Использование тематических моделей для парного сравнения коллекций научных статей.

Ф.В. Краснов, А.В. Диментов, М.Е. Шварцман

27 июня 2019 г.

1 Аннотация

Авторами предложена новая методика для парного сравнения коллекций научных статей с помощью тематической модели. Разработанная методика получила название Сравнительного Тематического Анализа (СТА). СТА позволяет получить не только количественную оценку похожести коллекций, но и структурные различия сравниваемых коллекций, как в количественном виде, так и с помощью средств визуализации, разработанных авторами. В данном исследовании проведено сравнение существующих подходов к тематическому моделирования применительно к рассматриваемой задаче сравнения коллекций научных статей. Рассмотрены вероятностные и генеративные тематические модели. Проведен анализ требований к текстовым коллекциям для корректного применения СТА. Методика СТА показала высокую эффективность на выделении структурных различий близких по тематике коллекций. Авторами разработана интегральная метрика «Коэффициент контентной аутентичности», позволяющая сравнивать коллекций между собой. В результате цифрового эксперимента, наиболее информативной показала себя тематическая модель с аддитивной регуляризацией (ARTM).

2 Введение

Научные статьи представляют передовые идеи человечества в структурированной форме исследований и результатов. Но объёмы производства научных текстов в мире растут экспоненциально [1]. Выбор оптимального набора журналов для наблюдения за развитием даже узкой области

исследований стал сложной задачей для исследователей из-за большого потока журналов и для библиотекарей из-за постоянного роста стоимости подписки. Всем приходится постоянно сравнивать различные коллекции научных журналов для выбора подходящих коллекций или отдельных журналов. Такой анализ сложно проводить вручную из-за большого объема сравниваемых массивов данных, поэтому без использования автоматизированных средств эту задачу решить трудно. Хорошим примером такого автоматизированного средства для анализа текстов служит программный пакет PullEnti [2], включающий алгоритмы морфологического и семантико-синтаксического анализа для выделения сущностей определенных типов из текстов естественного языка (персоны, организации, локации и другие целевые семантические объекты).

Анализ значительных объемов научных работ успешно выполняется средствами Интеллектуального анализа текста (ИАТ, Text Mining). Роль тематического моделирования, как инструмента для выделения латентных смыслов становится все более значительной [3—5] в ИАТ. Вторым по значимости современным направлением развития ИАТ де факто стало создание компактных (embedded) представлений текста и его частей: слов (Word2Vec [6]), частей слов (FastText [7]), предложений (Tree-LSTM [8]) и параграфов (paragraph2vec [9]).

Тематическое моделирование и выделение компактных представлений являются представителями методологии неконтролируемого обучения. Тематическое моделирование в силу своих родовых особенностей обладает тремя недостатками:

1. Заранее не известно какое количество тематик задать модели, так как количество тематик является свободным параметром;

2. В изначальной идее тематического моделирования нет механизмов настройки обучения на проблемный домен;

3. Полученные тематики нуждаются в интерпретации.

Эти недостатки не являются проблемами, лишь порождают задачи, которые необходимо решить в ходе конкретного исследования. В частности, настройка на конкретный проблемный домен текста решена в методике построения тематических моделей, названной ARTM [10]. Необходимость интерпретации тематик, в свою очередь, является препятствием для полной автоматизации некоторых задач по анализу текста.

Ключевыми группами задач ИАТ являются: категоризация текстов, извлечение информации и информационный поиск, обработка изменений в коллекциях текстов, а также разработка средств представления

информации для пользователя [11]. В настоящем исследовании авторы сосредоточились на группе задач по обработке изменений в коллекциях текстов, более конкретно на задаче сравнения коллекций слабо структурированных текстов — коллекций научных статей.

Наблюдение, которое стоит за подходом авторов, состоит в следующем: зачастую затраты на изучение коллекции научных журналов могут быть предприняты один раз, чтобы многократно получать дополнительную информацию, сравнивая уже известную коллекцию с новой.

Для реализации такого подхода необходимо иметь общий базис, в котором можно сравнивать коллекции. По гипотезе авторов таким базисом может быть тематическая модель, обученная особым способом. Метаалгоритм для сравнения коллекций, включающий подготовку коллекций, создание и обучение тематической модели и визуализацию результатов сравнения, назван авторами методикой Сравнительного Тематического Анализа (СТА). Дальнейшая структура исследования такова: в разделе «Методика парного сравнения коллекций научных статей» будет детально описан подход, который предлагают авторы, и даны примеры, в разделе «Экспериментальная проверка методики СТА» будет описана постановка и проведение эксперимента по проверке СТА на конкретных коллекциях научных статей, разделе «Заключение» будут приведены результаты исследования в целом и предложены прикладные применения для СТА.

3 Методика парного сравнения коллекций научных статей

Рассмотрим постановку задачи парного сравнения коллекций документов. Под коллекцией j, состоящей из M документов, будем понимать упорядоченную последовательность научных статей $j = (p_0, \ldots, p_M)$. Научная статья p в дальнейшем изложении – это документ, например в формате Дублинского Ядра (Dublin Core), состоящий из текста (d) и метаинформации (авторов (a), даты выпуска (y), названия статьи (s), названия журнала (jo)).

Задача сравнения коллекций может быть поставлена на разных уровнях детализации. В наиболее прямолинейной постановке мы должны найти метрику расстояния $Dist(\circ, \circ)$ от двух аргументов, при подстановке в которую коллекций j_0 и j_1 мы получим число, характеризующее близость двух коллекций. Такой подход обладает двумя преимуществами:

- Удобен для принятия бизнес-решений: принять решение на основании одного числа можно с помощью простого порогового правила. Например, считаем, что коллекции «похожи», если $Dist(\circ, \circ) > 0,5$;

- Позволяет сравнивать метрики $Dist(\circ, \circ)$ для разных пар коллекций и строить матрицу расстояний. Для этого расстояния должны быть нормированы – приведены к одному интервалу, например от 0 до 1. Расстояние равное нулю означает, что коллекции не отличимы. Также можно ввести метрику близости, обратную к метрике расстояния: $Sim(\circ, \circ) = 1 - Dist(\circ, \circ)$. Тогда коллекции будут «близкими» по метрике $Sim(\circ, \circ)$, если значение метрики $Sim(\circ, \circ) = 1$.

Так как коллекции обычно бывают разных размеров, как минимум содержат различное количество документов, но могут иметь и различную структуру метаинформации, то возникает задача выравнивания коллекций для сравнения. Под размером коллекции далее мы будем понимать пространство $\mathbb{R}$ рациональных чисел. В таком случае коллекция j может быть представлена как вектор пространстве $\mathbb{R}$.

Примером представления коллекции j в виде N-мерного вектора может быть словарь всех слов, входящих в тексты документов коллекции. В таком случае каждое слово получает порядковый номер, а размерность N будет определена общим для обоих сравниваемых коллекций j_0 и j_1 словарем. Обозначим такой словарь как V, тогда $N = \dim(V)$, а вектор коллекции будет описываться в пространстве $\mathbb{R}^N$. Другими словами, каждая из сравниваемых коллекций будет представлена вектором в пространстве $\mathbb{R}^N$, состоящим из единиц на тех позициях, для которых слово есть в данной коллекции, и нулей, если такого слова в коллекции нет.

Обозначим операцию представления коллекции j в виде вектора как $Emb(\circ)$. В векторном представлении коллекции в роли метрики близости может выступать, например, косинусная метрика. Выражения (1–2) описывают вид оператора $Emb(\circ)$ и метрики $Sim(\circ, \circ)$ для такого случая:

$$Emb(j_i) \to j_i^N, i \in (0, 1) \tag{1}$$

$$Sim(j_0^N, j_1^N) = \frac{j_0^N \cdot j_1^N}{\|j_0^N\| \cdot \|j_1^N\|} \tag{2}$$

Поясним смысл выражений (1–2) на примере:

Пример: П1

Пусть даны две коллекции $j_0 = (d_0, d_1)$ и $j_1 = (d_0, d_2)$. Для простоты изложения пусть коллекции состоят только из текстов. Документы d_0, d_1, d_2 состоят с следующих слов: $d_0 = (a, dog)$, $d_1 = (a, cat)$, и $d_2 = (a, tree)$. Общий словарь V для коллекций $j_i, i \in (0, 1)$ будет состоять из слов $V = (a, cat, dog, tree)$ и будет обладать размерностью $N = 4$. Таким образом, размерность векторного пространства для сравнения коллекций будет составлять $\mathbb{R}^4$. Другими словами, вектор каждой коллекции будет 4-мерным: $j_0^N = (1, 1, 1, 0)$ и $j_1^N = (1, 0, 1, 1)$. А оператор $Emb(\circ)$ будет действовать по следующему алгоритму:

1. Присвоить вектору коллекции нулевые значения,
2. Для каждого слова из коллекции необходимо определить порядковый номер по словарю V и установить значение 1 в векторном представлении коллекции по этому порядковому номеру.

Так, слово *cat* расположено в словаре V на втором месте, отсутствует в документах коллекции j_0, поэтому в векторном представлении j_0^N во второй позиции стоит 0, но слово *cat* присутствует в документе d_1 из коллекции j_1, поэтому в векторном представлении j_1^N во второй позиции стоит 1.

Векторное пространство $\mathbb{R}^N$ является удобным для сравнения словарного запаса коллекций, но не отражает структуры документов, авторов и другой метаинформации. Тем не менее, руководствуясь описанными принципами, можно расширить пространство $\mathbb{R}^N$ и усовершенствовать оператор $Emb(\circ)$ так, чтобы они учитывали больше информации, содержащейся в коллекции.

Для коллекций j_0, j_1, состоящих из M_0, M_1 документов, введем пространство $\mathbb{R}^{N \times M}$. Где M – это количество документов для объединённой коллекции $J = j_0 \cup j_1$. Коллекция J будет содержать $M = \dim(D_0 \cap D_1) \leq M_0 + M_1$ документов, где $D_i = (d_0, \ldots, d_{M_i}), i \in (0, 1)$. Словарь V для объединённой коллекции останется тем же, как и его размерность N. Тогда выражения (1–2) приобретут следующий вид (3–4):

$$Emb(j_i) \to j_i^{N \times M}, i \in (0, 1) \tag{3}$$

$$Sim(j_0^{N \times M}, j_1^{N \times M}) = \sum_{d=0}^{M} \frac{j_0^{(N \times M)[d]} \cdot j_1^{(N \times M)[d]}}{\|j_0^{N \times M}\| \cdot \|j_1^{N \times M}\|} \tag{4}$$

Оператор $Emb(\circ)$ в выражении (3) теперь преобразует коллекцию не в вектор с размерностью N, а в матрицу с размерностью $N \times M$. В свою очередь, вычисление близости $Sim(\circ, \circ)$ также будет оперировать матрицами. Рассмотрим, как будут выглядеть $j_0^{N \times M}$ и $j_1^{N \times M}$, на примере:

Пример: П2

Рассмотрим коллекции j_0 и j_1 из примера П1. Тогда $M = \dim((d_0, d_1) \cap (d_0, d_2)) = \dim(d_0, d_1, d_2) = 3$. Таким образом, размерность пространства $\mathbb{R}$ равна $dim(\mathbb{R}) = 4 \times 3$. Алгоритм работы оператора $Emb(\circ)$ для размерности документов будет аналогичен алгоритму для работы с размерностью словаря, описанной подробно в примере П1. Матричное представление коллекций $j_0^{N \times M}$ и $j_0^{N \times M}$ будет выглядеть следующим образом:

$$j_0^{N \times M} = \begin{vmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{vmatrix}$$

$$j_1^{N \times M} = \begin{vmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{vmatrix}$$

Разберем подробнее, как получены значения матрицы $j_0^{N \times M}$:

- Значение 1 с индексом $(0, 1)$ – первый ряд сверху, второй столбец слева. Индекс из размерности словаря равен 0. Под этим индексом в словаре V находится слово a. Индекс по размерности документов равен 1. Под индексом 1 в массиве документов D находится документ d_1. В документе d_1 есть слово a, поэтому значение матрицы $j_0^{N \times M}$ по индексу $(0, 1)$ равно 1;
- Значение 0 с индексом $(1, 0)$ - второй ряд сверху, первый столбец слева. Индекс из размерности словаря равен 1. Под этим индексом в словаре V находится слово cat. Индекс по размерности документов равен 0. Под индексом 0 в массиве документов D находится документ d_0. В документе d_0 нет слова cat, поэтому значение матрицы $j_0^{N \times M}$ по индексу $(1, 0)$ равно 0.

Описанная выше методика учета структуры документов при сравнении коллекций имеет ряд недостатков при практическом применении. Основным недостатком является то, что в реальных коллекциях почти не встречаются одинаковые документы. Поэтому матрицы $j_i^{N\times M}$ будут иметь высокую степень разреженности, что приводит к понижению чувствительности оператора $Emb(\circ)$. Значения косинусов в районе нуля слабо отличаются. Устранить данный недостаток может более компактное представление документов в пространстве меньшей размерности. Например, с помощью выделения групп похожих документов можно выделить подпространство $\mathbb{R}^C$ такое, что $C \ll M$, где C будет количеством групп, на которые разбиты документы D.

3.1 Обзор подходов к группировке текстов

Рассмотрим применимость различных алгоритмов группировки текстов для задачи сравнения коллекций. С точки зрения машинного обучения рассматриваемая задача относится к классу задач неконтролируемого обучения (unsupervised machine learning). Неконтролируемые методы обучения не требуют каких-либо обучающих данных, поэтому могут применяться к любым текстовым данным, без дополнительных усилий на разметку данных. Двумя основными неконтролируемыми методами обучения, обычно используемыми применительно к текстовым данным, являются кластеризация и тематическое моделирование. Проблема кластеризации заключается в том, чтобы разбить совокупность документов на разделы, каждый из которых соответствует тематическому кластеру. Проблемы кластеризации и тематического моделирования тесно связаны и корнями уходят к дистрибутивной гипотезе лингвистики, высказанной в работе [12]. Суть дистрибутивной гипотезы состоит в том, что слова с похожими значениями встречаются в похожих окружениях слов.

В тематическом моделировании мы используем вероятностную модель для определения мягкой кластеризации, в которой каждый документ имеет вероятность членства в кластере, а не жёсткую сегментацию документов. Тематические модели можно рассматривать как процесс кластеризации с генеративной вероятностной моделью. Каждую тему можно рассматривать как распределение вероятностей по словам, причем наиболее характерные слова имеют наибольшую вероятность. Каждый документ может быть выражен как вероятностная комбинация этих различных тем. Таким образом, тема может рассматриваться как аналог кластера, а принадлежность документа к кластеру носит вероятностный характер. Это также приводит к более элегантному представлению членства в кластере в тех случаях, когда известно, что документ содержит

различные темы. В случае жесткой кластеризации иногда бывает сложно присвоить документ одному кластеру. Кроме того, тематическое моделирование элегантно связано с проблемой сокращения измерений, где каждая тема обеспечивает концептуальное измерение, и документы могут быть представлены в виде линейной вероятностной комбинации этих различных тем. Таким образом, тематическое моделирование обеспечивает чрезвычайно общую структуру, которая относится как к проблемам кластеризации [13], так и к уменьшению размерности [14].

Проблема уменьшения размерности широко изучается в литературе как метод представления базовых данных в сжатом формате для индексации и поиска [15—18]. Вариант уменьшения размерности, который обычно используется для текстовых данных, известен как латентная семантическая индексация (LSI) [19]. Одной из интересных характеристик латентной семантической индексации является то, что она сохраняет ключевые семантические аспекты текстовых данных, что делает ее более подходящей для различных приложений интеллектуального анализа данных. Например, шумовые эффекты синонимии и полисемии уменьшаются из-за использования таких методов уменьшения размерности. Значительное место занимают подходы к уменьшению размерности, основанные на матричном исчислении, например, неотрицательное матричное разложение (NMF) [20; 21]. Другое семейство методов уменьшения размерности - это вероятностные тематические модели, в частности, PLSA [22], LDA [23] и их варианты; они выполняют уменьшение размера вероятностным способом с потенциально более значимыми представлениями темы, основанными на распределении слов.

Авторы настоящего исследования считают, что применение тематического моделирования для сравнительного анализа коллекций научных статей является перспективным направлением для более глубокого изучения. В целом, задача сравнения коллекций научных статей относится к направлению Сравнительного Текстового Анализа (comparative text mining) и включает в себя согласно [24]: обнаружение общих компонентов по всем коллекциям, выявление уникальных отличий между коллекциями по обнаруженным компонентам и количественный анализ отклонений по компонентам. Ключевыми исследованиями в этом направлении являются следующие: [25] – изучение связей двух корпусов текстов, [26] – сравнение коллекций по извлеченным знаниям, [27] – сравнение коллекций по ключевым фразам, [28] – гибридный подход с использованием связей «слово-слово» и сравнения извлеченных знаний. Также стоит разделить исследования информативности метрик для сравнения компонент коллекций и способов выявления компонент. Основными методами выделения компонент из текста являются: метод k-средних – [29],

метод сдвига среднего значения – [30; 31], спектральная кластеризация – [32], метод анализа плотности – [33], метод сбалансированного итеративного сокращения – [34].

3.2 Методика сравнительного тематического анализа

Для рассмотрения предлагаемой авторами методики Сравнительного Тематического Анализа (СТА) введём формальное определение тематической модели.

> Определение: О1
>
> Тематической моделью коллекции документов называются две стохастические матрицы, Φ и Θ, такие, что их векторное произведение равно матрице $P(V, D)$ вероятностей слов из словаря для каждого документа коллекции. Тематическую модель можно записать в виде матричного произведения: $P(V, D) = \Phi \cdot \Theta$. Размерность матрицы $\dim(\Phi) = N \times C$, размерность матрицы $\dim(\Theta) = C \times M$, а размерность матрицы $\dim(P) = N \times M$. Где $N = \dim(V)$ – количество слов в словаре, $C = \dim(T)$ – количество тематик, а $M = \dim(D)$ – количество документов в коллекции.

Развитие тематического моделирования, начиная с исследования [22], сосредоточено на повышении информативности тематик. В работе [35] показано, что человеческая оценка интерпретируемости тематик хорошо коррелирует с автоматизированной мерой качества, называемой когерентностью тематики $Coherence(t)$. В научной литературе предложено несколько различных автоматических методов ранжирования тем, которые измеряют $Coherence(t)$ и затем оцениваются путем сравнения с человеческими оценками. В самом общем виде для вычисления меры согласованности одной темы t берут N_t слов с наибольшими вероятностями $V^{(t)} = (w_0^{(t)}, \ldots, w_{N_t}^{(t)})$, суммируют меру информационных отношений по всем парам слов – $IR(w_i^{(t)}, w_j^{(t)})$ и нормируют. Мера информационных отношений $IR(w_i^{(t)}, w_j^{(t)})$ зависит от одной пары слов из множества $V^{(t)}$ и может быть вычислена разными способами, детально описанными далее. В дальнейшем мы будем опираться на метрику $Coherence(t)$ для определения качества тематической модели, поэтому введём необходимые понятия для пояснения того, как вычисляется метрика $Coherence(t)$.

В работе [36] предложена следующая формула для $Coherence(t)$:

$$C_{UCI}(t; V^t) = \frac{2}{N_t(N_t-1)} \sum_{i=1}^{N_t-1} \sum_{j=i+1}^{N_t} PMI(w_i^{(t)}, w_j^{(t)}) \ , \tag{5}$$

где $PMI(w_i, w_j) = \log \frac{N \cdot N_{w_i w_j}}{N_{w_i} \cdot N_{w_j}}$ – по-точечная взаимная информация, $N_{w_i w_j}$ – число документов, в которых слова w_i и w_j хотя бы один раз встречаются в окне k. N_{w_j} – число документов, в которых слово w_j встретилось хотя бы один раз, а N – это количество слов в словаре.

В работе [37] приведена другая формула для вычисления $Coherence(t)$:

$$C_{UMass}(t; V^t) = \frac{2}{N_t(N_t-1)} \sum_{i=1}^{N_t-1} \sum_{j=i+1}^{N_t} \log \frac{N_{w_i w_j} + \epsilon}{N_{w_j}} \tag{6}$$

Кроме C_{UMass} и C_{UCI} существуют ещё варианты определения когерентности тематик с использованием нормализованной метрики PMI (NPMI) [38].

$$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log(N_{w_i w_j} + \epsilon)} \tag{7}$$

$$C_{NPMI}(t; V^t) = \frac{2}{N_t(N_t-1)} \sum_{i=1}^{N_t-1} \sum_{j=i+1}^{N_t} NPMI(w_i^{(t)}, w_j^{(t)}) \tag{8}$$

Четвертым вариантом для расчета когерентности является подход на основе косинусной близости векторов из значений NPMI для ста наиболее близких слов из словаря для каждого слова из темы [35]. Такую меру когерентности будем обозначать C_v.

Но в работе [39] экспериментально доказано, что C_{UCI} показывает более точные результаты, чем NPMI. Для задач настоящего исследования достаточно выбрать один способ вычисления когерентности. Наиболее универсальным с методической точки зрения представляется C_{UCI}, но в экспериментальной части статьи мы проверим и другие варианты вычисления когерентности.

Основываясь на определении О1 мы можем преобразовать выражения (3–4) для использования тематической модели. Основное отличие будет состоять в уменьшении размерности пространства $\mathbb{R}^{N \times C}$, где $C = \dim(T)$ – это количество тематик для объединённой коллекции $J = j_0 \cup j_1$. Выражения (3–4) будут иметь следующий вид:

$$Emb(J) \to j_0^{N\times C}, j_1^{N\times C} \equiv \Phi, \Theta \tag{9}$$

$$Sim(j_0^{N\times C}, j_1^{N\times C}) = \sum_{t=0}^{C} \frac{j_0^{(N\times C)[t]} \cdot j_1^{(N\times C)[t]}}{\|j_0^{N\times t}\| \cdot \|j_1^{N\times t}\|} \tag{10}$$

В выражении (9) оператор $Emb(\circ)$ изменил алгоритм работы: для построения тематической модели необходимо обработать объединённую коллекцию J, чтобы получить матрицу Θ, содержащую документы из обоих коллекций j_0 и j_1. Поясним это изменение с помощью схемы (рис. 1–2):

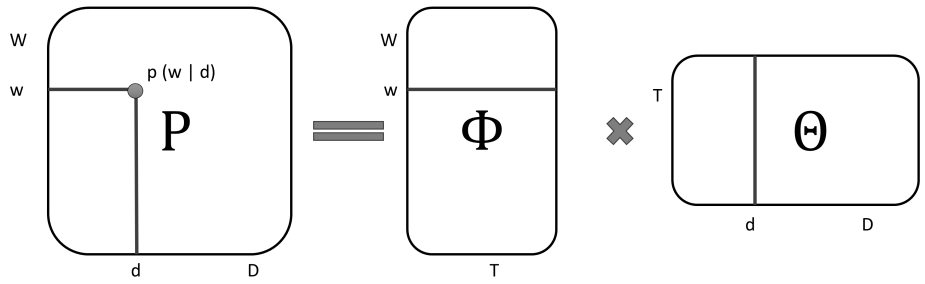

Рис. 1: Матричное представление тематической модели

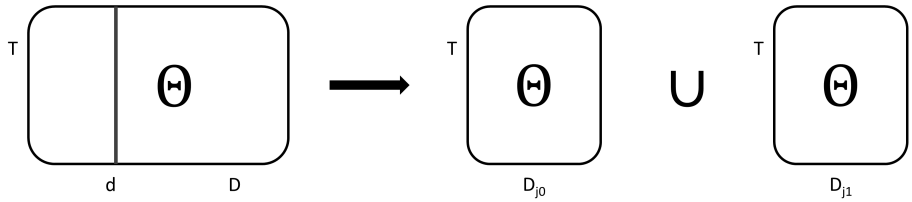

Рис. 2: Разложение матрицы Θ на коллекции j_0 и j_1

На схеме (рис. 2) показано, как матрица Θ раскладывается на две матрицы, каждая из которых относится к документам одной коллекции. На практике для повышения когерентности тематик при объединении коллекций оказалось целесообразно перемешивать документы в случайном порядке. Но при этом запоминать принадлежность документов к коллекциям.

В выражении (10) метрика $Sim(\circ, \circ)$ будет иметь размерность C. Другими словами, информация о близости коллекций будет получена в разрезе тематик $T = (t_0, \ldots, t_C)$. Эта возможность является одним из важных преимуществ при использовании разработанной авторами методики

Сравнительного Тематического Анализа для решения практических задач.

Рассмотрим, как будут выглядеть $j_0^{N\times C}$ и $j_1^{N\times C}$, на примере:

Пример: ПЗ

Пусть даны две коллекции $j_0 = (d_0)$ и $j_1 = (d_1, d_2)$. Для простоты изложения пусть коллекции состоят только из текстов.

d_0 = 'cat cat tree tree cat'
d_1 = 'dog dog dog bone bone bone'
d_2 = 'dog cat tree tree bone bone'

Общий словарь V для коллекций $j_i, i \in (0, 1)$ будет состоять из слов $V = (bone, cat, dog, tree)$ и будет обладать размерностью $N = \dim(V) = 4$.

Объединенная коллекция $J = j_0 \cup j_1$ будет состоять из документов $D = (d_0) \cap (d_1, d_2) = (d_0, d_1, d_2)$. Размерность D будет равна $M = \dim(D) = \dim(d_0, d_1, d_2) = 3$. Для простоты рассмотрения предположим, что нам будет достаточно двух тематик: $T = (t_0, t_1)$. Тогда размерность тематик C будет равна $C = \dim(T) = 2$. Таким образом, размерность пространства $\mathbb{R}$ равна $dim(\mathbb{R}) = M \times C = 4 \times 2$. Алгоритм работы оператора $Emb(\circ)$ будет решать оптимизационную задачу $\|P(V, D) - \Phi \cdot \Theta\|_{Fro} \to \min$, где $\|A\|_{Fro} = \sum_{ij} A_{ij}^2$ (Норма Фробениуса).

Матрица $P(V, D)$ для нашего примера представлена в таблице 1.

Таблица 1: Матрица $P(V, D)$

	d_0	d_1	d_2
bone	0,00	0,50	0,33
cat	0,60	0,00	0,17
dog	0,00	0,50	0,17
tree	0,40	0,00	0,33

Разберем подробнее, как получены значения матрицы $P(V, D)$:

- Значение 0,40 с индексом $(3, 0)$ - четвертый ряд сверху, первый столбец слева. Индекс из размерности словаря равен 3. Под этим индексом в словаре V находится слово $tree$. Индекс по размерности документов равен нулю. Под индексом 0 в массиве документов D находится документ d_0. В документе d_0 слово $tree$ встречается два раза, а сам документ содержит пять слов. Поэтому относительная частота слова $p(d_0|"tree") = 2/5 = 0,4$.
- Значение 0.33 с индексом $(3, 2)$ – четвертый ряд сверху, третий столбец слева. Индекс из размерности словаря равен 3. Под этим индексом в словаре V находится слово $tree$. Индекс по размерности документов равен 2. Под индексом 2 в массиве документов D находится документ d_2. Документ d_2 состоит из шести слов, а слово $tree$ встречается дважды, поэтому значение матрицы $P(V, D)$ по индексу $(3, 2)$ равно $2/6 = 0,33$.

В результате работы оператора $Emb(\circ)$ полученные матрицы Φ и Θ представлены в таблице 2 и таблице 3.

Таблица 2: Матрица Φ

	t_0	t_1
bone	0,02	0,70
cat	0,42	0,00
dog	0,00	0,30
tree	0,56	0,00

Матрица Φ представляет распределение вероятностей для каждого слова по двум тематикам. Сумма вероятностей всех слов для каждой тематики (строки матрицы Φ) равна единице. Так как по определению О1 матрица Φ строится как стохастическая матрица. Из таблицы 2 мы видим, что тема t_0 имеет наибольшие вероятности у слов $cat, tree$. А тема t_1 имеет наибольшие вероятности у слов $dog, bone$.

Ошибка вычисления Φ и Θ в данном примере вычисляется по формуле $\|P(V, D) - \Phi \cdot \Theta\|_{Fro} = 0,02$

Таблица 3: Матрица Θ

	d_0	d_1	d_2
t_0	1,00	0,00	0,64
t_1	0,00	1,00	0,36

3.3 Длина текста

В большинстве тематических моделей темы представлены в виде групп коррелированных слов, причем корреляция в основном соответствует паттернам совпадения слов в документах. Например, после наблюдения слов «dog» и «bone», часто встречающихся друг с другом, можно сказать, что они имеют общее употребление и, возможно, принадлежат к одной и той же теме, хотя их точного значения мы можем не знать. Традиционные генеративные тематические модели используют паттерны совместного использования слов, чтобы неявным образом выявлять скрытую семантическую структуру документов путем моделирования генерации слов в каждом документе. Такие подходы чувствительны к краткости документов, поскольку шаблоны совместного использования слова в одном коротком документе редки и ненадежны.

Отметим, что рассматриваемые в примере П3 документы очень короткие и приведены только в показательных целях. Большинство из упомянутых выше методик для построения тематических моделей не предназначены для работы с короткими текстами. Причина, по которой вероятностные и генеративные модели плохо работают с короткими текстами состоит, в том, что распределения плохо сбалансированы. Существует отдельный класс методик построения тематических моделей для работы с короткими (150–180 слов) и очень короткими (5–7 слов) документами. К наиболее продуктивным методикам для работы с несбалансированными текстами относятся Word Network Topic Model (WNTM, [40]) и Biterm Topic Model (BTM, [41]), которые используют совстречаемость (co-occurence) слов в узком окне.

В этих моделях создаются фиктивные документы для каждого слова, состоящие из слов, с которыми это слово употреблялось в окне, состоящем из трех слов. WNTM и BTM могут работать и с текстами обычной длины, но показывают низкую производительность.

С другой стороны, в настоящем исследовании мы можем рассматривать аннотации научных статей как короткие тексты, а полные тексты статей как полноценные текстовые документы. Аннотации научных статей являются открытой информацией, и создать большой датасет из ан-

нотаций технически проще. В исследовании [42] изучено поведение одной тематической модели LDA на датасетах из аннотаций и полных текстах научных статей. Авторами [42] показано, что метрика когерентности для аннотаций получается не хуже, чем для полных текстов научных статей. В настоящем исследовании мы можем продвинуться дальше и рассмотреть результаты нескольких тематических моделей применительно к коротким и длинным текстам для решения задачи сравнительного анализа коллекций.

3.4 Коэффициент контентной аутентичности

Для количественной оценки различия коллекций в рамках CTM авторы предложили использовать сумму модулей отклонений от равномерного распределения тематик, деленную на количество тематик – Коэффициент контентной аутентичности (κ).

$$\kappa(j_0, j_1) = \frac{1}{C} \sum_{t \in C} abs \left[\frac{\sum_{d \in j_0} \theta(d,t)}{\sum_{\hat{d} \in j_0} \sum_{\hat{t} \in C} \theta(\hat{d},\hat{t})} - \frac{\sum_{d \in j_1} \theta(d,t)}{\sum_{\hat{d} \in j_1} \sum_{\hat{t} \in C} \theta(\hat{d},\hat{t})} \right] \qquad (11)$$

Максимальное значение κ будет равно единице, когда каждая из выявленных тематик относится только к одной сравниваемых коллекции.

4 Эксперимент

4.1 Подготовка данных

Для участия в эксперименте было выбрано три коллекции научных журналов на русском языке – j_0, j_1, j_2. Коллекции j_0 и j_1 выбраны из одной области медицины, коллекция j_2 состоит их журналов по тематике наук о Земле. В таблице 4 собраны общие описательные данные по коллекциям.

Таблица 4: Коллекции научных текстов

Название	Направление	Язык	Кол-во
j_0	Медицина	русский	702
j_1	Медицина	русский	2599
j_2	Науки о Земле	русский	1123

В исходном виде коллекции находились в формате Дублинского ядра (Dublin Core). Затем коллекции были приведены к формату Vowpal Wabbit[1]. По сути это файл в котором одна строка соответствует одному документу. Для объединенной коллекции количество документов равно $M = \dim(D) = 4424$. Каждый документ был дополнен биграммами входящих в него слов. При образовании биграмм была сохранена последовательность слов в документе. Такой подход к построению корпуса документов называется «Мешок слов» (Bag of Words, BoF [12]). На следующем шаге подготовки коллекции к анализу был создан словарь V объединенной коллекции, размерность которого составила $N = \dim(V) = 2572622$. На рис. 3 и 4 приведены частотные гистограммы исходного и очищенного от шума словарей. Для очистки от шума к сло-

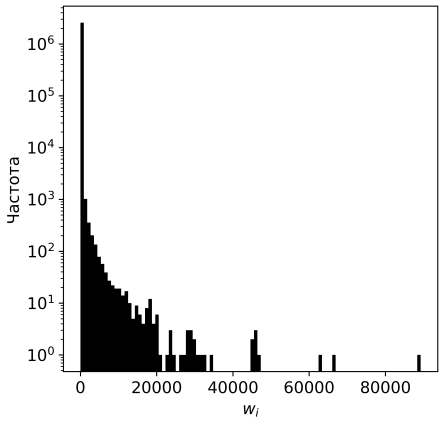

Рис. 3: Частотная гистограмма слов объединенной коллекции

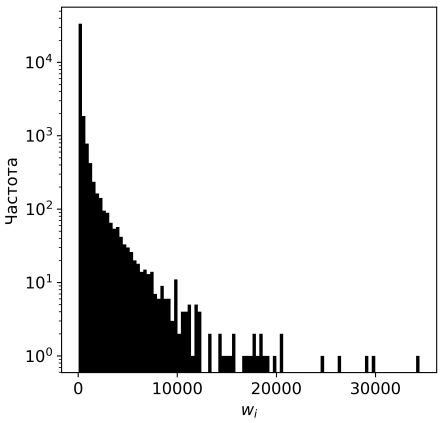

Рис. 4: Частотная гистограмма слов объединенной коллекции после удаления шума

варю были применены два фильтра: отброшены слова с частотой употребления меньше 40 и слова, которые встречались более чем в 60% документов. В очищенном словаре объединенной коллекции получилось $N = \dim(V) = 37\,910$ слов (униграмм и биграмм). Полученный набор данных доступен в Mendeley Data по адресу [2].

[1] https://github.com/JohnLangford/vowpal_wabbit/wiki

[2] `http://dx.doi.org/10.17632/s89sbmxwt4.1`

4.2 Выбор количества тематик

Независимо от того какую методику для построения модели мы выберем далее, основным и самым важным параметром для тематической модели является количество тематик C. Покажем экспериментально, что рассмотренная ранее метрика $Coherence(t)$ не может быть использована в для определения количества тематик C. Для этого проведем измерения $Coherence(t)$ для разных значений C и рассмотрим форму зависимости. Оптимальное количество тематик C соответствует максимуму когерентности. На рисунках 5–8 представлены зависимости $Coherence(t)$ от количества тематик C для различных методик тематического моделирования.

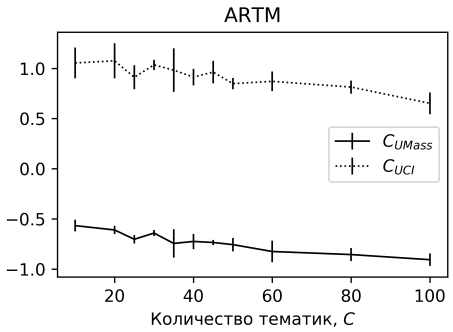

Рис. 5: ARTM

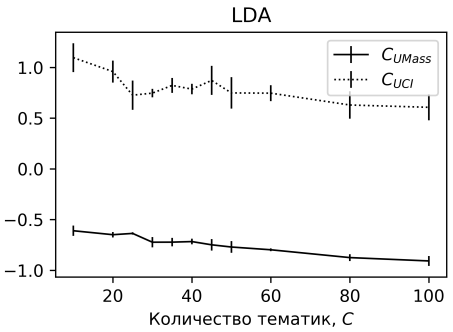

Рис. 7: LDA

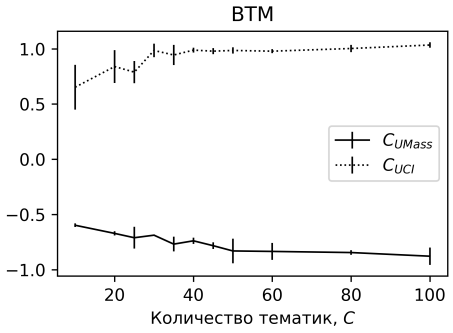

Рис. 6: BTM

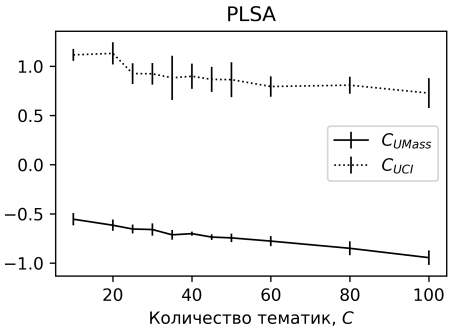

Рис. 8: PLSA

Мы рассмотрели методики ARTM, LDA, BTM, PLSA и построили зависимости когерентностей C_{Umass} и C_{UCI} для диапазона $C \in (10, 100)$. Ошибка измерений возникает при разном порядке документов в объединенной коллекции. Поэтому измерения проводились для десяти различных вариантов порядка документов, полученных случайным перемеши-

ванием. Из рисунков 5–8 видно, что когерентность модели убывает и не имеет экстремумов. Отсутствие максимума когерентности не позволяет определить оптимальное количество тематик. Что и требовалось доказать.

В исследовании [43] предложена методика для оценки количества тематик на основе качества полученных кластеров. Разработанная авторами [43] метрика качества $cDBI$ имеет максимум при оптимальном значении C. На рис. 9 приведена зависимость метрики $cDBI$ от C для объединенной коллекции J.

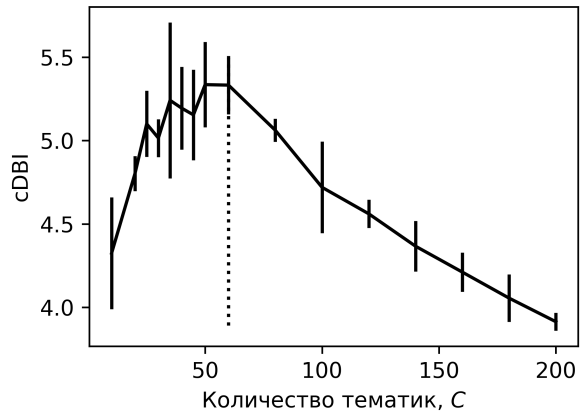

Рис. 9: Зависимость метрики $cDBI$ от количества тематик C

На рис. 9 мы можем наблюдать максимум качества выделяемых кластеров при значении $C = 60$.

4.3 Выбор методики тематической модели

Для выбора методики тематической модели были рассмотрены методики ARTM, LDA, BTM, PLSA. Критерием для выбора методики модели стала метрика когерентности. Когерентность была вычислена для каждой из рассматриваемых моделей в виде зависимости от количества итераций обучения. Так как когерентность зависит от порядка документов, то для каждой итерации обучения было вычислено десять значений когерентности для коллекций с различным случайным порядком документов. На зависимости когерентности от времени показано среднее значение для каждой итерации и среднеквадратическое отклонение.

На рисунках 5–8 приведена зависимость различных вариантов метрики $Coherence(t)$ от количества итераций обучения тематической модели

для объединенной коллекции J.

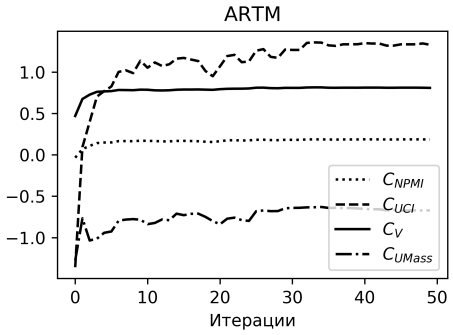

Рис. 10: ARTM

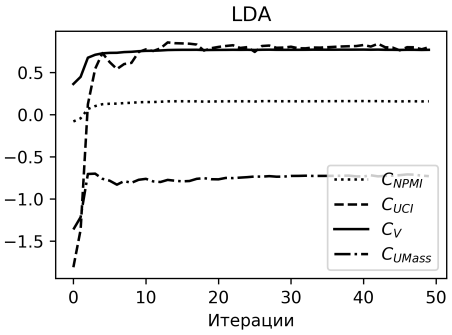

Рис. 12: LDA

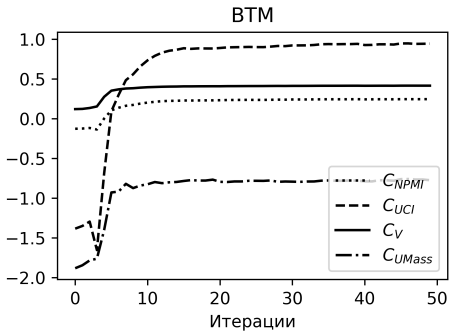

Рис. 11: BTM

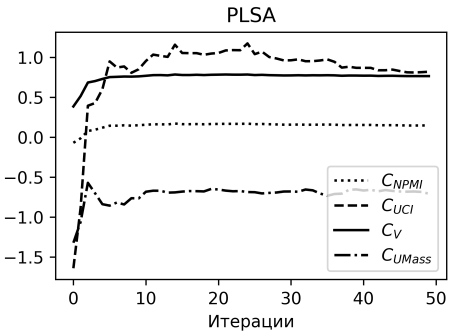

Рис. 13: PLSA

Из зависимостей на рисунках 5–8 можно сделать вывод, что наиболее информативным вариантом расчета метрики $Coherence(t)$ является вариант C_{UCI}.

На рисунке 14 приведены зависимости C_{UCI} для рассматриваемых методик тематического моделирования. Сравнение зависимостей позволяет сделать вывод о том, что наиболее когерентные тематики могут быть получены с помощью методики ARTM.

4.4 Стратегия обучения тематической модели

До этого мы не рассматривали такое важное свойство тематической модели, как разреженность матриц Φ и Θ. На практике представляется выгодной ситуация, когда для одного документа в столбце матрицы Θ представлен только один не нулевой коэффициент, как в примере П3 в таблице 3 для документов d_0 и d_1. Это означает, что содержание всего

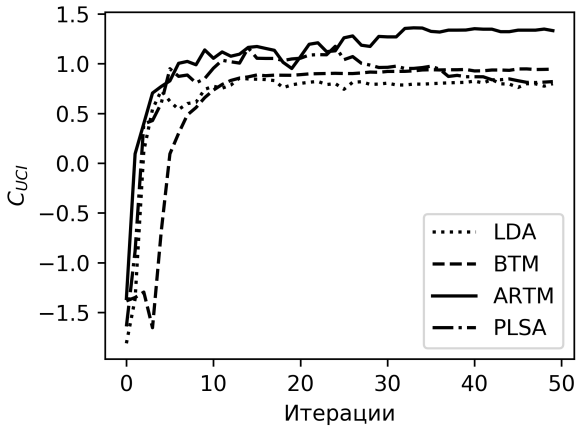

Рис. 14: Зависимость когерентности тематической модели $Coherence(t)$ от количества тематик итераций обучения модели

документа укладывается в одну тематику – строку из матрицы Φ. Следствием такой ситуации будет то, что матрицы Θ и Φ будут содержать много нулей или, другими словами, будут разреженными матрицами. На практике такого разреженного вида матриц Θ и Φ добиваются с помощью определенной последовательности обучения – стратегии обучения.

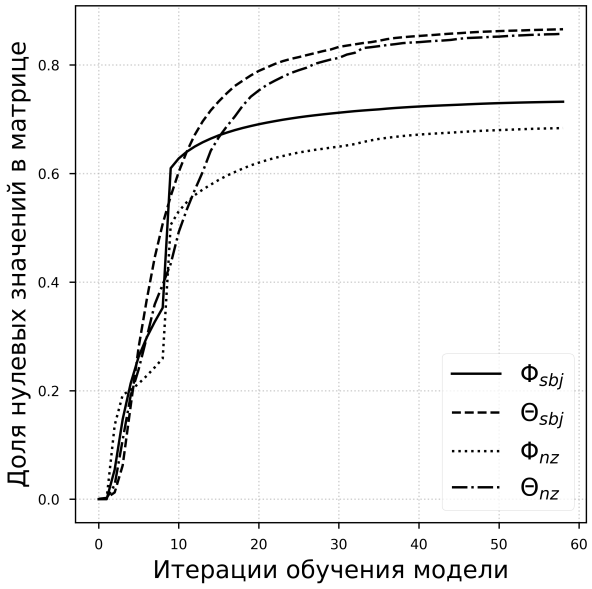

Рис. 15: Стратегия обучения тематической модели

На рисунке 15 графически отображена эволюция изменения разреженности матриц Φ и Θ. В данной стратегии обучения применен принцип разделения на основные тематики (sbj_i) и шумовые тематики (nz_i). Интуитивная логика, стоящая за данной стратегией, состоит в том, что в каждой статье, например, есть обзор научной литературы, содержащий наиболее значимые научные исследования, сделанные в этом направлении. Но основная тематика статей может отличаться. Таким образом, условно, обзор научной литературы может быть выделен в отдельные тематики, которые назовем шумовыми, так как они могут быть сразу во многих статьях, в отличие от основной тематики статьи, которая не должна повторяться.

Для достижения различной степени разреженности в методике ARTM используются соответствующие регуляризаторы. Добавление регуляризатора в качестве слагаемого к кост-функции в процессе оптимизации позволяет накладывать штраф на варианты матриц с неподходящими свойствами.

5 Сравнительный анализ

Теперь, когда выбрана методика тематической модели и количество тематик, мы можем приступить к испытанию методики Сравнительного Тематического Анализа (СТА). Для этого обучим модель для объединенной коллекции J в соответствии с изложенной в разделе (4.4) стратегией обучения.

В результате у нас получатся матрицы Φ и Θ. Матрица Θ отображена на рисунке 16.

Из рисунка 16 мы видим, что строки матрицы Θ, относящиеся к основным тематикам (sbj_i), достаточно разряжены (sparse), а строки, относящиеся к шумовым тематикам (nz_i), наоборот, плотные (dense). Это результат работы стратегии обучения тематической модели (4.4).

5.1 Различные коллекции

В разделе (4.1) в таблице 4 приведены три коллекции. Рассмотрим для сравнения две коллекции: «Медицина» (j_1) и «Науки о Земле»(j_2). При таких начальных условиях мы ожидаем, что СТМ выделит разные наборы тематик. На рисунке 17 представлены тематики, выделенные для коллекций «Медицина» и «Науки о Земле».

На основе визуального наблюдения рисунка 17 можно сделать вывод о том, что коллекции «Медицина» и «Науки о Земле» представлены

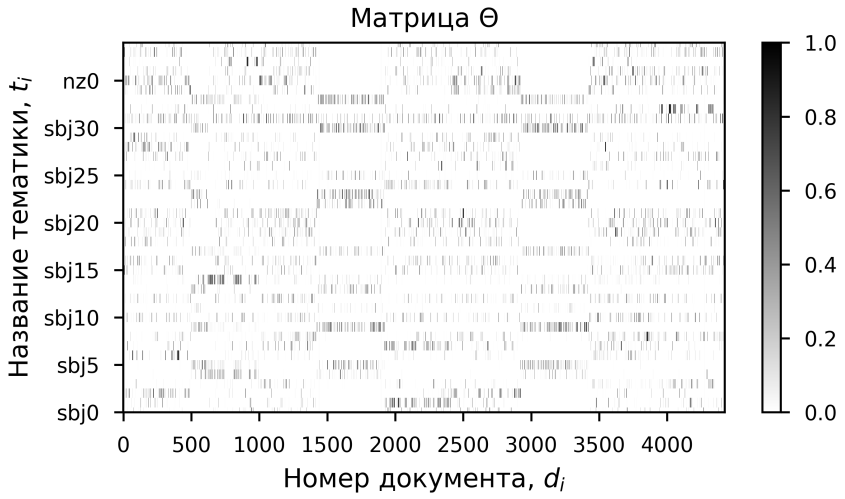

Рис. 16: Матрица Θ тематической модели объединенной коллекции J

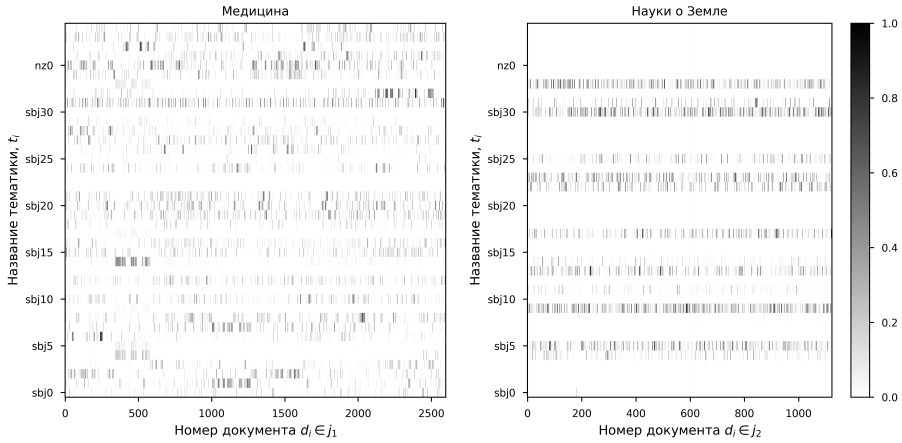

Рис. 17: Матрицы Θ_{j_1} и Θ_{j_2}

разными тематиками. Более наглядно это различие можно увидеть на гистограмме, изображенной на рисунке 18, отображающей суммарный вес для каждой тематики по каждой из коллекций.

Из гистограммы, изображенной на рисунке 18, мы видим, что большинство столбиков имеют один цвет – принадлежат к одной коллекции. Например, тематика sbj_9 полностью связана с «Науками о Земле». Слова с наибольшим весом в этой тематике следующие: «комплекс, порода, урал, зона, массив, состав, базальт». С другой стороны, в тематике sbj_{20} с

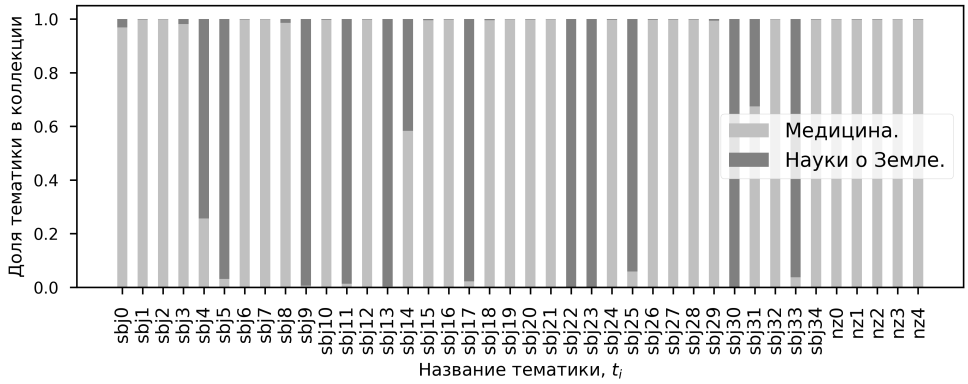

Рис. 18: Профили тематик для коллекций «Медицина» и «Науки о Земле»

наибольшими вероятностями будут только слова, относящиеся к «Медицине»: «препарат, эффективность, эффект, доза, мес, боль, применение». Но будут и смешанные тематики, в которых с большими значениями вероятности будут присутствовать слова, относящиеся к обоим коллекциям. Количественной мерой расстояния (отличия) коллекций является Коэффициент контентной аутентичности (κ). Для рассматриваемого случая различных по тематике коллекций $\kappa(j_0, j_1) = 0,95$.

5.2 Похожие коллекции

Результат, полученный в предыдущем разделе данной статьи, для заведомо различных коллекций интересен как предельный случай. Но так же важно понимать границы применимости СТМ и чувствительности к изменениям в коллекциях. Поэтому для следующего эксперимента были взяты две коллекции журналов по направлению «Медицина» с близкими тематиками – «Медицина 0» (j_0) и «Медицина 1» (j_1). При таких начальных условиях мы ожидаем, что СТМ выделит схожие наборы тематик. На рисунке 19 представлены тематики, выделенные для коллекций «Медицина 0» и «Медицина 1».

На основе визуального наблюдения рисунка 19 можно сложно вывод о том, что коллекции «Медицина 0» и «Медицина 1» представлены схожими по плотности тематиками. Более наглядно это можно увидеть на гистограмме, изображенной на рисунке 20, отображающей суммарный вес для каждой тематики по каждой коллекции.

Из гистограммы, изображенной на рисунке 20 мы видим, что боль-

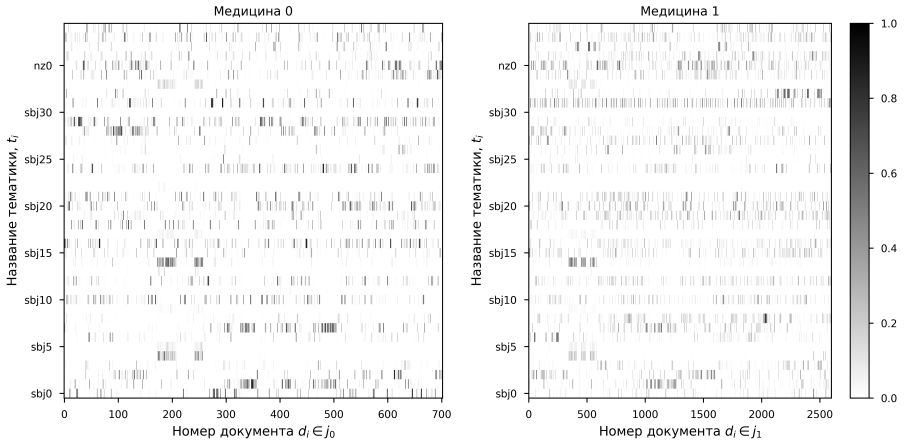

Рис. 19: Матрицы Θ_{j_0} и Θ_{j_1}

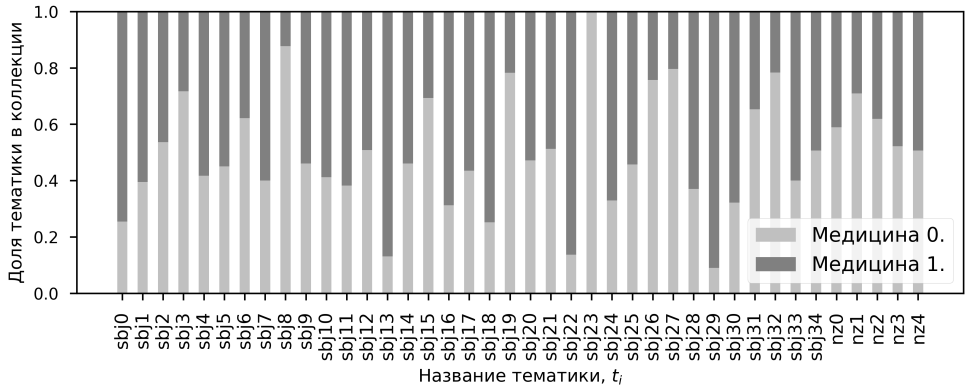

Рис. 20: Профили тематик для коллекций «Медицина 0» и «Медицина 1»

шинство столбиков имеют оба цвета примерно в одинаковой пропорции, другими словами, принадлежат к обеим коллекциям. Для рассматриваемого случая близких по тематике коллекций значение Коэффициент контентной аутентичности - $\kappa(j_0, j_1) = 0,32$. Так как Коэффициент контентной аутентичности отражаем дистанцию (разницу) между коллекциями, то чем ближе κ к нулю, тем более похожи коллекции.

6 Заключение

В данном исследовании авторы предложили методику для сравнения коллекций научных журналов (СТМ). Предложенная методика показала свою наглядность и количественную точность в ходе проведенного авторами эксперимента. В качестве метрики расстояния между коллекциями авторами предложен Коэффициент контентной аутентичности. В таблице 5 собраны значения Коэффициента контентной аутентичности для трех исследуемых коллекций.

Таблица 5: Коэффициент контентной аутентичности

κ	Медицина 0	Медицина 1	Науки о Земле
Медицина 0	0	0.32	0.94
Медицина 1	0.32	0	0.95
Науки о Земле	0.94	0.95	0

Результаты, отображенные в таблице 5, полностью согласуются с теорией.

Авторы рассмотрели четыре наиболее популярных методики тематического моделирования: ARTM, BTM, PLSA и LDA. Проведенный эксперимент по выбору методики тематического моделирования показал, что по метрике когерентности C_{UCI} наиболее высокое значение когерентности у методики ARTM.

В процессе методического обзора авторами было обнаружено, что на сновании имеющихся метрик когерентности тематических моделей не представляется возможным определить наиболее значимый параметр тематической модели – количество тематик. Авторы подтвердили эту находку экспериментально. Действительно метрика когерентности не имеет глобальных экстремумов. Важно отметить следующее: ни одна из рассмотренных метрик не учитывает, что порядок документов влияет на тематики и на их когерентность. Авторы показали в ходе эксперимента, что метрики тематической модели необходимо считать для различного порядка документов и усреднять полученные значения. Эта методическая находка позволяет оперировать средними значениями метрик и учитывать возникающую статистическую ошибку. Также важно, что ни одна из рассмотренных метрик когерентности не учитывает вероятности тематик (Ф), а только выбирает топ-N тематик. Поэтому для определения количества тематик авторами была использована метрика $cDBI$, учитывающая качество получаемых кластеров, порядок документов и полные

вектора тематик из матрицы Ф.

Список литературы

1. On the shoulders of giants: The growing impact of older articles / A. Verstak [и др.] // arXiv preprint arXiv:1411.0275. — 2014.
2. *Козеренко Е. Б.*, *Кузнецов К. И.*, *Романов Д. А.* Семантическая обработка неструктурированных текстовых данных на основе лингвистического процессора PullEnti // Информатика и её применения. — 2018. — Т. 12, № 3. — С. 91—98.
3. *Alghamdi R.*, *Alfalqi K.* A survey of topic modeling in text mining // Int. J. Adv. Comput. Sci. Appl.(IJACSA). — 2015. — Т. 6, № 1.
4. Structural topic models for open-ended survey responses / M. E. Roberts [и др.] // American Journal of Political Science. — 2014. — Т. 58, № 4. — С. 1064—1082.
5. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey / H. Jelodar [и др.] // Multimedia Tools and Applications. — 2018. — С. 1—43.
6. Distributed representations of words and phrases and their compositionality / T. Mikolov [и др.] // Advances in neural information processing systems. — 2013. — С. 3111—3119.
7. Bag of tricks for efficient text classification / A. Joulin [и др.] // arXiv preprint arXiv:1607.01759. — 2016.
8. *Tai K. S.*, *Socher R.*, *Manning C. D.* Improved semantic representations from tree-structured long short-term memory networks // arXiv preprint arXiv:1503.00075. — 2015.
9. *Le Q.*, *Mikolov T.* Distributed representations of sentences and documents // International conference on machine learning. — 2014. — С. 1188—1196.
10. *Vorontsov K.*, *Potapenko A.* Additive regularization of topic models // Machine Learning. — 2015. — Т. 101, № 1—3. — С. 303—323.
11. *Berry M. W.*, *Castellanos M.* Survey of text mining // Computing Reviews. — 2004. — Т. 45, № 9. — С. 548.
12. *Harris Z. S.* Distributional structure // Word. — 1954. — Т. 10, № 2/3. — С. 146—162.
13. *Kumar B. S.*, *Ravi V.* LDA Based Feature Selection for Document Clustering // Proceedings of the 10th Annual ACM India Compute Conference on ZZZ. — ACM. 2017. — С. 125—130.

14. *Onan A.*, *Bulut H.*, *Korukoglu S.* An improved ant algorithm with LDA-based representation for text document clustering // Journal of Information Science. — 2017. — Т. 43, № 2. — С. 275—292.

15. *Xie P.*, *Xing E. P.* Integrating document clustering and topic modeling // arXiv preprint arXiv:1309.6874. — 2013.

16. *Jolliffe I.* Principal component analysis. — Springer, 2011.

17. *Aggarwal C. C.*, *Zhai C.* A survey of text clustering algorithms // Mining text data. — Springer, 2012. — С. 77—128.

18. *Sajana T.*, *Rani C. S.*, *Narayana K.* A survey on clustering techniques for big data mining // Indian journal of Science and Technology. — 2016. — Т. 9, № 3. — С. 1—12.

19. Indexing by latent semantic analysis / S. Deerwester [и др.] // Journal of the American society for information science. — 1990. — Т. 41, № 6. — С. 391—407.

20. *Lee D. D.*, *Seung H. S.* Algorithms for non-negative matrix factorization // Advances in neural information processing systems. — 2001. — С. 556—562.

21. *Lee D. D.*, *Seung H. S.* Learning the parts of objects by non-negative matrix factorization // Nature. — 1999. — Т. 401, № 6755. — С. 788.

22. *Hofmann T.* Probabilistic latent semantic analysis // Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. — Morgan Kaufmann Publishers Inc. 1999. — С. 289—296.

23. *Blei D. M.*, *Ng A. Y.*, *Jordan M. I.* Latent dirichlet allocation // Journal of machine Learning research. — 2003. — Т. 3, Jan. — С. 993—1022.

24. *Zhai C.*, *Velivelli A.*, *Yu B.* A cross-collection mixture model for comparative text mining // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. — ACM. 2004. — С. 743—748.

25. *Lesk M. E.* Word-word associations in document retrieval systems // American documentation. — 1969. — Т. 20, № 1. — С. 27—38.

26. *Landauer T. K.*, *Dumais S. T.* A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. // Psychological review. — 1997. — Т. 104, № 2. — С. 211.

27. *Zha H.* Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering // Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. — ACM. 2002. — С. 113—120.

28. Corpus-based and knowledge-based measures of text semantic similarity / R. Mihalcea, C. Corley, C. Strapparava [и др.] // AAAI. Т. 6. — 2006. — С. 775—780.

29. *Arthur D.*, *Vassilvitskii S.* k-means++: The advantages of careful seeding // Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. — Society for Industrial, Applied Mathematics. 2007. — С. 1027—1035.

30. *Cheng Y.* Mean shift, mode seeking, and clustering // IEEE transactions on pattern analysis and machine intelligence. — 1995. — Т. 17, № 8. — С. 790—799.

31. *Comaniciu D.*, *Meer P.* Mean shift: A robust approach toward feature space analysis // IEEE Transactions on Pattern Analysis & Machine Intelligence. — 2002. — № 5. — С. 603—619.

32. *Ng A. Y.*, *Jordan M. I.*, *Weiss Y.* On spectral clustering: Analysis and an algorithm // Advances in neural information processing systems. — 2002. — С. 849—856.

33. A density-based algorithm for discovering clusters in large spatial databases with noise. / M. Ester [и др.] // Kdd. Т. 96. — 1996. — С. 226—231.

34. *Zhang T.*, *Ramakrishnan R.*, *Livny M.* BIRCH: an efficient data clustering method for very large databases // ACM Sigmod Record. Т. 25. — ACM. 1996. — С. 103—114.

35. *Röder M.*, *Both A.*, *Hinneburg A.* Exploring the space of topic coherence measures // Proceedings of the eighth ACM international conference on Web search and data mining. — ACM. 2015. — С. 399—408.

36. Automatic Evaluation of Topic Coherence / D. Newman [и др.] // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Los Angeles, California : Association for Computational Linguistics, 2010. — С. 100—108. — (HLT '10). — ISBN 1-932432-65-5. — URL: http://dl.acm.org/citation.cfm?id=1857999.1858011.

37. Optimizing semantic coherence in topic models / D. Mimno [и др.] // Proceedings of the conference on empirical methods in natural language processing. — Association for Computational Linguistics. 2011. — С. 262—272.

38. *Bouma G.* Normalized (pointwise) mutual information in collocation extraction // Proceedings of GSCL. — 2009. — С. 31—40.

39. *Lau J. H.*, *Newman D.*, *Baldwin T.* Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. — 2014. — С. 530—539.

40. *Zuo Y.*, *Zhao J.*, *Xu K.* Word network topic model: a simple but general solution for short and imbalanced texts // Knowledge and Information Systems. — 2016. — Т. 48, № 2. — С. 379—398.

41. Btm: Topic modeling over short texts / X. Cheng [и др.] // IEEE Transactions on Knowledge and Data Engineering. — 2014. — Т. 26, № 12. — С. 2928—2941.

42. *Syed S.*, *Spruit M.* Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation // 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). — IEEE. 2017. — С. 165—174.

43. *Krasnov F.*, *Sen A.* The Number of Topics Optimization: Clustering Approach // Machine Learning and Knowledge Extraction. — 2019. — Т. 1, № 1. — С. 416—426.