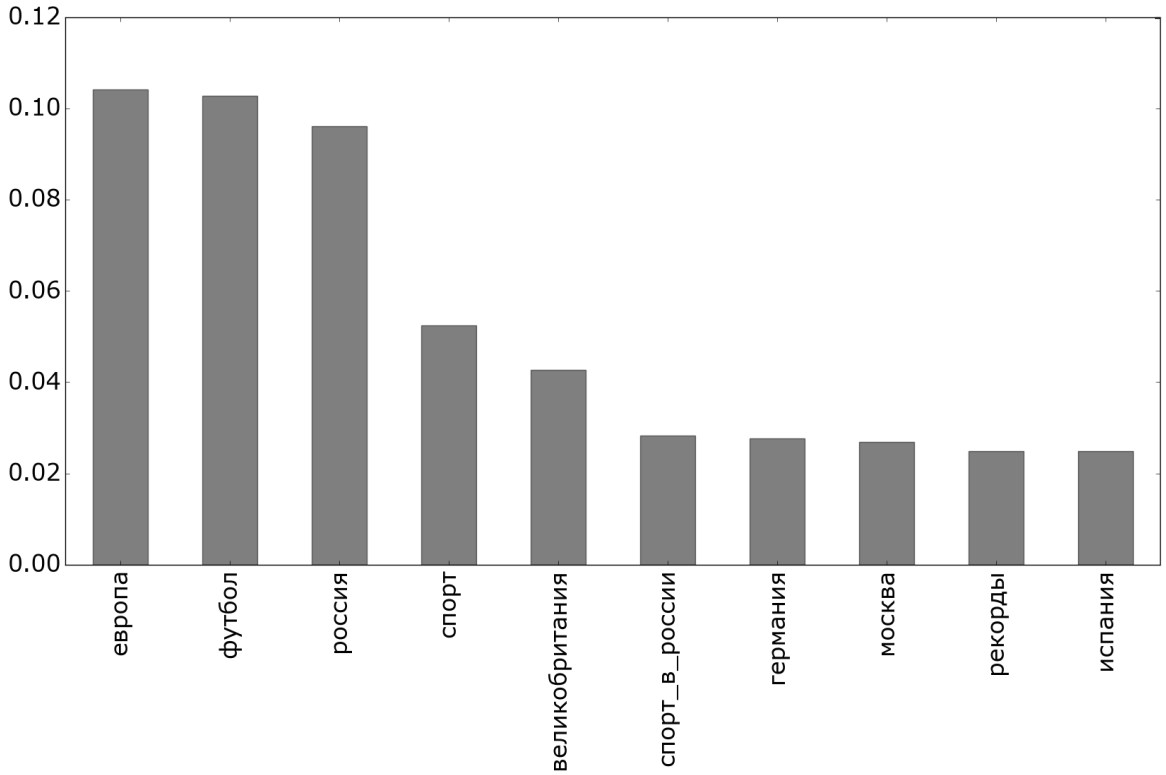


Рисунок 23 – Наиболее вероятные категории для слова «футбол»

3.4. Метод определения тем для «нового слова»

Важной задачей для динамических ВТМ является определение темы новых слов. Когда в потоке текстовых документов встречается слово, неизвестное для тематической модели, то необходимо принимать решение как учитывать вклад этого слова в тему нового документа. Существует метод добавления новых документов в уже построенную ВТМ, в котором новые слова пропускаются и никак не влияют на определение тем документа, где встретились.

Обзор существующих методов расширения словаря ВТМ

В работах «Динамическая тематическая модель» (Dynamic Topic Model) [49], «Многомасштабная томографическая модель» (Multiscale Topic Tomography) [82], «динамическая тематическая модель с непрерывным временем» (Continuous Time Dynamic Topic Models) [105] и «онлайн LDA» (Online Learning for Latent Dirichlet Allocation) [70] описаны подходы к построению динамических тематических моделей с фиксированным словарем. Эти ВТМ позволяют проследить изменение тематики во времени, но не позволяют оценить изменение словарного состава модели.

В работе Online Latent Dirichlet Allocation with Infinite Vocabulary [110] – описан алгоритм создания онлайн ВТМ с бесконечным словарем. Словарный состав ВТМ не меняется в размере, но изменяется по своему составу по мере добавления новых документов.

В работе On-line Trend Analysis with Topic Models: #twitter trends detection topic model online [74] – предложен алгоритм построения онлайн ВТМ с изменяемым словарем. В модели документы обрабатываются итеративно в рамках одного «окна». Новые слова, встретившиеся более 10 раз, добавляются в модель, а старые слова, встретившиеся менее 10 раз в «окне», удаляются из модели. Словарь ВТМ может изменяться по размеру и составу в ходе добавления новых документов.

Таким образом, существуют алгоритмы тематического моделирования с изменяемым словарем, в которых тематический вектор «нового слова» берется из

равномерного распределения Дирихле либо из процесса Дирихле, но не предложено подходов в определении тем «нового слова» по тематическим векторам документов, где это слово встретилось.

Методы определения темы «нового слова»

Допустимо предположение, что темы «нового слова», впервые появившегося в BTM, связаны с темами документа, где это слово встретилось. Для того чтобы определить эту связь, возьмем набор новых документов и определим их темы с помощью алгоритма ml-PLSI, предложенного в [18]. Для определения тем «нового слова» нужны тематические векторы документов, где эти слова встретились. Слова, редко встречающиеся в коллекции документов, не значимы для BTM. Чем больше документов содержат «новое слово», тем точнее определяется тематическая принадлежность этого слова.

Возможны два метода расчета тематического вектора «нового слова», по сумме тематических векторов документов, где это слово встретилось, и через произведение Адамара.

Рассмотрим подробнее результаты суммы и произведения Адамара векторов на примере. Предположим, что в BTM три темы и три документа с новым словом. Векторы тем новых слов равны {0.5 0.5 0.0}, {0.3 0.2 0.5}, {0.1 0.2 0.7}. Сумма векторов равна:

$$\begin{matrix} 0.5 \\ 0.5 \\ 0.0 \end{matrix} + \begin{matrix} 0.3 \\ 0.2 \\ 0.5 \end{matrix} + \begin{matrix} 0.1 \\ 0.2 \\ 0.7 \end{matrix} = \begin{matrix} 0.9 \\ 0.9 \\ 1.2 \end{matrix} \quad (23)$$

Результат произведения Адамара этих векторов:

$$\begin{matrix} 0.5 \\ 0.5 \\ 0.0 \end{matrix} * \begin{matrix} 0.3 \\ 0.2 \\ 0.5 \end{matrix} * \begin{matrix} 0.1 \\ 0.2 \\ 0.7 \end{matrix} = \begin{matrix} 0.015 \\ 0.02 \\ 0.0 \end{matrix} \quad (24)$$

Нормированные векторы вероятностей тем «нового слова» представлены на рисунке 24 для суммы (Summ) и произведения Адамара (Hadamard) векторов тем документов, где это слово встретилось.

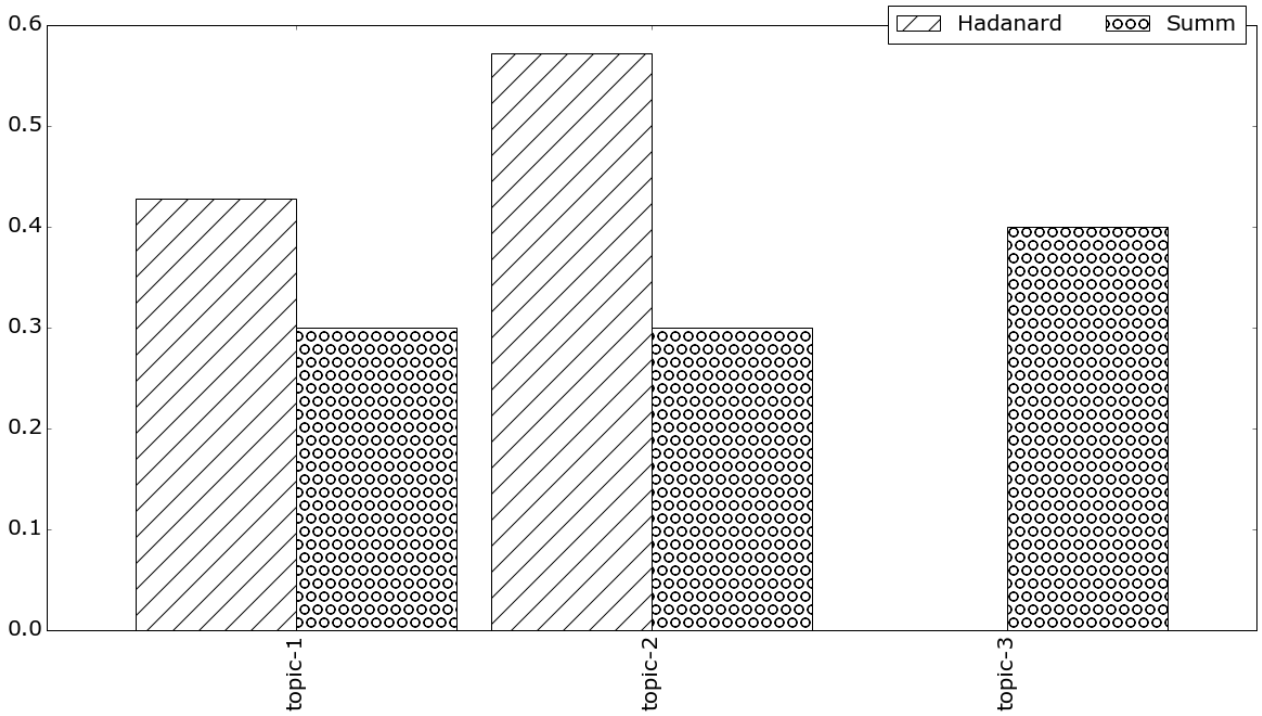


Рисунок 24 – Пример вектора суммы и произведения Адамара для трех тем

Суммы вероятностей для первой и второй темы равны, хотя значения вероятностных оценок в документах для этих тем сильно отличаются. Вывод: при суммировании векторов тем теряется значимая для определения тематики «нового слова» информация. По сумме векторов тем новое слово, с наибольшей вероятностью, относится к третьей теме, но третий документ не относится к третьей теме, следовательно, и слово, которое в нем встретилось, не относится к этой теме. Сумма вероятностей ошибочно связывает «новое слово» с одной из тем, а произведение Адамара обнуляет вероятность для этой темы. Таким образом, применение произведения Адамара для определения тематики «нового слова» точнее отражает тематическую принадлежность этого слова, так как обнуляет значение вероятности для непересекающихся векторов тем.

В методе с использованием произведения Адамара существует вероятность появления нулевых векторов из-за наличия ошибочно использованных слов в не соответствующих этим словам документах. Пример обнуления значений тематического вектора:

$$\begin{matrix} 0.1 & 0.0 & 0.0 \\ 0.9 * 0.0 = & & 0.0 \\ 0.0 & 1.0 & 0.0 \end{matrix} \quad (25)$$

Обнуление векторов тем «нового слова» возможно в случае ошибочного использования слов в несвойственной для этого слова теме. Необходимо идентифицировать ошибочное использование слова в наборе документов, где встретилось новое слово, и не учитывать его в дальнейшем.

Метод определения темы «нового слова» через произведения Адамара. Для расширения словаря BTM предлагается использовать следующий алгоритм (Листинг 3) (рисунок 25):

$$p_{new}(w|z) = p_{i=1}(d|z) \circ p_{i+1}(d|z) \circ \dots \circ p_n(d|z). \quad (26)$$

Листинг 3: Расширение словаря BTM

Вход: Корпус SCTM-ru, разделенный на обучающее и тестовое множество документов, тестовое множество как поток документов.

Выход: Тематическая принадлежность «нового слова» и BTM с бесконечным словарем.

1. Создаем BTM на обучающем множестве $D_1 \bigcap D_2 = \emptyset, D_1 + D_2 = D$.
2. Извлекаем «новые документы» из тестового множества D_2.
3. Для всех «новых документов» d_{new} определяем тематическую принадлежность с помощью алгоритма ml-PLSI:
$$p(d|z) = \sum_{w \in d} p(w|z).$$
4. Последовательно получаем документы из тестового множества:

a. Если встречаем «новое слово», то сохраняем слово и вектор тем документа, где оно встретилось, во временное хранилище данных;
b. Если встречаем «новое слово», которое ранее было сохранено, то считаем вектор тем для слова с помощью произведения Адамара сохраненного вектора тем и текущего вектора:

$$p_{new}(w|z) = p_{i=1}(w|z) \circ p_{i+1(w|z)}.$$

5. Рассчитываем косинусное расстояние между новым полученным вектором тем для «нового слова» и предыдущим значением.
   a. Если $\cos(p_i, p_{i+1}) > 0.99$, то считаем вектор корректным и стабилизировавшимся.
   b. Если $\cos(p_i, p_{i+1}) < 0.99$, то считаем вектор нестабильным, обновляем сохраненный вектор тем для «нового слова».
6. Для корректных, стабильных векторов обновляем матрицы «слово-тема»:
   a. Добавляем в матрицу «документ-тема» тематические вектора новых документов,
   b. Добавляем в матрицу «слово-документ» вектора новых документов с количеством слов в них,
   c. Добавляем в матрице «слово-тема» тематический вектор «нового слова».

Построение и регулярный пересчет BTM на корпусе текстов вычислительно сложная задача, поэтому при работе с динамическими и онлайн BTM актуально дополнение модели новыми словами и документами. Произведение Адамара – это бинарная операция над двумя векторами, сложность вычисления линейно зависит от длины векторов. Поэтому предложенный алгоритм [19] может быть использован в масштабных производственных системах.

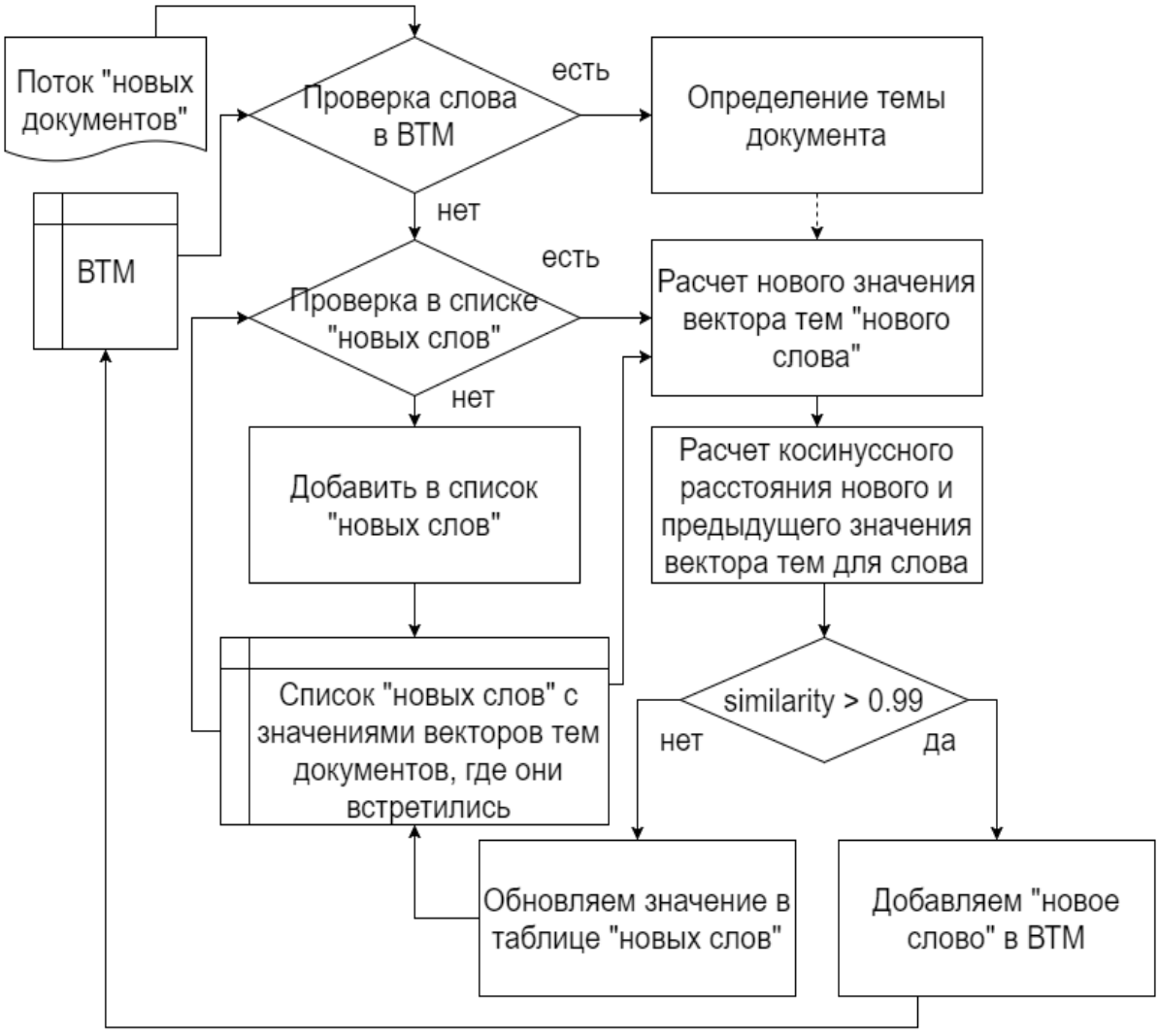


Рисунок 25 – Схема алгоритма определения тем «нового слова» и добавление его в ВТМ

Выводы к главе 3

- Предложен метод построения ВТМ через обучение на размеченных данных.
- Предложен метод многозначной классификации текстовых документов с использованием обученной тематической модели ml-PLSI. Для

предложенного алгоритма сделана экспериментальная оценка качества классификации.

- Предложено несколько методов определения тем «нового слова» в BTM с использованием суммы и произведения Адамара тематических векторов документов, где это слово встретилось. Методу определения тем «нового слова» через произведение Адамара дано теоретическое обоснование. Для предложенного метода приведены теоретические оценки вычислительной эффективности.
- Предложен алгоритм пополнения BTM, разработана BTM с бесконечным словарем.

4. Разработанный программный комплекс вероятностного тематического моделирования потока текстовых документов

Четвертая глава посвящена описанию выполненного процесса проектирования и разработки прототипа комплекса программных средств вероятностного тематического моделирования для анализа текстовых документов, рассмотрению методов применения системы для решения прикладных задач интеллектуального анализа текстов. Разработана архитектура комплекса, схемы отдельных частей, предложен сценарий использования BTM в задачах информационного поиска и в рекомендательных сервисах.

4.1. Архитектура программного комплекса

На основе требований к программному комплексу, изложенных в 2 главе для реализации программного комплекса, предложена микросервисная архитектура (Microservice Architecture).

Микросервисы – это небольшие, автономные, совместно работающие сервисы [34, с. 23].

В работе [62] дано следующее определение микросервисов. Микросервис – это минимальный независимый процесс, взаимодействующий с помощью сообщений. Микросервисная архитектура – это распределенное приложение, в котором все модули являются микросервисами.

Микросервисная архитектура – это архитектура на основе свободно сопряженных сервисов с ограниченными контекстами. Ограниченный контекст – это понятие явных границ вокруг какого-то бизнес-контекста [64].

Микросервисная архитектура является разновидностью Сервис-ориентированной архитектуры (Service-oriented Architecture SOA) [34, 64]. В работе [56] микросервисы указаны как один из стратегических технологических трендов на 2017 год.

В работах [34, 40, 64] выделены основные принципы, преимущества и свойства архитектуры, основанной на микросервисах.

Основные принципы микросервисов:

- микросервисы – небольшие, автономные, выполняют каждый свою единственную функцию, выполняются в своем процессе и взаимодействуют с остальными сервисами по протоколам HTTP и WebSockets;
- организационная культура включает в себя автоматизацию разработки, тестирования и поддержки, это снижает нагрузку на управление и снижает стоимость разработки;
- каждый микросервис – эластичный, легко меняющийся, элементарный и при этом законченный.

Преимущества, характерные для архитектуры микросервисов:

- Микросервисы можно легко заменить в любое время. Это свойство повышает устойчивость программных решений. Вышедший из строя микросервис не останавливает работу всей системы.
- Микросервисы организованы вокруг функций, каждый микросервис реализует специфические возможности в предметной области и свою бизнес-логику в рамках определенного контекста, должен разрабатываться автономно и независимо развертываться через автоматизированные механизмы.
- Микросервисы могут быть реализованы с помощью различных технологий, программных и аппаратных средств, обладают технологической разнородностью.
- К аппаратным преимуществам микросервисов относят независимую масштабируемость. Свойство масштабируемости касается отдельно

взятого микросервиса или приложения, но не всего программного комплекса.

- Независимая эволюция подсистем и микросервисов: микросервис может развиваться и ломать обратную совместимость, при этом в приложении может работать прежний микросервис еще некоторое время.
- Непрерывное развертывание – возможность быстро обновлять программный комплекс и вводить в эксплуатацию новые функции. Реализуется возможность частичного развертывания – возможности частичного обновления системы.
- Гибкость – функциональность разбивается на малые автономные микросервисы, масштабируется независимым развертыванием и репликацией этих микросервисов по серверам, виртуальным машинам, контейнерам (рисунок 26).
- Независимость данных каждого микросервиса продемонстрирована на рисунке 27. Различают микросервисы без состояния и микросервисы с состоянием. У каждого микросервиса должна быть своя модель предметной области. У микросервисов без состояния данные находятся во внешних базах данных. У микросервисов с состоянием данные находятся в том же микросервисе.

Отмеченные преимущества микросервисной архитектуры позволят разработать программный комплекс, удовлетворяющий требованиям, определенным в 2 главе, позволят создать приложения, соответствующие человеко-ориентированному, событийно-ориентированному и определяемому-данным подходам.

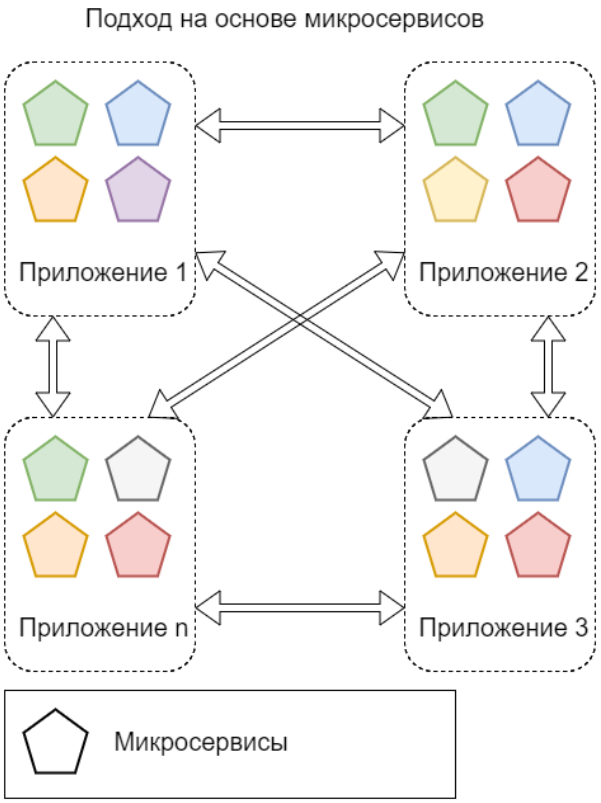


Рисунок 26 – Подход на основе микросервисов

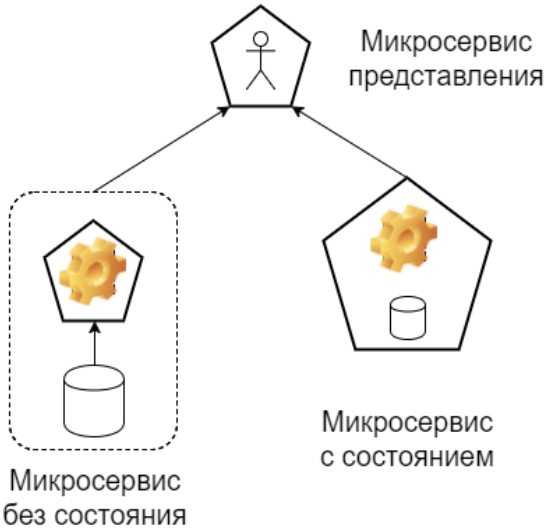


Рисунок 27 – Независимость данных в микросервисах

Программный комплекс строится как набор небольших сервисов, каждый из которых работает в собственном процессе и коммуницирует с остальными, используя легковесные механизмы, например HTTP. Пользователь управляет работой отдельных микросервисов и программного комплекса в целом, инициирует процессы, выполняемые в микросервисах. Взаимодействие между отдельными микросервисами осуществляется через API. Микросервисная архитектура позволяет изменить методы и алгоритмы, реализованные в отдельном микросервисе, безболезненно по отношению к целостности системы. Архитектура программного комплекса представлена на рисунке 28.

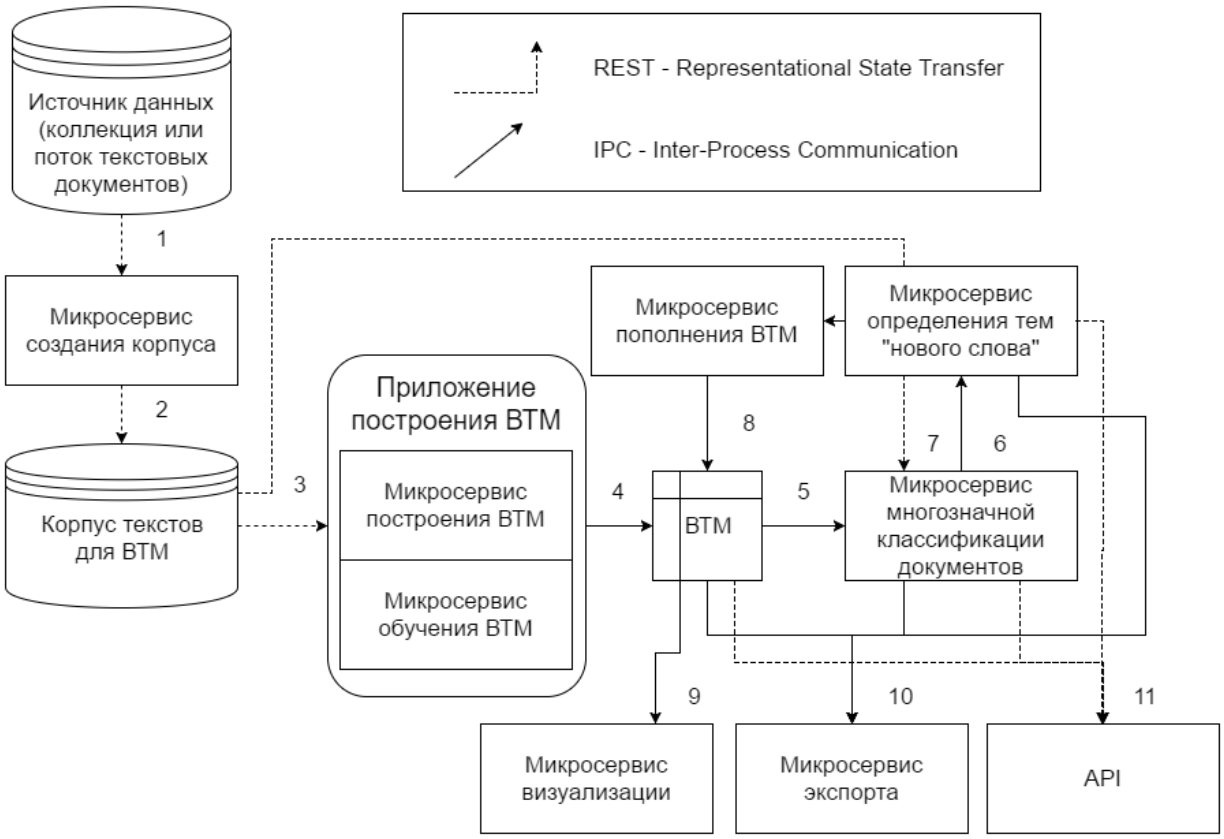


Рисунок 28 – Архитектура системы вероятностного тематического моделирования

Разработка программного комплекса велась в соответствии с правилами гибкой методологии разработки программного обеспечения (Agile) [30]. Гибкая методология разработки трактует изменения как внутренние аспекты проектирования программного обеспечения и предлагает «облегченные» методы для реагирования на изменения требований, выводя на передний план разработки вопросы программирования. Основные принципы сформулированы в Agile манифесте [46]:

- Люди и их взаимодействие важнее, чем процессы и средства.
- Работающее программное обеспечение важнее, чем исчерпывающая документация.
- Сотрудничество с заказчиком важнее, чем обсуждение условий контракта.
- Реагирование на изменения важнее, чем следование плану.

Гибкая методология разработки хорошо сочетается с концепцией микросервисной архитектуры. В Agile этапы жизненного цикла программного обеспечения – анализ, проектирование, реализация и развертывание – заменяются новыми терминами: пользовательские истории, примерочные тесты, рефакторинг, разработка через тестирование, непрерывная интеграция. Выбранная методология разработки позволила приступить к реализации отдельных микросервисов программного комплекса, как только появилась необходимость в этих микросервисах, без этапа планирования разработки всего программного комплекса. Планирование времени разработки велось на основе принципов «Схватки» (Scrum) [38]. Разработка велась равными итерациями, называемыми спринтами. За один спринт создавался работающий микросервис, либо производилось улучшение ранее созданного микросервиса. Такой подход позволил разрабатывать микросервисы автономно, по мере формирования функциональных требований, а затем объединить их в программный комплекс.

На схеме архитектуры программного комплекса (рисунок 27) «источник данных» – это один или несколько источников данных. Для работы программного комплекса источником данных может быть коллекция либо поток текстовых документов.

Микросервис «создания корпуса» предназначен для предварительной обработки входных данных в формат корпуса текстов, пригодного для построения BTM. В предварительную подготовку входит очистка данных от разметки, извлечение значимой для вероятностного тематического моделирования информации, морфологическая обработка.

Корпус текстов для BTM содержит набор текстовых документов, на которых будет строиться BTM в XML или CSV формате.

Приложение построения BTM включает два микросервиса. Один для построения классических BTM, второй для построения BTM на основе обучения на размеченных данных.

BTM – рабочая вероятностная тематическая модель.

Микросервис многозначной классификации документов предназначен для определения тем документов, поступающих в потоке после построения BTM. Микросервис имеет доступ к построенной на данный момент BTM и к потоку текстовых документов, прошедших предварительную обработку.

Микросервис определения тем «нового слова» предназначен для определения вектора тем для слова, которое ранее не встречалось в потоке текстовых документов и является новым для BTM. BTM не известно к каким темам относится это «новое слово».

Микросервис пополнения BTM выполняет важную роль добавления данных в ранее построенную BTM. Добавляется информация о новых документах, поступающих в потоке текстов, о новых словах, темах новых документов, темах новых слов.

Микросервис многозначной классификации определения тем «нового слова» и ВТМ умеют передавать результат своей работы внешним системам посредствам API.

Микросервис визуализации предназначен для представления результатов вероятностного тематического моделирования пользователю программного комплекса в графическом, интуитивно понятом виде, с помощью диаграмм, графиков и интерактивных средств.

Микросервис экспорта предназначен для экспорта результатов вероятностного тематического моделирования в необходимый для пользователя программного комплекса формат.

Для реализации программного комплекса в качестве базовых были выбраны следующие средства:

- языковая платформа Python
- интерактивная оболочка iPython
- дистрибутив python Anaconda
- библиотеки: gensim, pyldavis, matplotlib, wordcloud

Выбранные средства ни в коем случае не являются ограничением для разработки, т.к. микросервисная архитектура позволяет реализовывать каждый микросервис с использованием наиболее подходящего технологического стека.

Python – высокоуровневый язык программирования общего назначения, поддерживающий несколько парадигм программирования, в том числе структурное, объектно-ориентированное, функциональное, императивное и аспектно-ориентированное. Код в Python организовывается в функции и классы, которые могут объединяться в модули. Кроссплатформенное программное обеспечение [11, 26, 27].

IPython – интерактивная оболочка для языка программирования Python, которая предоставляет расширенную интроспекцию, дополнительный командный синтаксис, подсветку кода и автоматическое дополнение и является компонентом пакетов программ SciPy и Anaconda. Кроссплатформенное программное обеспечение.

Anaconda – дистрибутив языков программирования Python и R, включающий в себя набор библиотек для научных и инженерных расчетов, менеджер пакетов conda, интерактивную оболочку IPython.

Gensim – библиотека с открытым исходным кодом для моделирования в векторном пространстве и построения TM-LDA [16, 27].

Вместе выбранные программные средства и библиотеки обеспечивают необходимую гибкость и функциональность для реализации программного комплекса.

4.2. Микросервисы программного комплекса

Микросервис создания корпуса текстов для построения BTM

Микросервис создания корпуса включает:

1. извлечение данных из источника;
2. удаление из текстов извлеченных документов элементов разметки;
3. морфологическую предобработку.

В качестве источника данных использован международный новостной сайт «Русские Викиновости» (Викиновости), тексты статей которого распространяются по свободной лицензии Creative Commons Attribution 2.5 Generic, доступны для скачивания и анализа на любых компьютерах, в том числе на компьютерах без доступа в Интернет. В работах [22, 39] отмечены преимущества вики-ресурсов, таких как Викисловарь и Википедия, для использования в качестве источника данных в исследовательских целях. Вики-ресурсы – это сайты второго поколения Интернет, характеризуемые тем, что к их наполнению привлечено огромное количество рядовых пользователей, с помощью которых происходит пополнение и актуализация информации. Большой объем, постоянное пополнение, нейтральность во взглядах, доступность относятся к преимуществам всех вики-ресурсов, в том числе к Викиновостям.

Викиновости – это братский проект большой Википедии, предназначенный для написания новостных статей. Пример статьи Викиновостей представлен на рисунке 29. Отличительной особенностью сайта Викиновостей от любого другого новостного сайта является то, что каждый человек может принять участие в создании новости. Правила Викиновостей требуют писать новости с нейтральной точки зрения, в непредвзятом виде, выбирать существенные и актуальные темы, использовать достоверные источники.

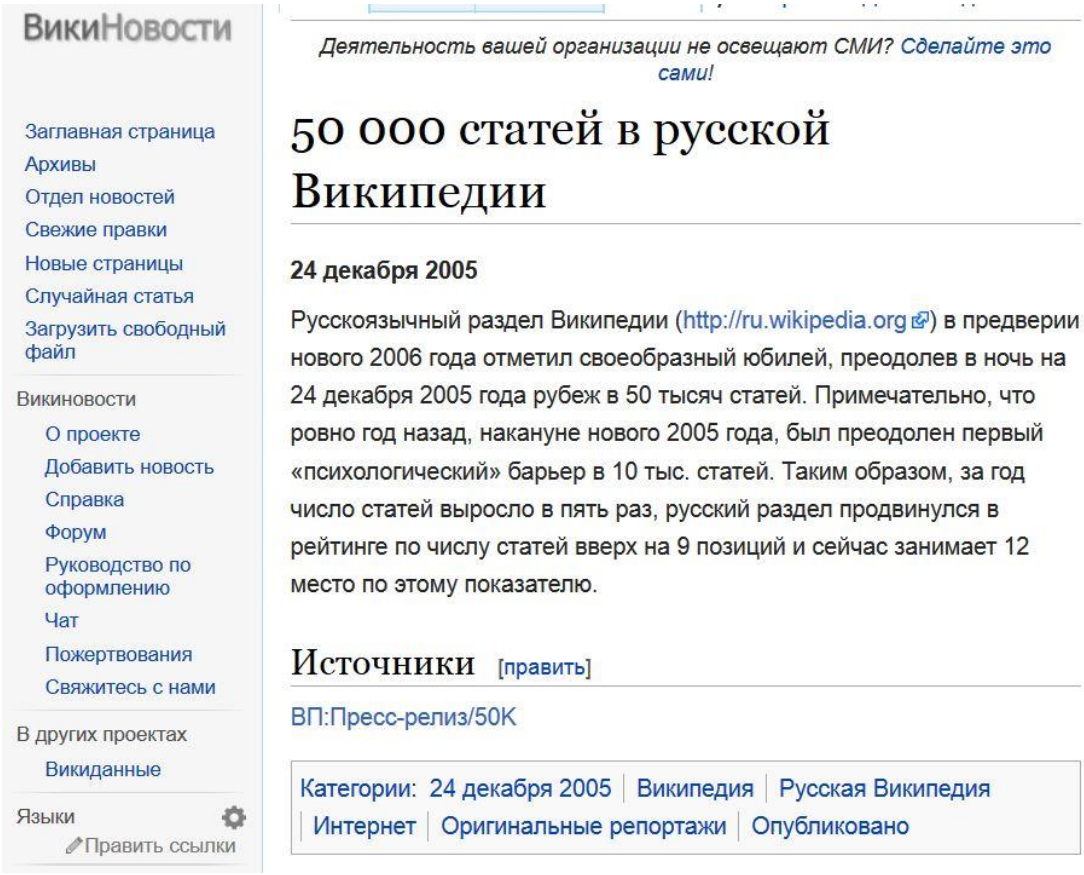


Рисунок 29 – Статья "50 000 статей в русской Википедии" на сайте русских Викиновостей

Пример части XML-дерева экспортного файла базы данных Викиновостей представлен на рисунке 30.

```
<?xml version="1.0" encoding="utf-8"?>
<page>
  <title>50 000 статей в русской Википедии</title>
  <ns>0</ns>
  <id>1838</id>
  <revision>
    <id>75780</id>
    <parentid>39949</parentid>
    <timestamp>2011-10-01T18:45:01Z</timestamp>
    <contributor>
      <username>Schekinov Alexey Victorovich</username>
      <id>2156</id>
    </contributor>
    <text xml:space="preserve">{{:Дата |24 декабря 2005}}
{{ВикипедияН
 |Язык = Русская
 }}
Русскоязычный раздел [[Википедия|Википедии]] (http://ru.wikipedia.org)
в предверии нового 2006 года отметил своеобразный юбилей, преодолев в
ночь на 24 декабря 2005 года рубеж в 50 тысяч статей. Примечательно,
что ровно год назад, накануне нового 2005 года, был преодолен первый
«психологический» барьер в 10 тыс. статей. Таким образом, за год число
статей выросло в пять раз, русский раздел продвинулся в рейтинге по
числу статей вверх на 9 позиций и сейчас занимает 12 место по этому
показателю.
{{оригинальный репортаж 2}}
== Источники ==
{{w|ВП:Пресс-релиз/50К}}
{{публиковать}}
[[Категория:Русская Википедия]]</text>
    <sha1>ezckutzzznn6tioszytixr2atkv0v4p</sha1>
    <model>wikitext</model>
    <format>text/x-wiki</format>
  </revision>
</page>
```

Рисунок 30 – Пример XML статьи "50 000 статей в русской Википедии" на сайте русских Викиновостей

В таблице 9 представлено описание дерева XML элементов файла экспорта Викиновостей, отмечены элементы, востребованные для построения BTM.

Таблица 9 – Дерево XML элементов экспортного файла Викиновостей

Родительский элемент	XML элемент	Востребованность для построения ВТМ	Описание
<root>	<page>	Да	группа элементов новостной статьи
<page>	<title>	Да	название статьи
<page>	<ns>	Нет	идентификатор или имя пространства имен (namespace), элемент предназначен для отделения основных статей от служебных, ноль соответствует основному пространству имен
<page>	<id>	Да	уникальный идентификатор статьи
<page>	<revision>	Да	ревизия – это группа элементов актуальной версии статьи
<revision>	<id>	Нет	первичный ключ ревизии, используется для контроля изменений статьи
<revision>	<parented>	Нет	идентификатор родительской статьи
<revision>	<timestamp>	Нет	дата и время создания ревизии статьи
<revision>	<contributor>	Да	группа элементов авторства статьи
<contributor>	<username>	Нет	имя автора статьи

Продолжение таблицы 9

<contributor>	<id>	Да	уникальный идентификатор автора статьи
<revision>	<text>	Да	текст статьи с элементами вики-разметки
<revision>	<sha1>	Нет	хеш код статьи, полученный алгоритмом криптографического хеширования SHA-1, используется для контроля версий
<revision>	<model>	Нет	модель контента статьи, в данном случае wikitext
<revision>	<format>	Нет	формат данных статьи, в данном случае text/x-wiki

Предварительная обработка данных Викиновостей

В экспортном файле Викиновостей статьи отсортированы по дате создания ревизии <timestamp>, эта дата не связана с датой описанных событий. Авторам рекомендуется указывать с помощью вики-разметки дату событий в тексте статьи. Пример вики-разметки даты {{:Дата | 24 декабря 2005}} представлен на рис. 30 внутри элемента <text>. Часть статей в экспортном файле Викиновостей не содержит дату событий в вики-разметке, но при этом она указана в тексте или в категории. Чтобы сохранить максимум востребованной в алгоритмах тематического моделирования информации, дата событий была по возможности восстановлена из текста и категорий. В 455 статьях не удалось восстановить дату событий, эти статьи являются подборками новостей, произошедших в один день, в разные годы и не представляют

ценности для построения тематических моделей, они были исключены из корпуса. Документы корпуса SCTM-ru отсортированы по дате событий, от старых к новым.

В экспортном файле базы данных Викиновостей содержится информация об авторе последней ревизии статьи. Используем эту информацию как идентификатор авторства для построения автор-тематических моделей. Так как 58 Викиновостей не содержат информацию об авторе, а статьи ценны, то было принято техническое решение присвоить этим статьям уникальный идентификатор автора – 2 и включить их в корпус SCTM-ru.

Текст статьи Викиновостей содержит оформленные специальным образом ссылки. Ссылки делятся на три группы: внутренние – инструмент связывания страниц внутри языкового раздела Википедии, межъязыковые ссылки (интервики) – средство для организации связей между различными вики-системами в сети Интернет и ссылки на страницы братских вики-проектов (например, на Википедию). Текст статьи, заключенный в двойные квадратные скобки, является внутренней ссылкой, пример [[Википедия|Википедии]] представлен на рисунке 29. Если падеж ссылающегося слова или словосочетания не совпадает с именительным падежом, то в двойных квадратных скобках стоит черта, слева от которой указан именительный падеж текста ссылки, а справа текст, соответствующий грамматике предложения. Алгоритмы тематического моделирования учитывают количество вхождений каждой леммы слова в текст. Во внутренних ссылках каждое слово имеет два вхождения в разных словоформах и будет дважды учтено в тематической модели, тем самым исказив частотные характеристики модели. В документах корпуса SCTM-ru оставлена только та часть ссылки, которая соответствует грамматике предложения.

Новости должны сопровождаться ссылками на документальный источник. Они обычно делятся на четыре вида: другие статьи Викиновостей, внешние ссылки на онлайн-источники, цитаты печатных изданий и веб-сайты со справочной или связанной информацией. Для раздела статьи «Источники» используют вики-разметку

== Источники == (см. пример на рисунок 29). Для целей тематического моделирования ссылки на источники не представляют большой ценности, поэтому было принято решение об их исключении из корпуса SCTM-ru.

Важным элементом разметки Викиновостей и важными данными для построения тематических моделей является информация о категориях, к которым статья имеет отношение. Категории статьи определяет ее автор.

Для предварительной обработки текстов был разработан микросервис на языке C#, среда разработки Visual Studio Express 2013. Для поиска по экспортному файлу Викиновостей использовались регулярные выражения [7, 42]. Пример задействованных регулярных выражений представлен в таблице 10. Программа многомодульная, каждый модуль выполняет одну определенную операцию. Программа получает на вход исходный XML-файл, специально подготовленные регулярные выражения последовательно обходят файл в поисках совпадения по шаблону, на выходе создается XML-файл с внесенными за одну итерацию изменениями. Для сохранения целостности первоначальных данных, каждый проход по исходному XML-файлу вносит лишь часть изменений, которые внимательно проверяет администратор системы, после чего программу запускают с другим модулем обработки.

Для подсчета статистики корпуса SCTM-ru была разработана многомодульная программа. Модуль подсчета документов осуществляет разбор XML-дерева корпуса, извлекает уникальные идентификаторы каждого документа и считает их общее количество. Модуль подсчета авторов извлекает список уникальных идентификаторов авторов статей Викиновостей и подсчитывает их количество. Модуль подсчета категорий извлекает уникальные категории из XML-дерева корпуса и считает их количество. Модуль обработки дат описанных в статьях событий осуществляет разбор XML-дерева корпуса, извлекает информацию о дате события

каждого документа, подсчитывает уникальные значения, находит самую раннюю и самую позднюю дату документа.

Таблица 10 – Регулярные выражения для предварительной обработки текста

Регулярное выражение	**Назначение поиска**
^(=)?=(\s+)?Источник(и)?(\s+)?(=)?=\n(^([^\n]+)\n)+^\n	блок источники
\{\{Категории|([^|]+)|([^|]+)\}\}	блок категории
\[\[([^|]+)|([^\]]+)\]\]	ссылки

Для подсчета словарного состава корпуса SCTM-ru был разработан модуль с использованием регулярных выражений и программы MyStem. Модуль берет текст из заданных элементов XML-дерева (title, text), регулярные выражения из текста извлекают все последовательности букв русского алфавита. При подсчете слов последовательность букв русского алфавита, отделенная от других букв не буквами (например, знаки препинания, пробел), считается словом. Для определения лемм слов использовалась программа MyStem. Программа MyStem производит морфологический анализ текста на русском языке. Для слов, отсутствующих в словаре, порождаются гипотезы [116].

Разметка корпуса SCTM-ru

В качестве формата хранения корпуса SCTM-ru выбран XML (eXtensible Markup Language — расширяемый язык разметки), как один из наиболее удобных форматов для использования в программной среде и конвертации данных в другие форматы. Возможности XML позволяют сохранить текст исходной статьи Викиновости и выделить дополнительные параметры документа.

XML-файл корпуса (SCTM-ru) состоит из следующих элементов:

- <page> - группа элементов документа;
  - <title> - название документа;
  - <id> - уникальный идентификатор документа;

- <userid> - уникальный идентификатор автора;
- <category> - категория документа;
- <date> - дата событий документа;
- <text> - текст документа;

Пример разметки одного документа в корпусе SCTM-ru представлен на рисунке 31.

Заголовок документа (title) отделен от текста документа, т.к. словам заголовка может придаваться большее значение при построении тематической модели.

Уникальный идентификатор автора статьи (userid) – это параметр, который необходим в автор-тематических моделях. Автор-тематическая модель во времени (Author-Topic over Time) [108] представляет собой расширение LDA. При построении модели оценивается распределение авторов, тем и документов по времени.

Категории документа (category) – это указанные автором статьи категории. Например, на рисунке 30 в статье "50 000 статей в русской Википедии" указана категория «Русская Википедия». Информация о категории важна для тематического моделирования, поэтому сохранена в корпусе SCTM-ru (см. рисунок 31). Наличие информации о принадлежности документов к категориям позволит автоматически проверять точность, полноту, аккуратность тестируемых алгоритмов тематического моделирования. Информация о категориях документа может быть использована в моделях Labeled LDA, описанных в [90].

Дата описанных в статье событий (date) используется при построении временных (temporal) тематических моделей. Пример модели, использующей дату под названием «Тематики во времени» (Topic over Time - TOT) представлен в работе [104]. При построении временной модели наряду со стандартными распределениями слов по темам и тем по документам оцениваются распределения каждой темы по времени, что позволяет проследить и отобразить динамику изменения тем во времени.

```
<?xml version="1.0" encoding="utf-8"?>
<page>
  <title>50 000 статей в русской Википедии</title>
  <id>1838</id>
  <userid>2156</userid>
  <category>Русская Википедия</category>
  <data>24 декабря 2005</data>
  <text>
    Русскоязычный раздел Википедии (http://ru.wikipedia.org) в предверии
    нового 2006 года отметил своеобразный юбилей, преодолев в ночь на 24
    декабря 2005 года рубеж в 50 тысяч статей. Примечательно, что ровно
    год назад, накануне нового 2005 года, был преодолен первый
    «психологический» барьер в 10 тыс. статей. Таким образом, за год число
    статей выросло в пять раз, русский раздел продвинулся в рейтинге по
    числу статей вверх на 9 позиций и сейчас занимает 12 место по этому
    показателю.
  </text>
</page>
```

Рисунок 31 – Пример XML-документа "50 000 статей в русской Википедии" в корпусе SCTM-ru

Текст документа (text) соответствует тексту исходной статьи. Мы целенаправленно оставляем исходный текст без изменения, без преобразования его в модель «мешка слов», без лингвистической обработки для возможности исследования уникальных особенностей русского языка. Информация о последовательности слов в тексте документа используется в моделях, учитывающих взаимную встречаемость слов. Например, модель под названием «Скрытая тематическая Марковская модель» (Hidden Topic Markov’s Model - HTMM), описанная в работе [69], основана на предположениях, что слова в составе предложения, а также сами предложения связаны одной общей темой и темы слов в документе образуют цепь Маркова. В результате работы HTMM уменьшает неоднозначность слов, расширяет понимание темы.

Реализован микросервис предварительной обработки Викиновостей для удаления невостребованных при построении BTM данных и извлечения

необходимых. Предложена XML схема разметки корпуса. Сделано статистическое исследование корпуса SCTM-ru. Источником данных корпуса является международный новостной сайт «Русские Викиновости». Корпус SCTM-ru состоит из 7 тыс. документов, 185 авторов, почти 12 тыс. уникальных категорий. События, описанные в документах, распределены по более чем 2 тыс. уникальным датам, с ноября 2005 года по июнь 2014 года. В корпусе SCTM-ru 2,4 млн словоупотреблений, состоящих только из русских букв. Словарный состав корпуса – 150,6 тыс. уникальных словоформ, 59 тыс. уникальных лемм. В 2017 году был обновлен микросервис для создания корпуса текстов SCTM-ru, обновление было реализовано на языке Python, и был обновлен корпус текстов. Корпус SCTM-ru 2.0 состоит из 12 тыс. документов, 320 авторов. События, описанные в документах, распределены по датам с ноября 2005 года по январь 2017 года. В корпусе SCTM-ru более 2,5 млн словоупотреблений, состоящих только из русских букв. Словарный состав корпуса – 262 тыс. уникальных словоформ. В обновленный корпус включены дополнительные значения данных, а именно: выделенные из текстов документов русскоязычные слова и выделенные из текстов документов имена существительные, приведенные к нормальной словоформе.

Объем созданного корпуса дает основания предположить его репрезентативность для различных задач автоматической обработки текстов на естественном языке. Как отмечено в работе [14] «Неразумно ждать пока кто-то по-научному сбалансирует корпус, перед теЗм как его использовать, и неосмотрительно было бы оценивать результаты анализа корпуса как «малодостоверные» или «неуместные» просто потому, что нельзя доказать, что используемый корпус «сбалансирован»». Разнообразие описанных в корпусе SCTM-ru событий и огромный коллектив авторов статей (21 тыс. участников) обосновывают предположение о его сбалансированности. Убедиться в сбалансированности корпуса можно после проведения анализа его внутренних признаков и построения тематических моделей.

Микросервис построения BTM

Приложение включает в себя два микросервиса. Один предназначен для построения BTM по коллекции текстовых документов, второй для построения BTM с помощью обучения на размеченных данных.

Для построения BTM по коллекции текстовых документов использована библиотека Gensim [113]. Данные в микросервис поступают из корпуса текстов для построения BTM в виде списка документов. Прежде чем построить модель, текстовые данные необходимо токенизировать, сконвертировать в формат, поддерживаемый библиотекой Gensim и создать словарь BTM. Для этих целей в библиотеке Gensim реализован класс corpora, который поддерживает несколько форматов корпусов для построения BTM, Market Matrix format [54], Joachim's SVMlight format [73] и GibbsLDA++ format [117]. В библиотеке Gensim реализовано несколько методов статистической обработки текста, при реализации микросервиса использованы те, которые позволяют создавать BTM – это ldamodel (Latent Dirichlet Allocation) [51] и hdpmodel (Hierarchical Dirichlet Process) [111]. Каждый метод обладает своим набором параметров вероятностного тематического моделирования, описание основных параметров представлено в таблице 11.

Основные методы класса LdaModel:

- clear() предназначен для очистки состояния модели;
- get_document_topics () возвращает список тем, вероятных для документа;
- get_term_topics () возвращает наиболее вероятные темы для слова;
- get_topic_terms () возвращает наиболее вероятные слова для темы;
- log_perplexity () предназначен для расчета правдоподобия;
- save () сохраняет модель в файл;
- show_topic () возвращает наиболее вероятные слова для темы;
- show_topics () возвращает список тем с наиболее вероятными словами для каждой темы;

- top_topic () рассчитывает согласованность темы для каждой темы [80];
- update () предназначен для обновления BTM без обновления словаря модели.

Таблица 11 – Параметры построения BTM в библиотеке genism

Параметр	Описание	Реализован в ldamodel	Реализован в hdpmodel
num_topics	Количество скрытых тем, которые необходимо извлечь	Да	Нет
id2word	Отображение идентификаторов слов (целые числа) в слова (строки)	Да	Да
Alpha	Гиперпараметр, который влияет на разреженность распределений документов	Да	Да
Eta	Гиперпараметр, который влияет на разреженность распределений слово-тема	Да	Да
Distributed	Параметр, необходимый для включения распределенных вычислений	Да	Нет
minimum_probability	Параметр, который контролирует фильтрацию тем для документа	Да	Нет

На рисунке 32 представлена диаграмма классов для LdaModel и HdpModel.

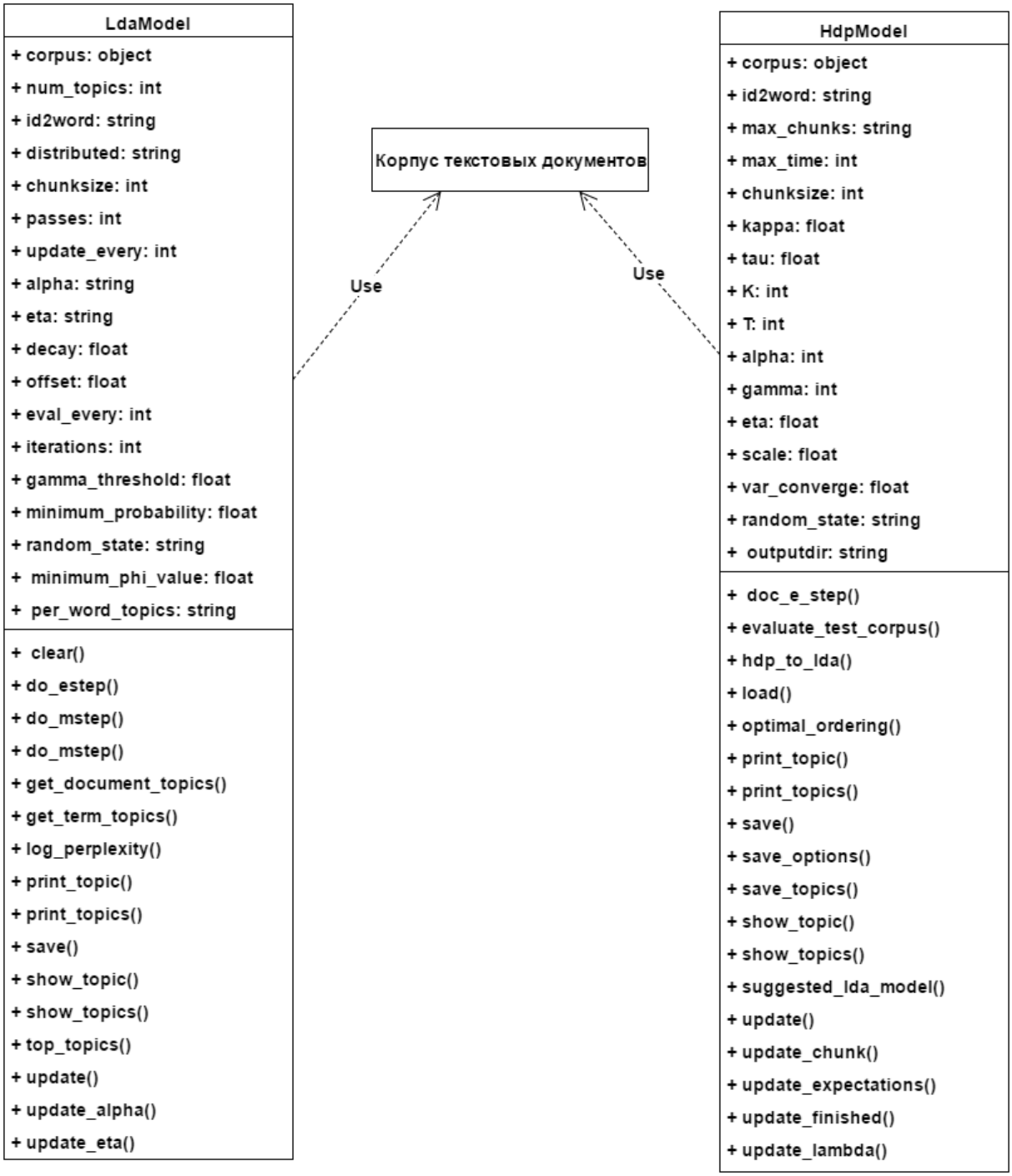


Рисунок 32 – Диаграмма классов LdaModel и HdpModel

Основные методы класса HdpModel:

- load () предназначен для загрузки, ранее сохраненной BTM;
- save () сохраняет модель в файл;
- show_topic () возвращает наиболее вероятные слова для темы;
- show_topics () возвращает список тем с наиболее вероятными словами для каждой темы;
- update () предназначен для обновления BTM без обновления словаря модели.

В микросервисе для построения BTM также может быть использована библиотека BigARTM [114].

В микросервисе построения BTM с обучением на размеченных данных реализован алгоритм, описанный в главе 3.2. Для построения BTM через обучение, корпус текстов, помимо текстов документов, должен содержать информацию о том, к каким темам относятся эти документы. В корпусе SCTM-ru содержится информация о принадлежности документов к категориям, поэтому корпус может быть использован для построения BTM через обучение на размеченных данных.

В таблице 12 приведены наиболее характерные и наиболее часто встречаемые слова для трех категорий: спорт, происшествия и политика корпуса SCTM-ru. Стоит заметить, что слова эти однозначно характеризуют категорию, к которой относятся, но не являются часто употребимыми во всем корпусе. Зачастую это имена собственные и фамилии либо редкие по написанию слова с буквой «ё»

Для оценки качества BTM используются встроенные методы log_perplexity (), top_topic (), и существует возможность реализовать расчет Перплексии [44, 51] и Когерентности [80] самостоятельно. Перплексия построенной методом обучения BTM равна $Perplexity = 0.1118$. Чем значение Перплексии меньше, тем лучше построенная BTM, в данном случае оценка получилась заниженной из-за метода построения модели [51].

Таблица 12 – Пример слов характерных для категорий

Спорт	**Происшествия**	**Политика**
Зимнему	Даги	Реймер
Виндсёрфинг	Дагов	Халип
Перепёлкин	Гоа	Гольман
Трофименко	Моди	Минтимер
Заезда	Зингер	Муртаза
Кайтингу	Ксанте	Рахимова
Полумарафона	Ока	Дарькина
Фонак	Реанимация	Спутникам
Гимнаст	Стропила	Хамитов

Построение ВТМ – важный шаг в анализе коллекции и потока текстовых документов, от его результата зависит репрезентативность представления вероятностного тематического моделирования. Следует уделять большое внимание подбору параметров вероятностного тематического моделирования для получения наиболее наглядного результата.

Микросервис многозначной классификации текстовых документов

Микросервис предназначен для многозначной классификации текстовых документов. Для реализации алгоритма классификации необходим набор данных для обучения, а именно коллекция документов, у которых определены темы, к которым они относятся. Для обучения классификатора использовался корпус текстов SCTM-ru, в котором автор каждой новости самостоятельно определил список тем, к котором относится документ. Обучение ВТМ выполнено с помощью микросервиса построения ВТМ.

Для оценки эффективности предложенного алгоритма проведен вычислительный эксперимент, сравнивающий алгоритмы ml-PLSI, случайный лес (Random forest), логистическую регрессию (Logistic regression). Для реализации вычислительного эксперимента созданы отдельные микросервисы, реализующие классификации случайный лес и логистическую регрессию. Схема разделения корпуса на обучающее и тестовое множество документов представлено на рисунке 21. Алгоритм ml-PLSI обучался на 5000 документах менее чем за 30 минут, обучить на таком же количестве документов Random forest, Logistic regression за 1 час не удалось, поэтому выборку для обучения сократили до 2000 документов. Оценка делалась на 100 новостях, для сравнения брались первые 50 предсказанных тем. Рассчитывалась функция потерь Хэмминга – отношение отрицательных предсказаний темы к общему количеству предсказаний. Доля предсказаний, в которых первая предсказанная тема входила в список тем документа, может восприниматься как оценка точности классификации. Доля предсказаний, в которых тема из первых 50 предсказанных входила в список тем документа, может интерпретироваться как оценка полноты классификации. Результат представлен в Таблице 13, алгоритм ml-PLSI точнее предсказывает первую наиболее вероятную тему и показывает лучшую полноту. Функция потерь Хэмминга высока у всех алгоритмов, для уменьшения ее значения следует сократить количество предсказываемых тем.

Таблица 13 – Сравнение качества 3 алгоритмов многозначной классификации

Метрика качества	ml-PLSI	Random forest	Logistic regression
Функция потерь Хэмминга	0,70	0,96	0,95
Первая предсказанная тема совпала с одной из тем документа	62%	11%	11%
Процент верных предсказаний из 50 первых тем	96%	53%	84%

Построенная в ходе эксперимента на корпусе SCTM-ru BTM позволяет определить наиболее вероятные темы не только для документа, но и для отдельной фразы или слова. Микросервис многозначной классификации, реализующий алгоритм mlPLSI в отличие от других алгоритмов многозначной классификации, может работать с неразмеченными данными, для этого сначала строится классическая BTM, а затем для каждого нового документа определяется список латентных тем из построенной тематической модели.

Микросервис добавления «нового слова»

С каждым годом словарный состав, используемый в повседневной жизни человека, меняется. Изменяются частотные характеристики употребления слов и выражений, устаревающие выходят из обихода, появляются новые. С ростом количества новых документов по темам растет словарный состав тем. Анализ корпуса SCTM-ru показал, что словарный состав отдельных тем изменяется неравномерно. Существуют темы, в которых от года к году появляется больше новых слов, в то же время присутствуют темы с незначительным изменением по словарному составу. На рисунке 33 представлено изменение количества документов и изменение количества уникальных слов в корпусе SCTM-ru по годам, в некоторые годы количество уникальных слов превышает 20 тыс. за год или порядка 60 новых слов в день.

BTM, построенная на текстах до 2010 года, не будет содержать информацию о половине слов BTM 2013 года, так как словарь новостей в 2013 году в два раза больше словаря новостей 2010 года. Можно заметить, что рост количества уникальных слов произошел вследствие увеличения количества документов в Викиновостях. На рисунке 34 представлено количество документов в тематиках «Некрологи», «Россия», «Украина», «Политика». Хорошо заметен рост количества новостей по тематикам «Россия» и «Политика». На рисунке 35 представлено количество уникальных слов в этих же тематиках.

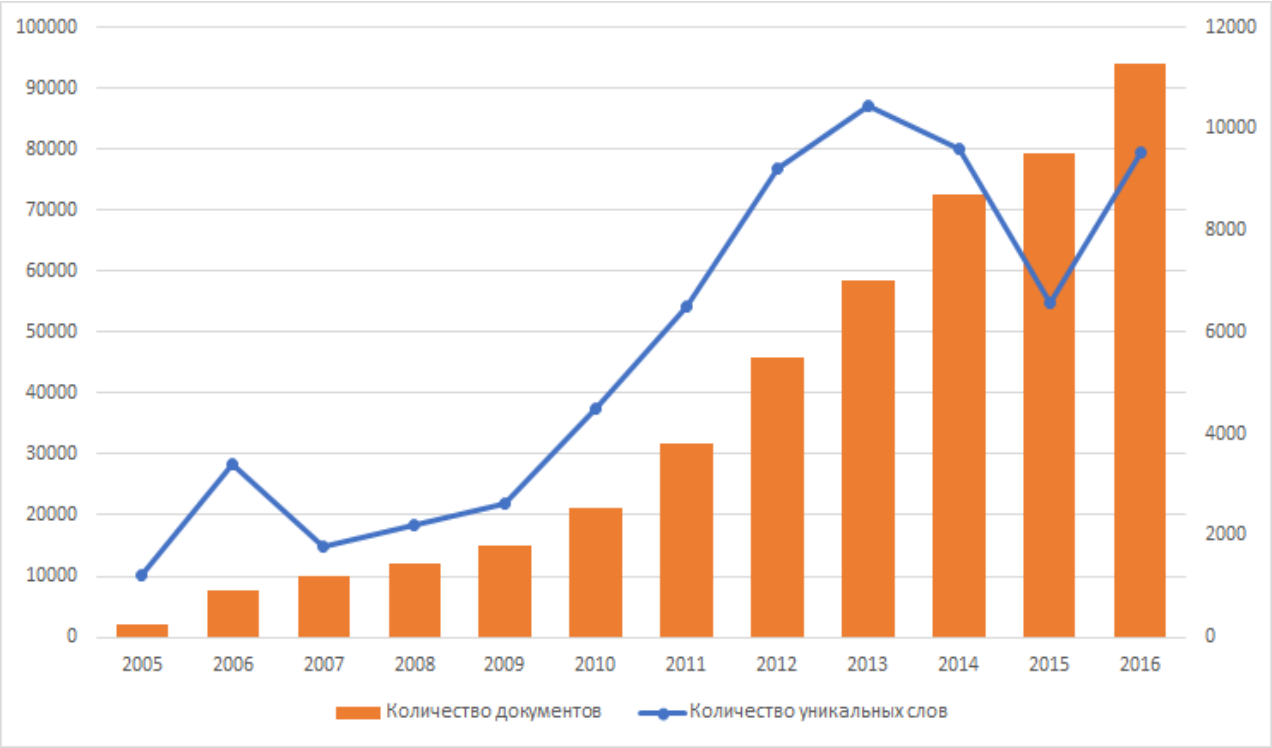


Рисунок 33 – Количество уникальных слов и количество документов в корпусе SCTM-ru по годам

Из графиков видно, что количество новых документов в тематиках «Россия» и «Политика» от года к году росло равномерно, при этом словарный состав тематики «Россия» рос быстрее, и по количеству уникальных слов тематика «Россия» опережает тематику «Политика». Тематика «Украина», в 4-е раза уступающая лидерам по количеству документов, по своему словарному составу лишь в 2-а раза меньше этих же тематик лидеров.

Задача микросервиса – определить темы для «нового слова» в BTM. Для этого в микросервисе реализован метод, описанный в главе 3.4. Микросервис реализован с помощью дистрибутива Anaconda, язык разработки Python, программные библиотеки для обработки и анализа данных – pandas, numpy, scikit-learn. Сначала, используя алгоритм ml-PLSI, строится BTM на данных для обучения. Затем последовательно обрабатываются все новости из тестовых данных. Запоминается каждое новое слово

и тематический вектор документа, где это слово встретилось. Для новых слов рассчитывается тематический вектор через сумму и через произведение Адамара тематических векторов документов, где это слово встретилось.

Исследование новых слов проводится на корпусе SCTM-ru [17]. Для проведения исследования разделим корпус на две части, над первой частью построим ТМ, вторую будем использовать в качестве объекта исследования. Новости с 2005 до конца 2012 года возьмем в качестве данных для обучения. Применим алгоритм multi-label Probabilistic Latent Semantic Indexing (ml-PLSI), описанный в работе [18] для обучения ТМ, который позволяет предсказывать тематическую принадлежность новых документов. Тестовый набор данных формируется из новостей за 2013 год.

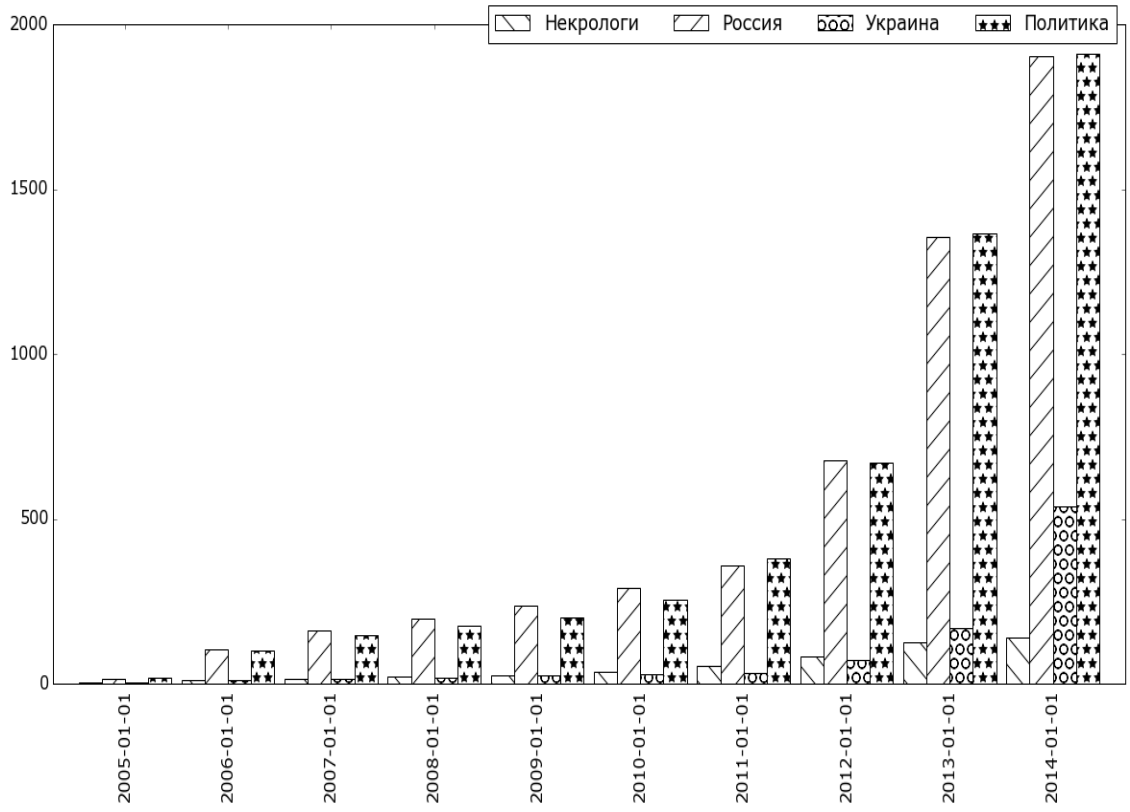


Рисунок 34 – Количество документов по темам в корпусе SCTM-ru, с 2005 по 2014 год

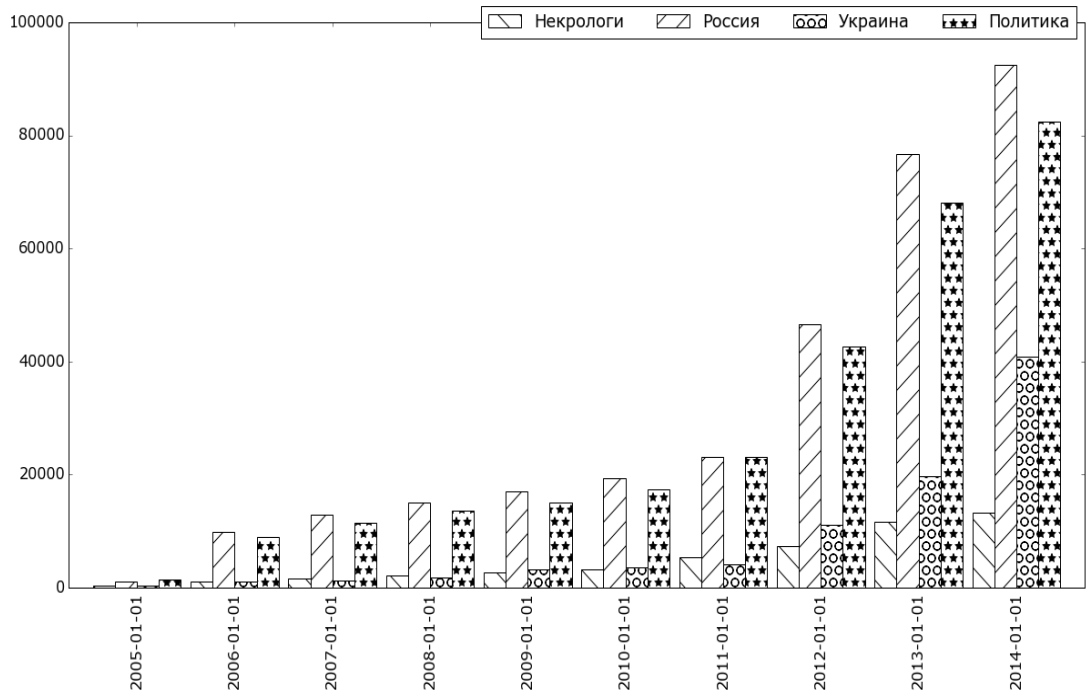


Рисунок 35 – Количество слов по темам в корпусе SCTM-ru, с 2005 по 2014 год

По определению тематического моделирования задача построения ТМ рассматривается как задача одновременной кластеризации слов и документов по одному и тому же множеству кластеров, называемых темами, поэтому размерность тематических векторов документов и слов одинакова. Алгоритм ml-PLSI определяет вектор тем нового документа по словарному составу этого документа: новые слова, которые впервые встретились в ТМ, никак не влияют на определение тематики нового документа. Резонно предположить, что тематика «нового слова», впервые появившегося в ТМ, каким-то образом связана с тематикой документа, где это слово встретилось. Для того чтобы определить эту связь, возьмем набор новых документов и определим их тематическую принадлежность. Для определения тематики «нового слова» нужны тематические векторы документов, где эти слова встретились.

Слова, редко встречающиеся в коллекции документов, не значимы для ТМ. Чем больше документов содержат новое слово, тем точнее определяется тематическая принадлежность этого слова по векторам тем документов. Пример новых слов приведен в таблице 14.

Таблица 14 – Пример новых слов в BTM

Слово	**Количество документов, содержащих слово**
Кипру	14
Олигарха	11
Царнаева	10
Тамерлан	9
Анастасиадис	7
Метеорита	6
Вальц	5

Для определения тематики «нового слова» «вальц» подсчитаем сумму тематических векторов документов, где это слово встретилось:

$$p_{new}(w|z) = \sum_{i=1}^{n} p_i(d|z), \tag{27}$$

где

- $p_{new}(w|z)$ – вектор тем «нового слова»,
- $p_i(d|z)$ – вектор тем документа, где встретилось слово,
- n – количество документов, где встретилось новое слово.

Для визуального представления отнесения «нового слова» ко всем тематикам модели отобразим нормированный вектор на графике. Результат представлен на рисунке 36, где по оси абсцисс – темы, по оси ординат – вероятности этих тем. На рисунке 37 показаны пять наиболее значимых тем, к которым относится слово «вальц».

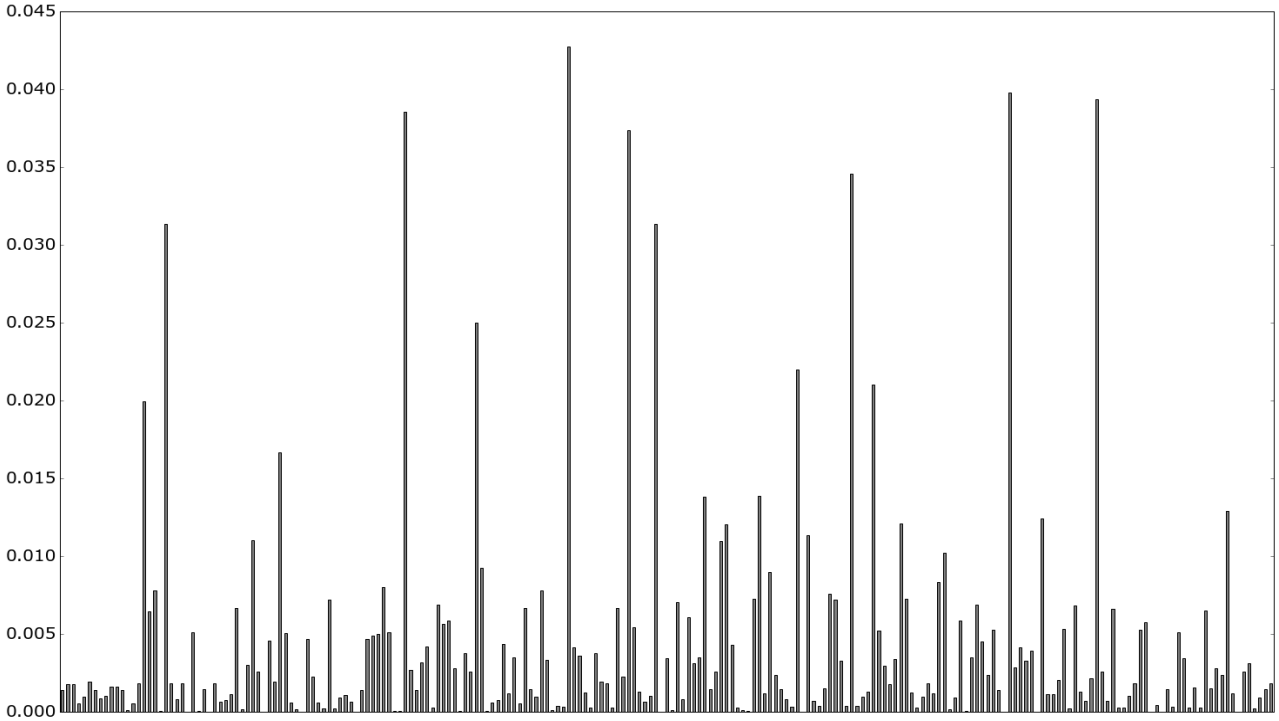


Рисунок 36 – График распределения суммы вероятностей тем для слова «вальц»

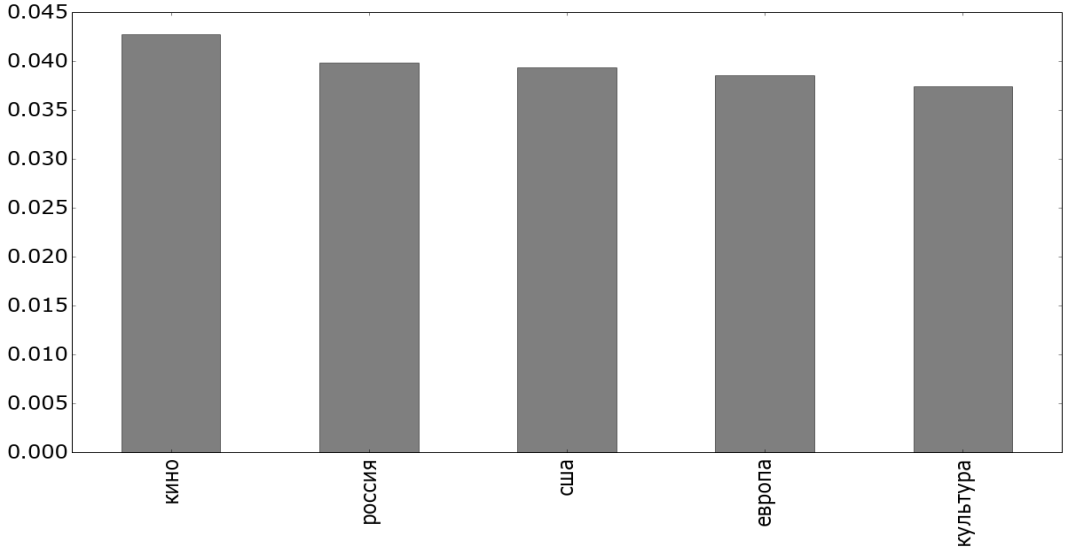


Рисунок 37 – Значимые темы от суммы вероятностей тем для слова «вальц»

Используем произведение тематических векторов документов, где встретилось новое слово, а именно покомпонентное произведение Адамара [72]:

$$p_{new}(w|z) = p_{i=1}(d|z) \circ p_{i+1}(d|z) \circ \dots \circ p_n(d|z). \quad (28)$$

На рисунке 38 представлен график нормированного распределения вероятностей тем для слова «вальц», полученный произведением Адамара тематических векторов документов, где это слово встретилось, а на рисунке 39 представлены наиболее вероятные темы для этого слова.

Как было отмечено в главе 3.2., при использовании произведения Адамара для определения тем «нового слова» возможны ошибки, связанные с ошибками во входных данных. Например, когда слово используется в несвязанном по теме документе, произведение Адамара может обнулить вектор тем этого слова. На тестовых данных можно подсчитать количество нулевых векторов, полученных от произведения Адамара для случая, когда учитываем все новые слова, встретившиеся в двух и более документах. Таких слов 2193. Нулевых векторов нет. Следовательно, в корпусе SCTM-ru нет ошибок, связанных с некорректным использованием слова в документе, и произведение Адамара может быть выбрано как лучший способ определения тематического вектора для «нового слова».

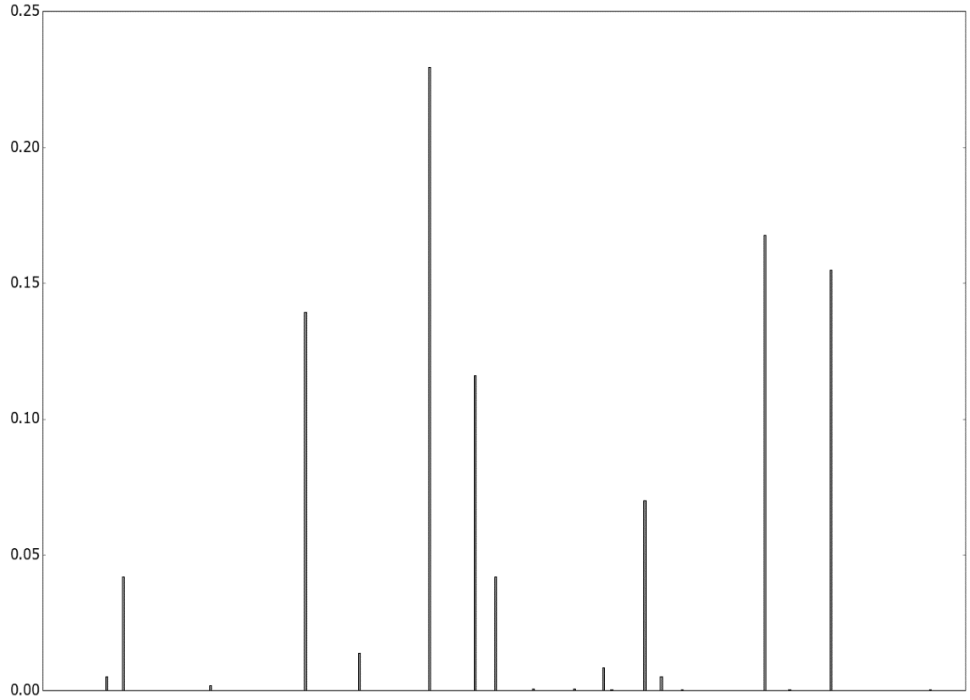


Рисунок 38 – График распределения произведения Адамара вероятностей тем для слова «вальц»

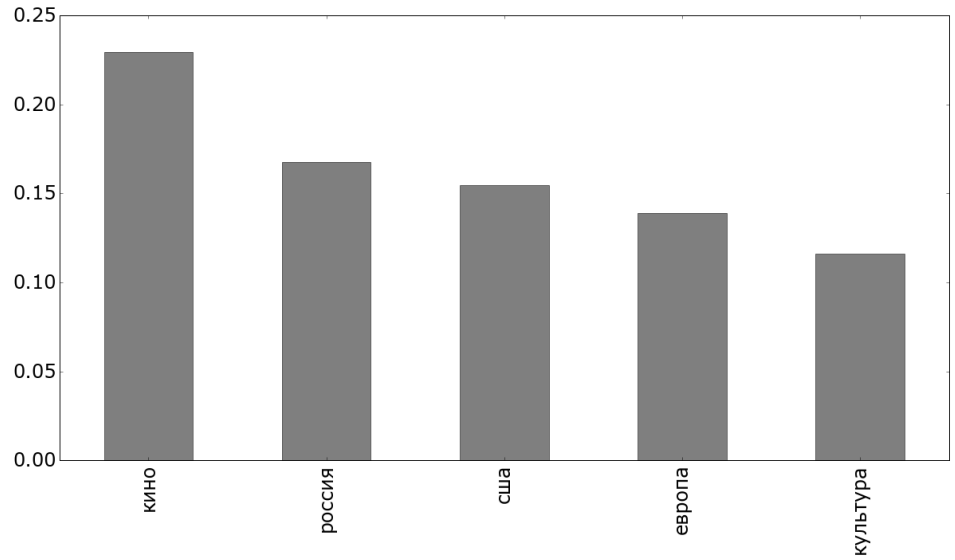


Рисунок 39 – Значимые темы от произведения Адамара вероятностей тем для слова «вальц»

Микросервис визуализации результатов вероятностного тематического моделирования

Микросервис предназначен для представления результатов тематического моделирования пользователю программного комплекса. При создании ВТМ пользователю необходимо оценить качество построенной модели, получить представление о распределении слов и тем, выявить ошибки. Востребованы системы для построения ВТМ и визуализации результатов тематического моделирования, позволяющие управлять параметрами тематической модели, перестраивать ВТМ, отображать в интуитивно понятном виде распределения слов и тем. Удобный интерфейс для ознакомления с результатами тематического моделирования позволит пользователю быстрее получить представление о качестве построенной ВТМ, провести анализ коллекции текстов и принять решение.

На рисунке 40 представлена архитектура системы визуализации ВТМ с помощью iPython. Источником данных является корпус текстов SCTM-ru, микросервис построения ВТМ с помощью библиотеки genism создает ВТМ, модуль расчет данных для представления осуществляет предварительный расчет данных для

визуализации ТМ на временном ряду, модуль представление результата ВТМ отвечает за представление результата вероятностного тематического моделирования.

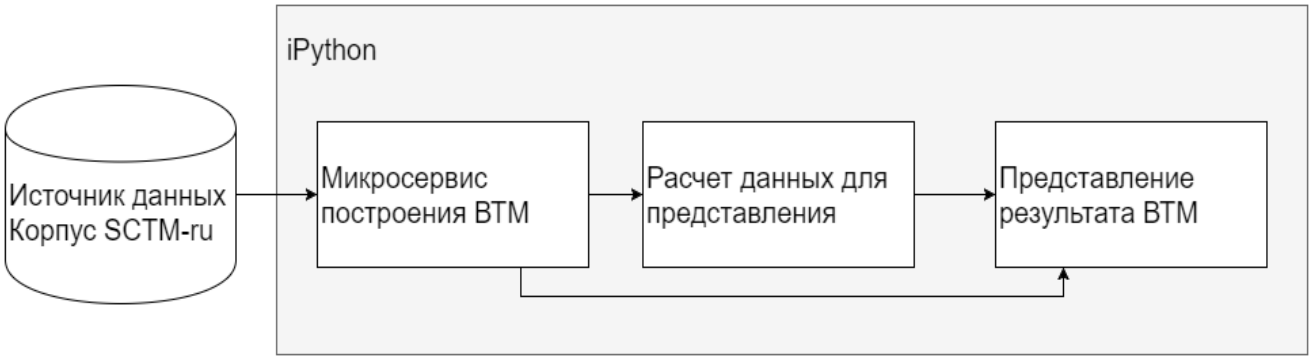


Рисунок 40 – Архитектура системы визуализации ВТМ

В качестве примера визуализации построим ВТМ на специальном корпусе текстов SCTM-ru. Первая версия корпуса была создана на новостях Википедии с 2005 по 2014 год. Используем вторую версию корпуса, в который добавлены новости до 2017 года. Корпус SCTM-ru 2.0 содержит оригинальный текст новости, текст, состоящий только из слов на кириллице и латинице без числительных и других символов, текст, состоящий только из имен существительных в нормальной словоформе, 10748 документов содержат информацию о дате описанных в новостях событий. Для ТМ используем имена существительные тех документов, у которых есть время описанных событий.

Стандартный вывод ТМ — это список наиболее вероятных слов для каждой темы с численной оценкой вероятности, пример представлен на рисунке 41. Такое представление нельзя назвать интуитивно понятным. С помощью библиотеки wordcloud визуализируем полученные темы в более удобном виде. Результат представлен на рисунке 42.

Каждому слову в тематике сопоставлена вероятность, размер шрифта в визуализации соответствует значимости слова для темы: чем выше вероятность, тем

больше размер шрифта у отображаемого слова. Для наглядности представления шрифты слов отличаются по цвету. Метод удобен для оценки промежуточных результатов и представления финальных результатов тематического моделирования.

```
for topic in lda.show_topics(7, 5):
    print (topic[0], topic[1])
```

```
0 0.017*"мир" + 0.012*"год" + 0.012*"игра" + 0.011*"россия" + 0.009*"участник"
1 0.015*"год" + 0.007*"новость" + 0.005*"жирон" + 0.005*"сочи" + 0.005*"викиучебник"
2 0.011*"президент" + 0.010*"россия" + 0.010*"сша" + 0.007*"страна" + 0.007*"человек"
3 0.019*"россия" + 0.013*"украина" + 0.013*"год" + 0.011*"путин" + 0.009*"человек"
4 0.012*"год" + 0.012*"компания" + 0.006*"человек" + 0.006*"пользователь" + 0.006*"сайт"
5 0.018*"новость" + 0.015*"сша" + 0.006*"клинтон" + 0.006*"человек" + 0.005*"время"
6 0.014*"год" + 0.011*"сентябрь" + 0.009*"россиипреступность" + 0.007*"барселона" + 0.006*"август"
```

Рисунок 41 – Вывод списка тем и слов построенной BTM в среде Anaconda

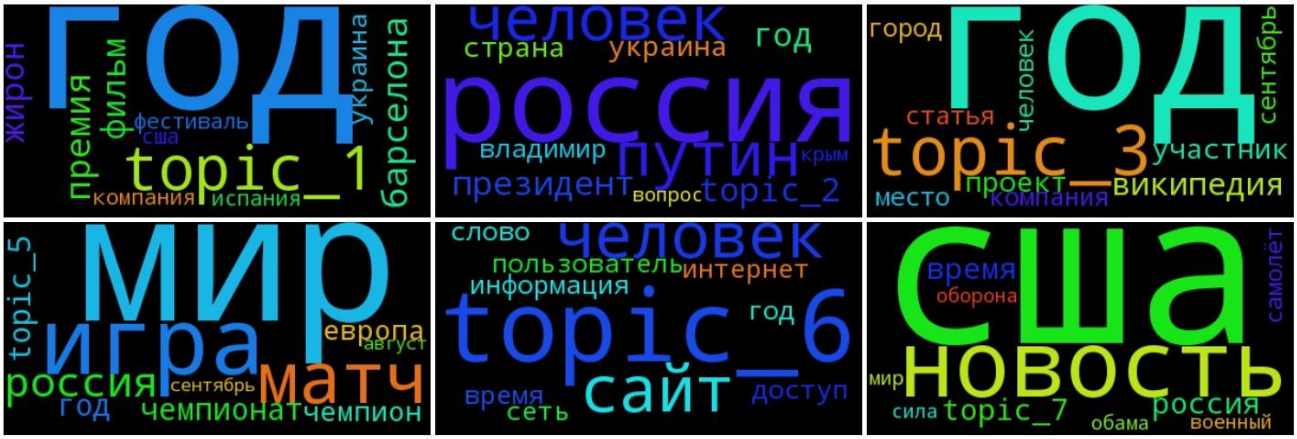


Рисунок 42 – Визуализация тематической модели wordcloud

На рисунке 43 представлен интерактивный высокоуровневый интерфейс, созданный с помощью библиотеки pyLDAvis. В левой части с помощью диаграммы Элерра-Венна представлены темы. Представление позволяет оценить, насколько близки темы. Те, что расположены рядом, находятся ближе, те, в которых общие слова встречаются реже, находятся дальше друг от друга. С правой стороны представлен набор слов, наиболее характерных для выбранной темы. Такое представление

позволяет не только увидеть значимость слова для выбранной темы, но и отражает, насколько слово часто встречается в коллекции. Интерфейс интерактивный. Специалист может выбрать интересующую его тему, для этого достаточно кликнуть на кружке отображения темы либо использовать фильтр. Выбранная тема и слово выделяются цветом. Если кликом выбрать интересующее слово, то представление тем слева обновится. на рисунке 44 продемонстрировано отображение для слова «путин».

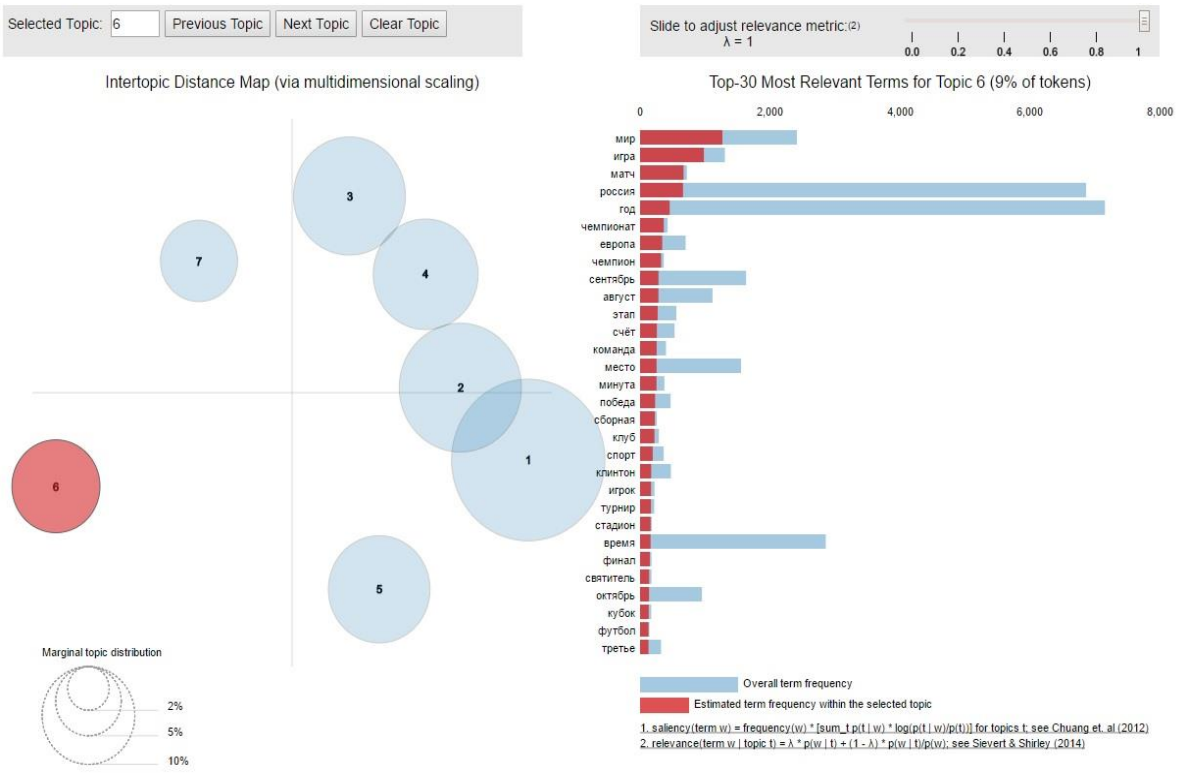


Рисунок 43 – Высокоуровневый интерфейс, построенный с библиотекой pyLDAvis

В практических задачах часто приходится иметь дело с анализом потока текстовых документов, например, потока новостей. В данной работе под текстовым потоком понимается последовательность текстовых документов с определенным для каждого документа временем создания. Под обработкой потока текстовых документов понимается комплексная задача кластеризации поступающих документов и анализа тематических характеристик этих документов. ТМ, анализирующие поток текстовых документов, называют темпоральными ТМ. Чтобы визуализировать

динамику изменения тем во времени, необходимо провести предварительный расчет данных TM, псевдокод в листинге 4.

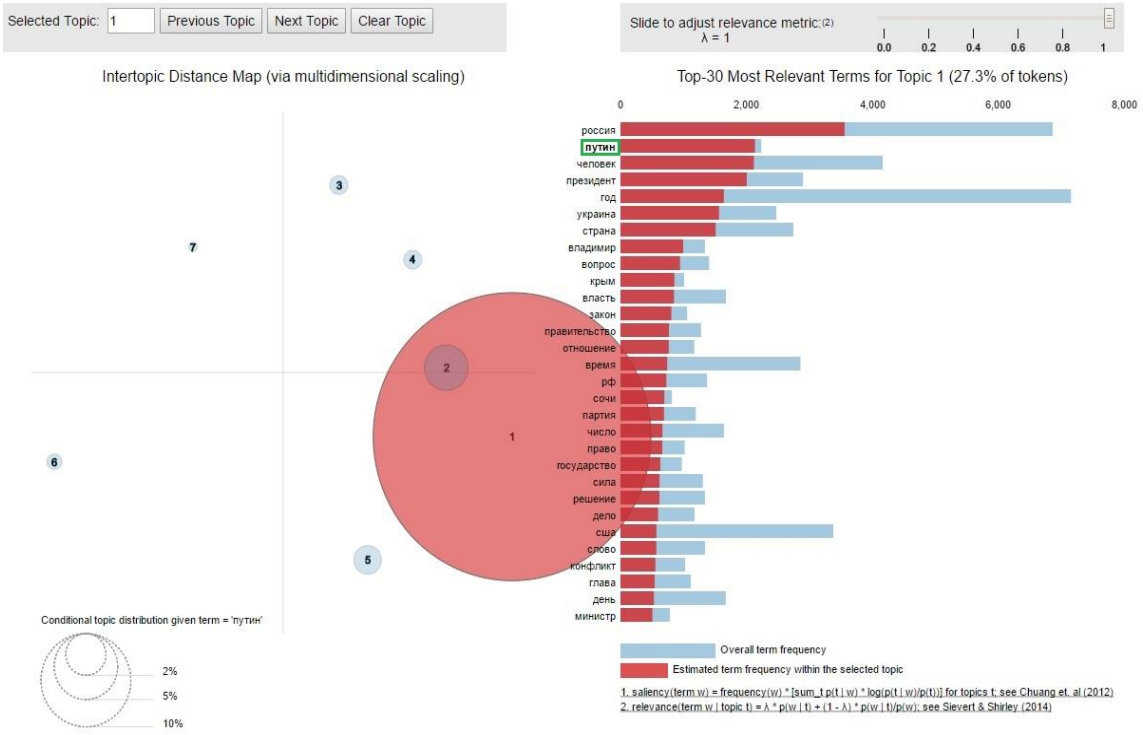


Рисунок 44 – Интерактивность при выборе слова, для высокоуровневого интерфейса

Листинг 4: Подготовка данных для темпоральной визуализации

Вход: D – коллекция документов, где у каждого документа есть дата описанных событий $t(d), z$ – темы документов,

Выход: рассчитанные, нормированные и ненормированные значения вероятностей тем на дату.

1. Для всех $d \in D$.
2. Для всех $z(d)$, $p(z,t) = \sum_{d \in D} p(z|d)$.
3. Считаем матрицу тема-дата, значение которой – это сумма вероятностей темы за дату.
4. Нормируем значения суммы вероятностей тем на каждую дату, сумма всех вероятностей тем за день равна единице.

Визуализация результатов тематического моделирования на временной шкале возможна несколькими методами. На рисунке 45 представление с помощью простой диаграммы-линии, реализованное библиотекой matplotlib. Каждая линия, отражающая тему, окрашена в свой цвет. Представление позволяет увидеть всплески популярности отдельных тем в потоке новостей. На представленной диаграмме заметно появление нескольких популярных тем в 2014 и 2016 годах.

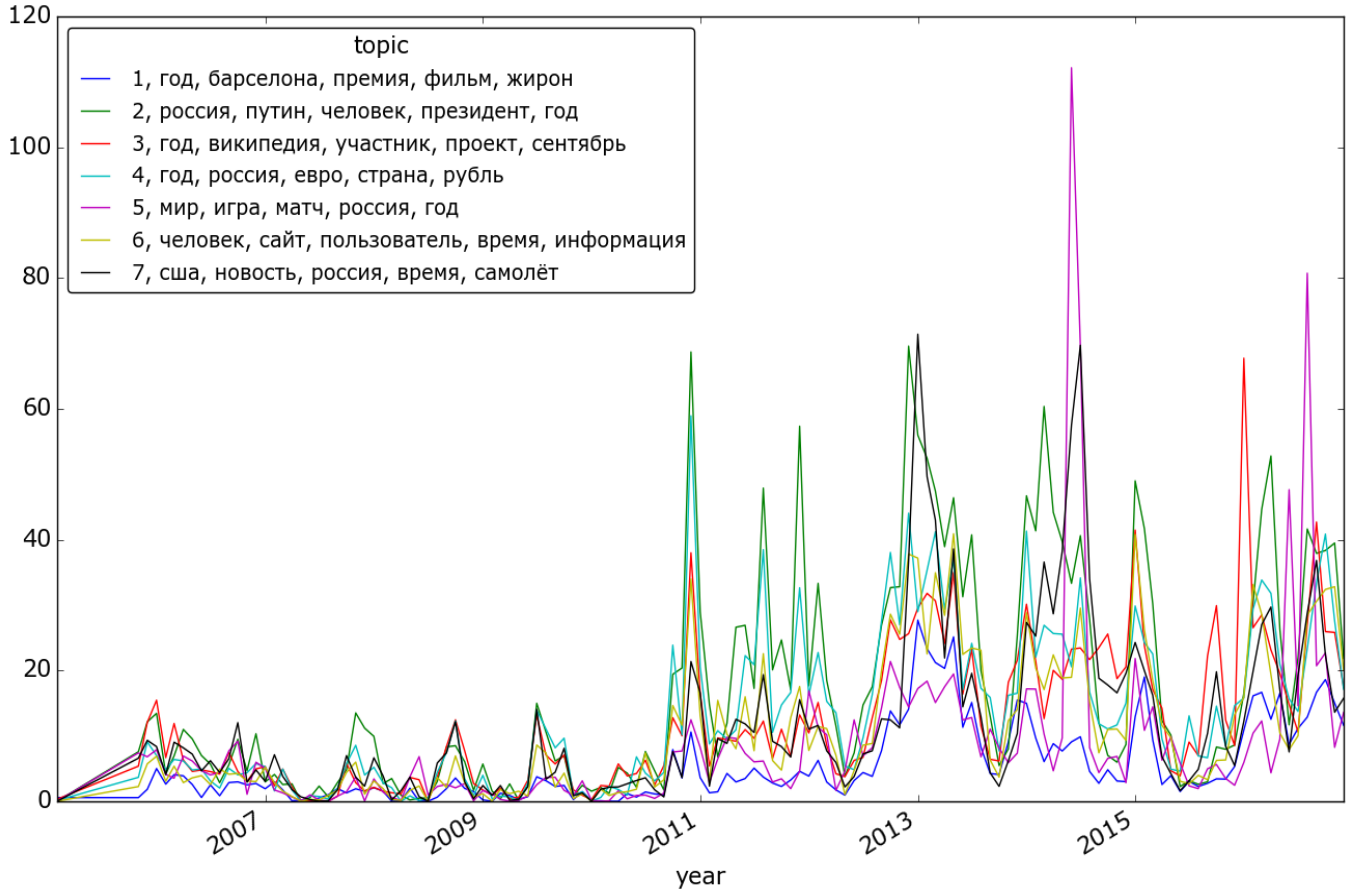


Рисунок 45 – Представление результатов тематического моделирования во времени

Второй метод представления результатов тематического моделирования во времени – это использование диаграммы-области. Метод реализован библиотекой matplotlib, результат представлен на рисунке 46. У каждой темы свой цвет на диаграмме. На графике выведены все темы, значение по шкале ординат – это сумма вероятностей темы на конкретную дату. Отчетливо видны всплески по отдельным

темам, прошедшие в 2014 и 2016 годах. Заметно увеличение количества новостей с 2011 года. Такое представление позволяет проследить изменение популярности каждой темы во времени и увидеть общий рост количества документов.

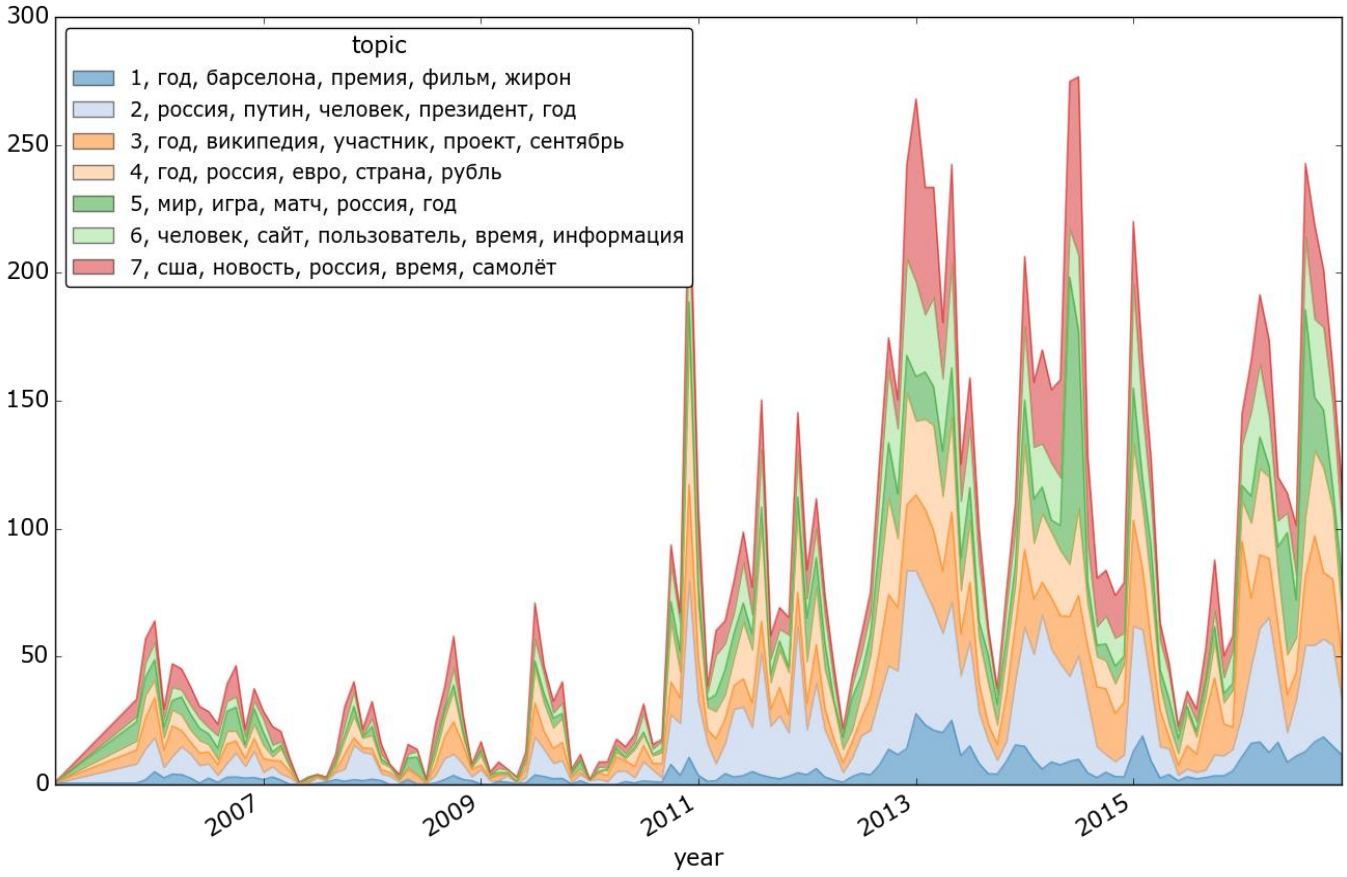


Рисунок 46 – Ненормализованное представление тематического моделирования во времени

На рисунке 47 представлен третий метод визуализации ТМ во времени, диаграмма-области построена на нормированных данных. Так же, как и на предыдущем представлении каждой теме соответствует свой цвет на диаграмме. Представление позволяет проследить повышение популярности или зарождение новой темы и снижение популярности другой. Заметны изменения популярности нескольких тем в 2014-2016 годах. На временной шкале с 2005 года по 2017 тяжело разобраться, какая тема в какой месяц значительно изменилась. Система располагает всем необходимым для того, чтобы сделать срез данных по времени, взяв диапазон с 2014 по конец 2016 года, и визуализировать его. Результат представлен на рисунке 48.

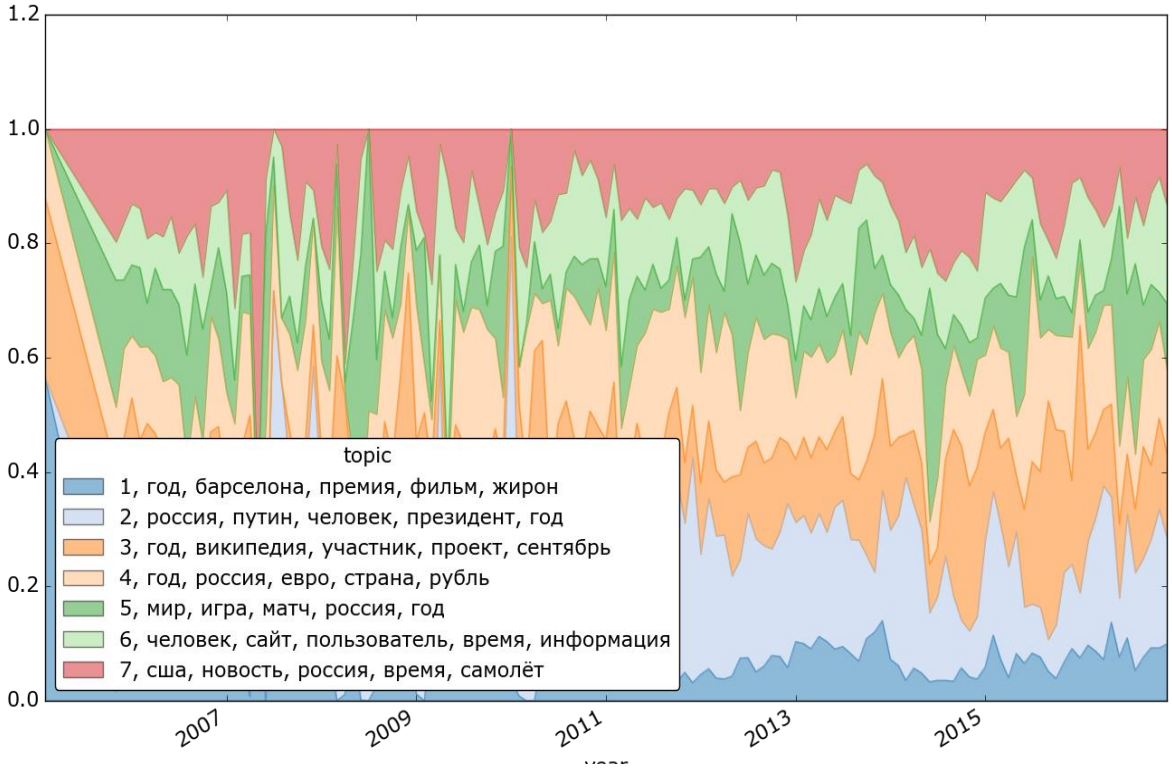


Рисунок 47 – Нормализованное представление тематического моделирования во времени

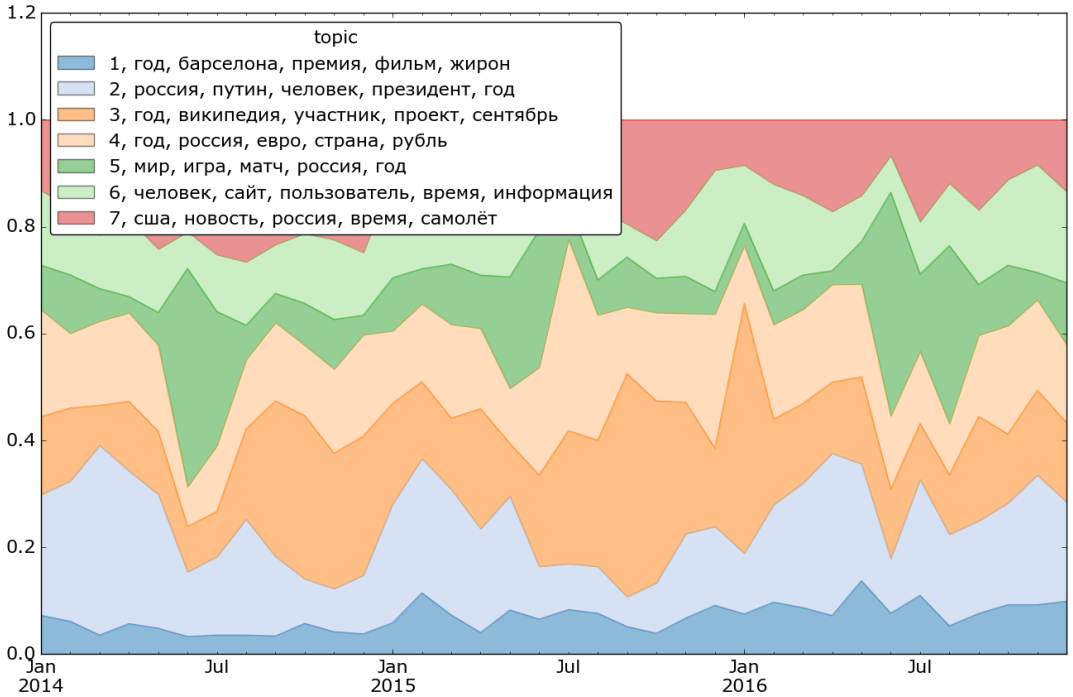


Рисунок 48 – Нормализованное представление тематического моделирования во времени 2014-2016 года

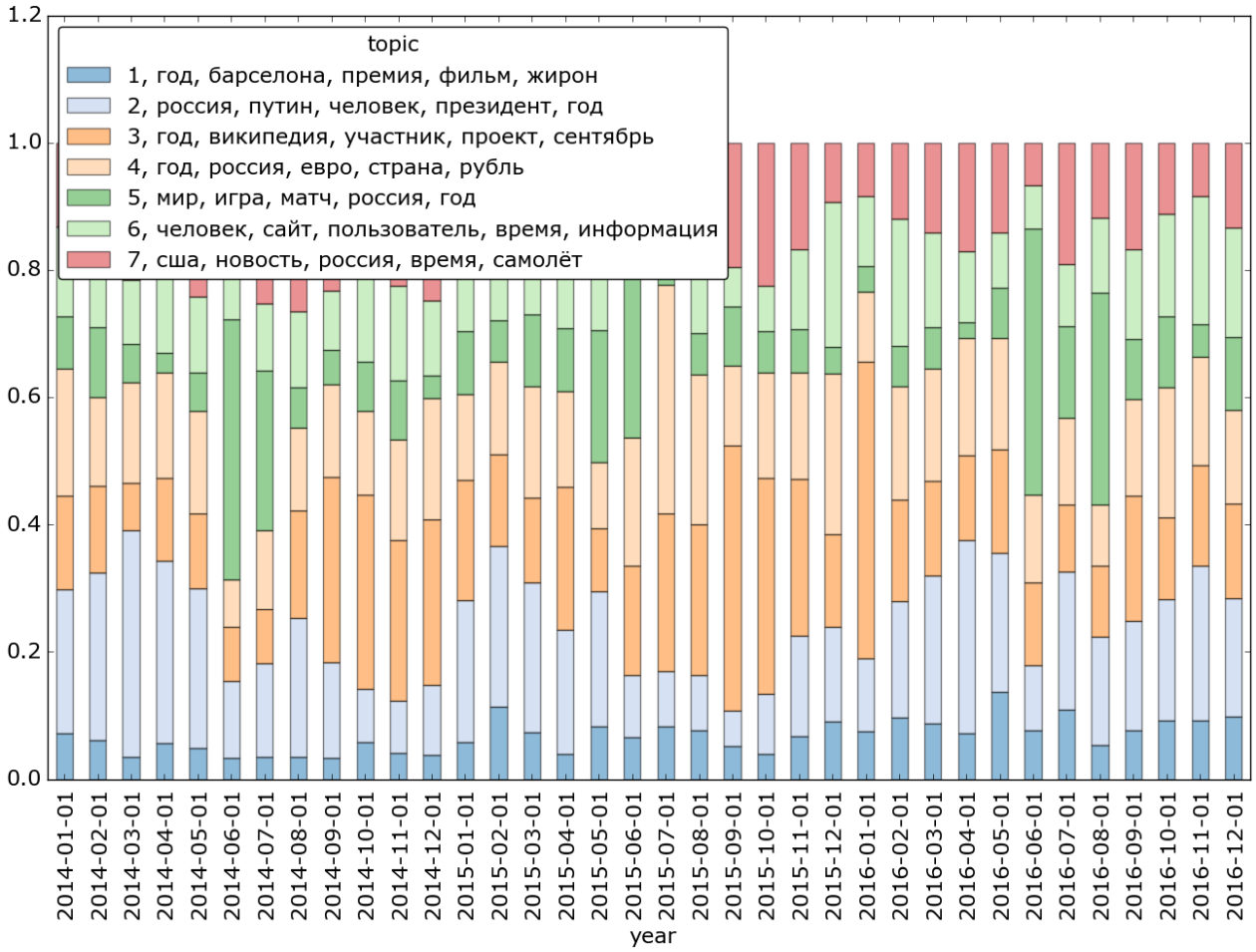


Рисунок 49 – Нормализованное представление тематического моделирования во времени

Четвертый метод реализован с помощью гистограмм. Результат представлен на рисунке 49. Такое представление удобно для отслеживания изменений популярности темы в конкретную дату. На гистограмме отчетливо видно, что тема 1 (зеленый цвет на диаграмме) резко выросла в июне 2016 года, а затем ее популярность стала снижаться.

Особенностями подхода к визуализации являются гибкие возможности по настройке и управлению процесса создания BTM, выбор источника данных и использование подходящих библиотек для визуализации. Пользователь, работающий с системой, может проанализировать текущий результат, вернутся к параметрам BTM, изменить их и перестроить модель заново. При анализе изменения популярности тем

во времени важно наличие интуитивно понятных диаграмм, позволяющих не только специалисту по анализу текстовых данных получить представление об особенностях коллекции текстов, но и обычному пользователю интуитивно понять смысл представленных на диаграмме данных. Наличие высокоуровневого интерфейса позволяет специалисту взаимодействовать с результатами тематического моделирования, в дружелюбном для пользователя виде представлять BTM. Существуют другие подходы к визуализации результатов вероятностного тематического моделирования, которые рассмотрены в статье [95].

4.3. Применение программного комплекса в практических задачах

Разработанный программный комплекс на базе микросервисной архитектуры позволяет использовать его для решения практических задач в автоматизированном и автоматическом режиме. Для анализа коллекции или потока текстовых документов, пользователь программного комплекса настраивает параметры построения BTM, в соответствии с поставленными целями последовательно задействует необходимые микросервисы, визуализирует или экспортирует результат вероятностного тематического моделирования. Использование программного комплекса в автоматическом режиме требует настройки выполнения микросервисов в соответствии с необходимым сценарием обработки текстовых данных.

Методы вероятностного тематического моделирования применяются для решения задач лексической многозначности (дизамбигуации) наряду с другими алгоритмами дистрибутивного анализа, такими как word2vec [66, 78], sense2vec [98], ada-gram [45].

В [28] рассмотрены основные задачи информационного поиска. Важной подзадачей информационного поиска является автоматическая классификация поисковых запросов для повышения релевантности информационного поиска. На

рисунке 50 представлен сценарий применения BTM для решения подзадачи классификации в информационном поиске.

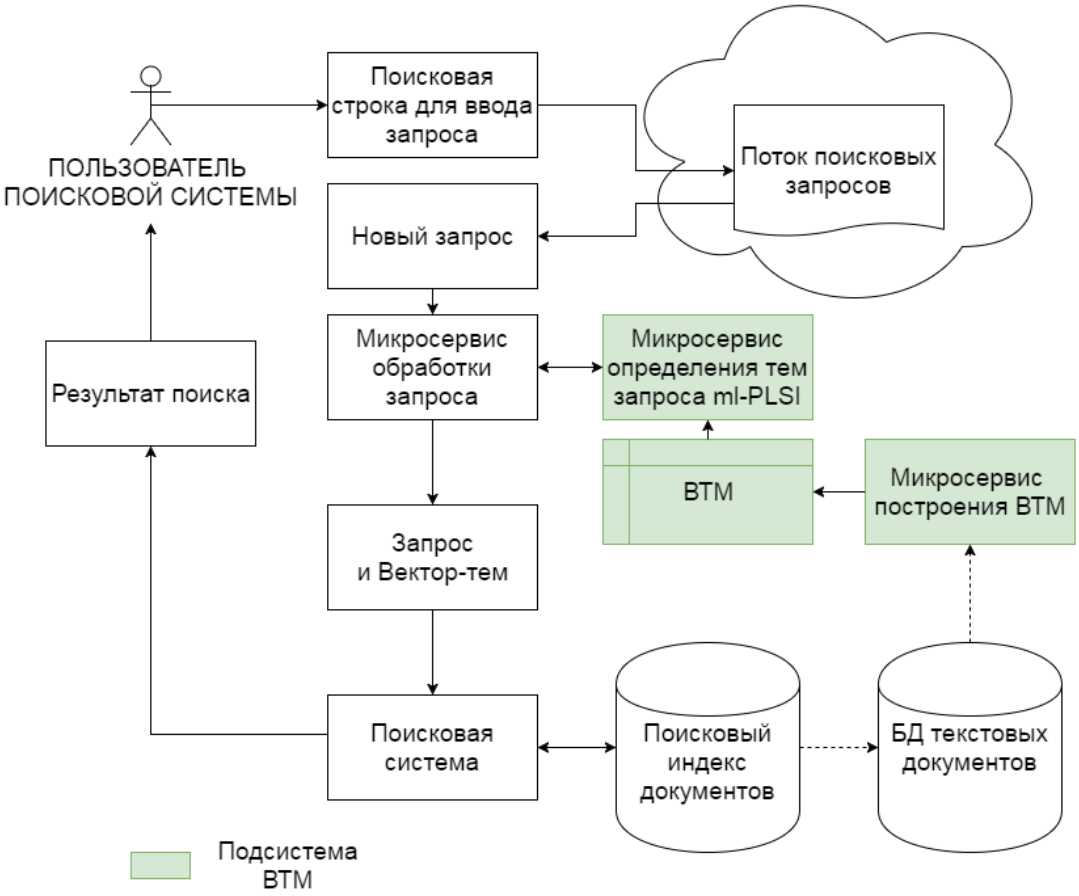


Рисунок 50 – Сценарий применения BTM для решения задач информационного поиска

BTM применяется для одновременной кластеризации коллекции текстовых документов (поискового индекса) и запросов пользователей поисковой системы. Сначала на коллекции текстовых документов строится BTM с помощью микросервиса для построения BTM, формально эта модель характеризует набор тем предметной области поиска. Пользователи поисковой системы задают запросы, микросервис обработки запроса обращается к микросервису определения тем запроса. Получив вектор тем, в поисковую систему отправляется запрос с характерным набором тем документов, соответствующих запросу. Поисковая система, опираясь на вектор тем,

формирует ответ из наиболее вероятных документов, которые также наиболее вероятны теме запроса, что в результате повышает релевантность результатов поиска.

Второй сценарий автоматического использования программного комплекса представлен на рисунке 51, он относится к использованию BTM в рекомендательном сервисе. Существуют подходы к использованию алгоритмов BTM в рекомендательных сервисах, один из них представлен в работе [76]. Преимущество предложенной схемы заключается в использовании алгоритма дополнения BTM новыми словами вместо перестроения при появлении новых документов в базе данных. На базе потока текстовых документов, это могут быть страницы веб-сайтов сети интернет или страницы одного сайта, создается БД текстовых документов. Когда в базе данных накопилось достаточное количество документов, включается микросервис построения BTM. Для каждого документа, который посетил пользователь, определяется список тем, к которым относится посещенный документ. Микросервис извлечения документов по темам запрашивает список тем в построенной BTM, обращается к хранилищу документов с целью извлечь документы в соответствии с полученным списком тем и формирует персональный список рекомендуемых к посещению документов для пользователя. Построенная на начальном наборе документов BTM может быть перестроена по мере необходимости. Поток документов обрабатывается с помощью микросервиса определения тем новых документов и микросервиса пополнения BTM.

Исследования, отраженные в диссертации, проведены в рамках НИР № 714630 «Интеллектуальные технологии для социо-киберфизических систем», проводимой в Университете ИТМО (государственная программа поддержки ведущих университетов РФ, субсидия 074-U01). Результаты, полученные в ходе исследования, применяются в системе анализа новостного потока принятой к использованию в ООО «Олимп» (Правительство Москвы) и в сервисе многозначной классификации поисковых запросов пользователей, принятом к использованию в ООО

«Rambler&Co». Результаты диссертационной работы использованы в учебном процессе кафедры информационных систем федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики». Основные понятия, определения и результаты диссертационного исследования используются при изучении дисциплины «Управление знаниями» магистерской программы по специальности 38.04.05. — Информационные системы бизнеса.

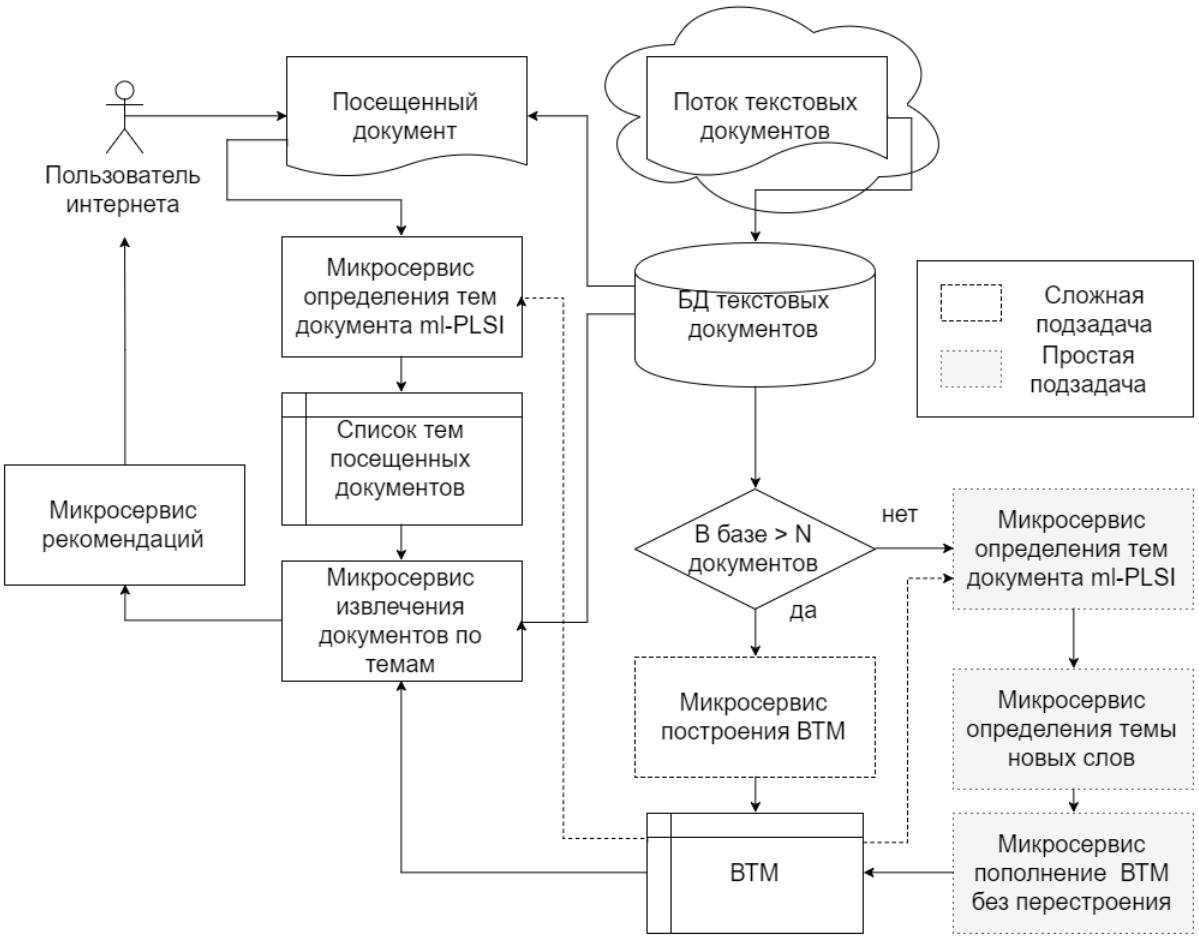


Рисунок 51 – Сценарий применения BTM в сервисе рекомендаций

По разработанной системе было получено свидетельство о регистрации программы для ЭВМ «Система для анализа текстовых документов с использованием вероятностного тематического моделирования // Карпович С.Н.» №2017615118 от 3 мая 2017.

Выводы к главе 4

- Глава посвящена проектированию программного комплекса для вероятностного тематического моделирования. Предложен подход на базе микросервисной архитектуры. Рассмотрены основные принципы микросервисов. Выделены преимущества, характерные для архитектуры микросервисов.
- Разработана архитектура системы вероятностного тематического моделирования. Преимущества микросервисной архитектуры позволяют создать программный комплекс, удовлетворяющий требованиям, определенным во 2 главе, позволяют создать приложения, соответствующие человеко-ориентированному, событийно-ориентированному и определяемому-данным подходам.
- Описан процесс разработки микросервисов программного комплекса. С использованием микросервиса построения корпуса текстов для BTM создан корпус SCTM-ru. Микросервис построения BTM позволяет использовать классические алгоритмы для построения BTM, а также позволяет строить BTM методом обучения на размеченных данных. Описаны микросервисы для многозначной классификации текстовых документов и определения темы «нового слова», необходимые для обработки потока текстовых документов и построения динамических тематических моделей.
- Уделено внимание микросервису визуализации результатов вероятностного тематического моделирования. Представлены методы визуализации статичных

коллекций и потоков текстовых документов. Представлен способ темпорального представления ВТМ.

- Описаны сценарии применения программного комплекса для решения практических задач. Определена возможность автоматического и автоматизированного использования отдельных микросервисов. Рассмотрены задачи анализа коллекций и потоков текстовых документов, задачи информационного поиска и применение ВТМ в рекомендательных системах.
- Разработанный прототип программного комплекса полностью удовлетворяет требованиям, обозначенным во 2 главе. Позволяет создавать и использовать в практических задачах вероятностные тематические модели. Предоставляет специалисту средства гибкой настройки параметров используемых алгоритмов в процессе работы, обладает удобными средствами визуализации и экспорта промежуточных и окончательных результатов вероятностного тематического моделирования.

Заключение

В диссертационной работе предложено решение актуальной научно-технической задачи по разработке комплекса математических и программных средств интеллектуального анализа потока текстовых документов с использованием вероятностного тематического моделирования, основанного на микросервисной архитектуре и позволяющего обеспечить специалиста необходимыми средствами анализа, возможностью выбирать источники данных, задавать параметры вероятностного тематического моделирования. В процессе решения данной задачи были получены следующие результаты:

1. Создан русскоязычный корпус текстов SCTM-ru, позволяющий исследовать алгоритмы вероятностного тематического моделирования и отличающийся от других корпусов наличием оригинального текста документа и метатекстовой разметки: автор, время описанных событий, тема. Текст и метатекстовая разметка необходимы для построения BTM различных видов. Источником данных корпуса является сайт «Русские Викиновости».
2. Разработан метод расчета матриц BTM на основе обучения с учителем (авторами документов), учитывающий заданные связи между документами и темами, позволяющий упростить построение BTM за счет отсутствия итераций.
3. Разработан алгоритм многозначной классификации текстовых документов ml-PLSI, основанный на вероятностном тематическом моделировании, заключающийся в использовании матрицы «слово-тема» BTM для классификации документов, что позволяет определять темы «новых документов» при анализе потока текстовых документов в динамической тематической модели.
4. Разработан метод определения тем «нового слова», основанный на использовании произведения Адамара тематических векторов документов, содержащих это слово, позволяющий определять вектора тем для «новых слов»

в потоке текстовых документов при построении динамической тематической модели с эффективностью, превосходящей существующие аналоги.

5. Разработан прототип комплекса программных средств для анализа потока текстовых документов с использованием вероятностного тематического моделирования, отличающийся использованием микросервисной архитектуры и позволяющий предоставить вариативность выбора подходящих способов решения конкретных практических задач, а также возможность визуализации промежуточных и конечных результатов вероятностного тематического моделирования.

Разработанный прототип комплекса программных средств рекомендован к использованию для построения и визуализации BTM. Дальнейшее исследование вероятностного тематического моделирования потока текстовых документов возможно за счет улучшения функциональности комплекса программных средств. Перспективным направлением исследования является использование BTM в задачах ассоциативной классификации в комбинации с другими классификаторами, а также использование BTM как решателя в алгоритмах коллективного распознавания.

Полученные результаты соответствуют п. 3 «Модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем», п. 4 «Системы управления базами данных и знаний» паспорта специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Литература

1. Баранов А.Н. Введение в прикладную лингвистику / А.Н. Баранов, Москва: Эдиториал УрСС., 2001.

2. Барсегян А. Анализ данных и процессов / А. Барсегян, 3-е изд., Санкт-Петербург: БХВ-Петербург, 2009.

3. Воронцов, К.В. Потапенко А.А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. 2012. № 4 (4). С. 693–706.

4. Воронцов, К.В. Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. 2013.

5. Воронцов К.В. Вероятностное тематическое моделирование // Москва. 2013.

6. Воронцов К.В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. № 3 (455). С. 268–271.

7. Гойвертс, Я. Левитан С. Регулярные выражения. Сборник рецептов / С. Гойвертс, Я. Левитан, Санкт-Петербург: Символ-Плюс, 2013.

8. Городецкий, В. И. Серебряков С.В. Методы и алгоритмы коллективного распознавания // Автоматика и Телемеханика. 2008. № 11. С. 3.

9. Городецкий, В. И. Тушканова О.Н. Ассоциативная классификация: аналитический обзор. Часть 1. // Труды СПИИРАН. 2015. № 38 (1). С. 183–203.

10. Городецкий, В. И. Тушканова О.Н. Ассоциативная классификация: аналитический обзор. Часть 2. // Труды СПИИРАН. 2015. № 39 (2). С. 212–240.

11. Доусон М. Программируем на Python / М. Доусон, Санкт-Петербург: Питер, 2014.

12. Дударенко М.А. Регуляризация многоязычных тематических моделей // Вычислительные методы и программирование. 2015. № 1 (16). С. 26–38.

13. Журавлёв, Ю.И. Зенкин, А.А. Зенкин, А.И. Исаев, И.В. Кольцов, П.П. Кочетков, Д.В. Рязанов В.В. Задачи распознавания и классификации со стандартной обучающей информацией // Вычислительная математика и математическая физика. 1980. № 5 (20).

С. 1294–1309.

14. Захаров, В.П. Азарова И.В. Параметризация специальных корпусов текстов. // Структурная и прикладная лингвистика: Межвузовский сборник. 2012. (9). С. 176–184.

15. Захаров В.П. Международные стандарты в области корпусной лингвистики // Структурная и прикладная лингвистика. 2012. (9). С. 201–221.

16. Ингерсолл, Г. Мортон, Т. Фэррис Э. Обработка неструктурированных текстов. Поиск, организация и манипулирование. / Э. Ингерсолл, Г. Мортон, Т. Фэррис, под ред. А.А. Перевод с английского Слинкин, Москва: ДМК Пресс., 2015.

17. Карпович С.Н. Русскоязычный корпус текстов SCTM-RU для построения тематических моделей // Труды СПИИРАН. 2015. № 39 (2). С. 123.

18. Карпович С.Н. Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI // Труды СПИИРАН. 2016. № 47 (4). С. 92–104.

19. Карпович С.Н. Тематическая модель с бесконечным словарем // Information & Control Systems/Informazionno-Upravlyaushie Sistemy. 2016. № 6 (85).

20. Кормен, Т. Лейзерсон, Ч. Ривест, Р. Штайн К. Алгоритмы: построение и анализ. / К. Кормен, Т. Лейзерсон, Ч. Ривест, Р. Штайн, 2-е изд., Москва: Вильямс, 2005. 1296 с.

21. Коршунов, А. Гомзин А. Тематическое моделирование текстов на естественном языке. // Труды ИСП РАН. 2012.

22. Крижановский, А.А. Смирнов А.В. Подход к автоматизированному построению общецелевой лексической онтологии на основе данных викисловаря. // Известия РАН. Теория и системы управления. 2013. (2). С. 53–63.

23. Ландэ Д.В. Основы моделирования и оценки электронных информационных потоков. // Инжиниринг. 2006.

24. Лапшин В.А. Онтологии в компьютерных системах. Роль онтологий в

современной компьютерной науке. // RSDN MAGAZINE. 2009. (4). С. 61–67.

25. Лукашевич Н.В. Тезаурусы в задачах информационного поиска / Н.В. Лукашевич, Москва: Издательство МГУ, 2011.

26. Лутц М. Программирование на Python. / М. Лутц, Москва: Символ-Плюс, 2011.

27. Маккинли У. Python и анализ данных. / У. Маккинли, Москва: ДМК Пресс., 2015.

28. Маннинг, К. Рагхаван, П. Шютце Х. Введение в информационный поиск. / Х. Маннинг, К. Рагхаван, П. Шютце, Москва: Вильямс, 2011.

29. Марманис, Х. Бабенко Д. Алгоритмы интеллектуального интернета. / Д. Марманис, Х. Бабенко, Санкт-Петербург: Символ-Плюс, 2011. Передовые методики сбора, анализа и обработки данн с.

30. Мацяшек Л.А. Анализ и проектирование информационных систем с помощью UML 2.0. / Л.А. Мацяшек, Москва: Вильямс, 2008. 816 с.

31. Николаев, И.С. Митренина, О.В. Ландо Т.М. Прикладная и компьютерная лингвистика. / Т.М. Николаев, И.С. Митренина, О.В. Ландо, 2-е изд., Москва: URSS, 2017.

32. Нокель, М.А. Лукашевич Н.В. Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами . // Вычислительные методы и программирование2. 2015. № 2 (16). С. 215–234.

33. Нокель М.А. Методы улучшения вероятностных тематических моделей текстовых коллекций на основе лексико-терминологической информации 2015.

34. Ньюмен С. Создание микросервисов. / С. Ньюмен, Санкт-Петербург: Питер, 2016.

35. Омельченко В.В. Общая теория классификации. / В.В. Омельченко, Москва: ИПЦ «Маска», 2008.

36. Пфеффер А. Вероятностное программирование на практике / А. Пфеффер, Москва: Manning Publications, 2017.

37. Розенфельд, Л. Морвиль П. Информационная архитектура / П. Розенфельд, Л. Морвиль, Санкт-Петербург: Символ-Плюс, 2005.

38. Сазерленд Д. Scrum: Революционный метод управления проектами / Д. Сазерленд, Москва: Манн, Иванов и Фербер, 2015.

39. Смирнов, А.В. Круглов, В.М. Крижановский, А.А. Луговая, Н.Б. Карпов, А.А. Кипяткова И.С. Количественный анализ лексики русского WordNet и викисловарей // Труды СПИИРАН. 2012. (23). С. 231–253.

40. Торре, С. Сингх, К.Д. Туречек В. Microsoft Azure - Azure Service Fabric и архитектура микросервисов [Электронный ресурс]. URL: https://msdn.microsoft.com/ru-ru/magazine/mt595752.aspx (дата обращения: 01.07.2017).

41. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах, Москва: Litres, 2017.

42. Фридл Д. Регулярные выражения / Д. Фридл, под ред. А. Переводчики Матвеев, Е. Киселев, Санкт-Петербург: Символ-Плюс, 2008. 608 с.

43. Aggarwal C.C. Data streams: models and algorithms // Springer Science & Business Media. 2007. (31).

44. Asuncion, A. Welling, M. Smyth, P. Teh Y.W. On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artifiial Intelligence. 2009. С. 27–34.

45. Bartunov, S. Kondrashkin, D. Osokin, A. Vetrov D. Breaking sticks and ambiguities with adaptive skip-gram // Artificial Intelligence and Statistics. 2016. С. 130–138.

46. Beck K. Manifesto for agile software development 2001.

47. Blei, D. Chaney A. Visualization Topic Models // ICWSM. 2012.

48. Blei, D. McAuliffe J. Supervised topic models // Advances in neural information processing systems 20. 2008. С. 121–128.

49. Blei, D.M. Lafferty J.D. Dynamic topic models // Proceedings of the 23rd international conference on Machine learning. – ACM. 2006. С. 113–120.

50. Blei, D.M. Moreno P.J. Topic segmentation with an aspect hidden Markov model // Proceedings of the 24th annual international ACM SIGIR conference on Research and

development in information retrieval. – ACM. 2001. С. 343–348.

51. Blei, D.M. Ng, A.Y. Jordan M.I. Latent Dirichlet Allocation // Journal of machine Learning research. 2003. (3). С. 993–1022.

52. Blei D.M. Probabilistic topic models // Communications of the ACM. 2012. № 4 (55). С. 77–84.

53. Blevins C. Topic modeling Martha Ballard's diary [Электронный ресурс]. URL: http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/ (дата обращения: 01.07.2017).

54. Boisvert, R.F. Pozo, R. Remington K.A. The matrix market exchange formats: Initial design // National Institute of Standards and Technology Internal Report, NISTIR. 1996. (5935).

55. Boyd-Graber, J.L. Blei, D.M. Zhu X.A. Topic Model for Word Sense Disambiguation // EMNLP-CoNLL. 2007. С. 1024–1033.

56. Cearley, D.W. Walker M.J. Top 10 Strategic Technology Trends for 2017 2016.

57. Chang, J. Boyd-Grabber, J. Wang, C. Gerrich, S. Blei D. Reading tea leaves: How human interpret topic models // Proceedings of the 24th Annual Conference on Neural Information Processing Systems. 2009. С. 288–296.

58. Chuang, J. Manning, C.D. Heer J. Termite: Visualization techniques for assessing textual topic models // Proceedings of the International Working Conference on Advanced Visual Interfaces. – ACM. 2012. С. 74–77.

59. Daud A. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of computer science in China. 2010. № 2 (4). С. 280–301.

60. Dempster, A.P. Laird, N.M. Rubin D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm // Journal of the Royal Statistical Society. 1977. № 1 (39). С. 1–38.

61. Ding, C. Li, T. Peng W. On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing // Computational Statistics and Data Analysis. 2008. (52). С. 3913–3927.

62. Dragoni N. [и др.]. Microservices: yesterday, today, and tomorrow // preprint arXiv:1606.04036. 2016.

63. Feldman, R. Sanger J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. / J. Feldman, R. Sanger, Cambridge University Press, 2006. 422 c.

64. Fowler, M. Lewis J. Microservices a definition of this new architectural ter [Электронный ресурс]. URL: http://martinfowler.com/articles/microservices.html (дата обращения: 01.07.2017).

65. Ganesan A. [и др.]. LDAExplore: Visualizing Topic Models Generated Using Latent Dirichlet Allocation // preprint arXiv:1507.06593. 2015. 2015.

66. Goldberg Y., Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method // preprint arXiv:1402.3722. 2014.

67. Griffiths, T.L. Steyvers M. Finding scientific topics // Proceedings of the National academy of Sciences. 2004. № 1 (101). С. 5228–5235.

68. Griffiths T.L. Integrating topics and syntax // Advances in neural information processing systems. 2005. С. 537–544.

69. Gruber, A. Rosen-Zvi, M. Weiss Y. Hidden Topic Markov Models // Proceedings of Artificial Intelligence and Statistics (AISTATS). 2007. (2). С. 163–170.

70. Hoffman, T. Blei, D. Bach F. Online learning for latent Dirichlet allocation // Neural Information Processing Systems. 2010.

71. Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. – ACM. 1999. С. 50–57.

72. Horn R.A. The Hadamard Product // Proc. Symp. Appl. Math. 1990. (40). С. 87–169.

73. Joachims T. SVM-Light Support Vector Machine // University of Dortmund. 1999. № 4 (19).

74. Lau, J.H. Collier, N. Baldwin T. On-line Trend Analysis with Topic Models:\# twitter Trends Detection Topic Model Online // COLING. 2012. С. 1519–1534.

75. Li, W. Wang, X. McCallum A. A continuous-time model of topic co-occurrence trends // Event Extraction and Synthesis. 2006. С. 48–53.

76. Liu, D. Chen Y. Biterm-LDA: A Recommendation Model for Latent Friends on Weibo // Journal of Residuals Science & Technology. 2017. № 3 (14).

77. McCallum, A. Corrada-Emmanuel, A. Wang X. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. 2005.

78. Mikolov T. [и др.]. Efficient Estimation of Word Representations in Vector Space 2013.

79. Mimno, D. Wallach, H.M. Naradowsky, J. Smith, D.A. McCallum A. Polylingual topic models // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP'09. 2009. С. 880–889.

80. Mimno, D. Wallach, H.M. Talley, E. Leenders, M. McCallum A. Optimizing semantic coherence in topic models // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — EMNLP '11. 2011. С. 262–272.

81. Mimno D. Computational historiography: Data mining in a century of classics journals // Journal on Computing and Cultural Heritage (JOCCH). 2012. № 1 (5). С. 3.

82. Nallapati R.M. [и др.]. Multiscale topic tomography New York, New York, USA: ACM Press, 2007. 520 с.

83. Newman, D. Bonilla, E.V. Buntine W.L. Improving topic coherence with regularized topic models // Advances in Neural Information Processing Systems 24. 2011. С. 496–504.

84. Newman, D. J. Block S. Probabilistic topic decomposition of an eighteenth-century American newspaper // Journal of the American Society for Information Science and Technology. 2006. № 6 (57). С. 753–767.

85. Newman, D. Lau, J. Grieser, K. Baldwin T. Automatic evaluation of topic coherence // Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010. С. 100–108.

86. Newman, D. Noh, Y. Talley, E. Karimi, S. Baldwin T. Evaluating topic models for digital

libraries // Proceedings of the 10th annual Joint Conference on Digital libraries. — JCDL ’10. 2010. С. 215–224.

87. Ni, X. Sun, J.T. Hu, J. Chen Z. Mining multilingual topics from Wikipedia // Proceedings of the 18th International Conference on World Wide Web, WWW’09. 2009. С. 1155–1156.

88. Padmanabhan D. [и др.]. Topic Model Based Multi-Label Classification from the Crowd 2016.

89. Papadimitriou C.H. Latent semantic indexing: A probabilistic analysis // Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. – ACM. 1998. С. 159–168.

90. Ramage, D. Hall, D. Nallapati, R. Manning C.D. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 Volume 1. — EMNLP ’09. 2009. С. 248–256.

91. Rosen-Zvi M. The author-topic model for authors and documents // Proceedings of the 20th conference on Uncertainty in artificial intelligence. 2004. С. 487–494.

92. Rubin, T.N. Chambers, A. Smyth, P. Steyvers M. Statistical topic models for multilabel document classification // Machine Learning. 2012. № 1–2 (88). С. 157–208.

93. Sasaki, K. Yoshikawa, T. Furuhashi T. Online topic model for Twitter considering dynamics of user interests and topic trends // EMNLP. 2014. С. 1977–1985.

94. Smet, W.D. Moens M.F. Cross-language linking of news stories on the web using interlingual topic modelling // Proceedings of the 2nd ACM Workshop on Social Web Search and Mining, SWSM’09. 2009. С. 5764.

95. Smirnov, A. Karpovich, S. Teslya, N. Grigorev A. Topic Model Visualization With IPython // In Proceedings of the 20th Conference of FRUCT association. 2017. С. 131–137.

96. Teh, Y.W. Jordan, M.I. Beal, M.J. Blei D.M. Hierarchical Dirichlet Processes // Journal of the American Statistical Association. 2006. С. 1566–1581.

97. Templeton C. Topic modeling in the humanities: An overview // Maryland Institute for

Technology in the Humanities Blog. 2011.

98. Trask A., Michalak P., Liu J. sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings 2015.

99. Tsoumakas G., Katakis I. Multi-Label Classification // International Journal of Data Warehousing and Mining. 2007. № 3 (3). С. 1–13.

100. Vorontsov K.V. Additive regularization for topic models of text collections // Doklady Mathematics. – Pleiades Publishing,. 2014. № 3 (89). С. 301–304.

101. Wallach H.M. Topic modeling: beyond bag-of-words // Proceedings of the 23rd international conference on Machine learning. 2006. С. 977–984.

102. Wang, C. Paisley, J. Blei D. Online variational inference for the hierarchical Dirichlet process // Artificial Intelligence and Statistics. 2011.

103. Wang, X. McCallum, A. Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval // Proceedings of the 7th IEEE International Conference on Data Mining. 2007. С. 697–702.

104. Wang, X. McCallum A. Topics over time: a non-Markov continuous-time model of topical trends // Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM. 2006. С. 424–433.

105. Wang C., Blei D.M., Heckerman D. Continuous Time Dynamic Topic Models 2012.

106. Wei, X. Croft W.B. LDA-based document models for ad-hoc retrieval // Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM. 2006. С. 178–185.

107. Xu, W. Liu, X. Gong Y. Document clustering based on non-negative matrix factorization // Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM. 2003. С. 267–273.

108. Xu S. Author-Topic over Time (AToT): a dynamic users' interest model // Mobile, Ubiquitous, and Intelligent Computing. – Springer. 2014. С. 239–245.

109. Yeh, J. Wu M. Recommendation based on latent topics and social network analysis //

Computer Engineering and Applications (ICCEA). 2010. (1). С. 209–213.

110. Zhai, K. Boyd-Graber J.L. Online Latent Dirichlet Allocation with Infinite Vocabulary // ICML. 2013. (28). С. 561–569.

111. Zhang J. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM. 2010. С. 1079–1088.

112. Сайт Machine Learning for Language Toolkit MALLET [Электронный ресурс]. URL: http://mallet.cs.umass.edu/topics.php (дата обращения: 01.07.2017).

113. Сайт библиотеки статистической обработки текстов Gensim [Электронный ресурс]. URL: https://radimrehurek.com/gensim/ (дата обращения: 01.07.2017).

114. Сайт BigARTM [Электронный ресурс]. URL: http://bigartm.org/ (дата обращения: 01.07.2017).

115. Сайт Национального корпуса русского языка НКРЯ [Электронный ресурс]. URL: http://www.ruscorpora.ru/ (дата обращения: 01.07.2017).

116. Сайт программы морфологического анализа текстов на русском языке MyStem [Электронный ресурс]. URL: https://tech.yandex.ru/mystem/ (дата обращения: 01.07.2017).

117. Сайт GibbsLDA++ format [Электронный ресурс]. URL: http://gibbslda.sourceforge.net/ (дата обращения: 01.07.2017).

Приложение А. Акты внедрения

Акционерное общество “Олимп”

121099, Москва, ул. Новый Арбат, д. 36. Телефон: (495) 690-77-22. Факс: (495) 697-28-04. E-mail: olymp@olymp-arbat.ru. www.olymp-arbat.ru

От 19.05.2017 № 10-7-565/17

На № _____

Акт
Об использовании результатов кандидатской диссертационной работы
Карповича Сергея Николаевича

Настоящий акт составлен в том, что результаты диссертационной работы «Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов» использованы при проведении анализа текстов новостей и поисковых запросов.

Использованы следующие результаты работы:

1. Многозначная классификация ml-PLSI
2. Вероятностное тематическое моделирование
3. Создание темпоральной вероятностной тематической модели
4. Представление результатов тематического моделирования.

Создание темпоральной вероятностной тематической модели позволило оценить рост и снижение популярности новостных тем связанных с сезонными и ежегодными мероприятиями, что в свою очередь в два раза сократило время, затрачиваемое на составление плана новостных публикаций, связанных с регулярными событиями. Использование многозначной классификации ml-PLSI для потока поисковых запросов, позволило на 5% повысить точность классификации. На основе приведенных результатов были разработаны:

1. Сервис анализа потока новостей, позволяющий обрабатывать большие объемы данных, автоматически группировать схожие новости, отслеживать динамику изменения популярности тем во времени, визуализировать результаты тематического моделирования для последующего представления этих результатов в отчетах.
2. Сервис классификации поисковых запросов, позволяющий определять наиболее вероятные тематические группы очень коротких текстов, таких как запрос в поисковую систему.

Генеральный директор,
кандидат экономических наук **А.Н.Фомочкин**

рамблер

Общество с ограниченной ответственностью «Рамблер Интернет Холдинг»
Место нахождения: 115280, г. Москва, ул. Ленинская Слобода, д. 19
Почтовый адрес: Россия, 117105, г. Москва, Варшавское ш., 9, стр. 1
Телефон/ Факс +7 495 785 17 00 / +7 495 785 17 01
ОКПО 53779934 ОГРН 1037725059800 ИНН/КПП 7725243282/772501001

Акт

Об использовании результатов кандидатской диссертационной работы
Карповича Сергея Николаевича

Настоящий акт составлен в том, что результаты диссертационной работы «Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов» использованы при проведении анализа поисковых запросов.
Использованы следующие результаты работы:

1. Многозначная классификация ml-PLSI
2. Вероятностное тематическое моделирование
3. Создание темпоральной вероятностной тематической модели
4. Представление результатов тематического моделирования.

Использование многозначной классификации ml-PLSI для большой коллекции поисковых запросов, позволило на 10% повысить точность классификации, что на 20% сократило время работы специалиста необходимого для анализа поисковых запросов. На основе приведенных результатов был разработан:

- Сервис классификации поисковых запросов, позволяющий определять наиболее вероятные тематические группы очень коротких текстов, таких как запрос в поисковую систему.

Антон Сучков

Директор по интернет-маркетингу
Rambler&Co

УНИВЕРСИТЕТ ИТМО

МИНОБРНАУКИ РОССИИ

федеральное государственное автономное
образовательное учреждение высшего образования
«Санкт-Петербургский национальный
исследовательский университет
информационных технологий,
механики и оптики» (Университет ИТМО)

Кронверкский проспект, д. 49, г. Санкт-Петербург,
Российская Федерация, 197101
тел.: (812) 232-97-04 | факс: (812) 232-23-07
od@mail.ifmo.ru | www.ifmo.ru

20.06.2017 № 4-25/625/1

УТВЕРЖДАЮ

Проректор по научной работе
Университета ИТМО
д.т.н., профессор
В.О. Никифоров

«__» _____ 20__

АКТ О ВНЕДРЕНИИ

результатов диссертационной работы

на соискание ученой степени

кандидата технических наук

Карповича Сергея Николаевича

Комиссия в составе: председателя — д-р техн. наук, проф. Парфенова В.Г., заведующего кафедрой информационных систем; и членов комиссии:

- канд. физ.-мат. наук, Зубка Д.А., доцента кафедры информационных систем,
- канд. пед. наук, доц. Маятина А.В., доцента кафедры информационных систем,

Составила настоящий акт, подтверждающий, что результаты диссертационной работы Карповича С.Н. «**Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов**» используются в учебном процессе кафедры информационных систем федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики».

Основные понятия, определения и результаты диссертационного исследования используются при изучении дисциплины «Управление знаниями» магистерской программы по специальности 38.04.05. — Информационные системы бизнеса.

В данной дисциплине используются следующие результаты диссертационной работы:

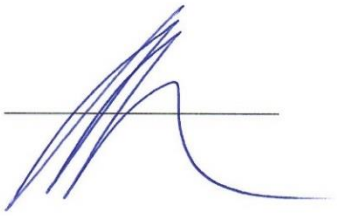
1. Русскоязычный корпус текстов SCTM-ru, позволяющий исследовать алгоритмы вероятностного тематического моделирования и отличающийся от других

корпусов наличием оригинального текста документа и метатекстовой разметки: автор, время описанных событий, тема.

2. Метод расчета матриц вероятностной тематической модели на основе обучения с учителем (авторами документов) с учетом заданных связей документов и тем.
3. Алгоритм многозначной классификации текстовых документов ml-PLSI, основанный на вероятностном тематическом моделировании и заключающийся в использовании матрицы «слово-тема» вероятностной тематической модели для классификации документов, позволяющий определять темы «новых-документов» при анализе потока текстовых документов в динамической тематической модели.

Использование указанных результатов в учебном процессе позволило предоставить студентам актуальные знания о методах построения вероятностных тематических моделей. В материалах курса демонстрируются современные методы расчета матриц «слово-тема» и «документ-тема» вероятностной тематической модели, включая разработанный в диссертации метод на основе обучения с учителем. Подробно описывается алгоритм для многозначной классификации текстовых документов, позволяющий классифицировать по темам ранее не встречавшиеся документы из непрерывного потока документов.

Председатель
Заведующий кафедрой ИС
д-р техн. наук, проф. Парфенов В.Г.



Члены комиссии:

канд. физ.-мат. наук, Зубок Д.А.

канд. пед. наук, доц. Маятин А.В.

Приложение Б: Копия свидетельства об интеллектуальной собственности.

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2017615118

Система для анализа текстовых документов с использованием тематического моделирования

Правообладатель: ***Карпович Сергей Николаевич (RU)***

Автор: ***Карпович Сергей Николаевич (RU)***

Заявка № **2017612769**
Дата поступления **21 марта 2017 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***03 мая 2017 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Г.П. Ивлиев