МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

На правах рукописи

Алексеев Василий Антонович

МЕТОДЫ ОЦЕНИВАНИЯ И УЛУЧШЕНИЯ ИНТЕРПРЕТИРУЕМОСТИ, УСТОЙЧИВОСТИ И ПОЛНОТЫ ВЕРОЯТНОСТНЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

Специальность 1.2.1 —
«Искусственный интеллект и машинное обучение»

Диссертация на соискание учёной степени
кандидата технических наук

Научный руководитель:
д-р физ.-мат. наук
Воронцов Константин Вячеславович

Москва — 2024

Оглавление

Введение

Тематическое моделирование — область машинного обучения, связанная с анализом коллекции текстовых документов, целью которого является выявление *тем*, которые затрагиваются в коллекции в целом и в каждом конкретном документе коллекции в частности. Темы скрыты и заранее не известны. (Более того, не известно, что такое вообще есть тема. Точнее, понятие темы в зависимости от задачи может определяться поразному.) Таким образом, тематическая модель принимает на вход коллекцию текстовых документов, и на выходе выдаёт набор тем, где каждая тема характеризуется словами, по которым можно понять смысл темы; и информацию о том, в каких документах какие темы встречаются. В настоящее время тематическое моделирование применяется в различных областях, например в категоризации документов [1], разведочном поиске [2], биологии [3].

Идеи и гипотезы, принимаемые в тематическом моделировании, позволяют в конечном итоге свести задачу нахождения тем в документах к задаче матричного разложения, которая решается итерационным методом. Проблема в том, что задача матричного разложения *некорректно поставлена*: множество её решений бесконечно. Результат работы итерационного алгоритма зависит от начального приближения матриц — решение ещё и неустойчиво [4]. Если несколько раз обучать тематические модели на одной и той же коллекции документов, но при разной начальной инициализации, то итоговые темы могут быть разными в зависимости от модели.

В работе [5] предложен подход к обучению тематических моделей, названный аддитивной регуляризацией тематических моделей (Additive Regularization of Topic Models, ARTM). Регуляризаторы позволяют сокращать допустимое множество решений задачи матричного разложения до тех решений, которые удовлетворяют определённым свойствам. Например, с помощью регуляризаторов можно накладывать ограничение не модель, такое чтобы темы модели были различными. Но помимо того, что регуляризация используется для получения решения с заданными свой-

ствами, она служит также и для повышения устойчивости тем модели. Тем не менее, даже не смотря на применение регуляризации, неустойчивость и неполнота всё равно присущи тематическим моделям.

Отчасти вытекает из полноты и неустойчивости, но и сам по себе важен вопрос об автоматической или полуавтоматической оценке качества тематических моделей. Существующие подходы: перплексия, разреженность, чистота, когерентность — не позволяют должным образом оценить интерптерируемость тем тематической модели. Таким образом, при работе с тематическими моделями исследователю часто приходится глазами просматривать темы, что долго и не удобно. Неполнота и неустойчивость приводят к тому, что при многократном обучении моделей (например, при поиске лучших гиперпараметров) некоторые темы могут с небольшими изменениями возникать в разных моделях, некоторые могут присутствовать в одной модели, но отсутствовать в других. Таким образом, кроме (полу-)автоматической оценки качества тем, при проведении экспериментов есть необходимость в том, чтобы (полу-)автоматически выявлять, сохранять и в дальнейшем использовать интерпретируемые темы. Использовать либо при анализе тем вновь обученной модели (чтобы повторно не просматривать похожие темы), либо при обучении новой модели (чтобы уже найденные интерпретируемые темы точно в ней присутствовали).

Целью данной работы является разработка комплекса программ, позволяющего в рамках АРТМ подхода с помощью множественного обучения тематических моделей получать полные и устойчивые тематические модели.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Предложить новый вид когерентности как способа автоматической оценки качества тематических моделей, учитывающий распределение темы по всему тексту; провести эксперименты по сравнению с когерентностью по совстречаемостям самых частых слов темы.

2. Разработать комплекс программ для автоматической оценки качества тематических моделей по ряду внутренних критериев, включая новую когерентность.
3. Исследовать возможность получения полного набора тем с помощью множественного обучения тематических моделей.
4. Разработать регуляризаторы в рамках подхода к тематическому моделированию АРТМ, предназначенные для улучшения тематической модели в процессе множественного обучения.
5. Сравнить по ряду внутренних критериев качества модель, полученную с помощью множественного обучения, с другими тематическими моделями.

Основные положения, выносимые на защиту:

1. Предложена внутритекстовая когерентность как метод оценки интерпретируемости темы по распределению её слов в тексте.
2. Реализованы алгоритмы вычисления когерентности и обучения интерпретируемых тематических моделей в рамках библиотеки `TopicNet`.
3. Разработана библиотека `OptimalNumberOfTopics` для оценки качества тематических моделей по внутренним критериям, предоставляющая для использования наибольшее число критериев по сравнению с аналогичными библиотеками.
4. Представлен метод `TopicBank` оценки качества тематических моделей с учётом их неустойчивости и неполноты.
5. Предложен многопроходной алгоритм улучшения тематической модели `ITAR`, повышающий устойчивость и полноту итоговой модели по сравнению с одиночными моделями.

Научная новизна:

1. Впервые использована внутритекстовая когерентность.
2. Предложен оригинальный способ сравнения разных функций когерентности с помощью полусинтетических сегментированных данных.
3. Подготовлены и опубликованы оригинальные датасеты документов на естественном языке для обучения и оценки качества тематических моделей.

4. Впервые в рамках одной библиотеки приведены реализации большого числа внутренних критериев качества тематических моделей, включая внутритекстовую когерентность.
5. Впервые в рамках АРТМ использован регуляризатор, предназначенный для использования при множественном обучении моделей.

Теоретическая значимость заключается в развитии методологии оценки качества тематических моделей и АРТМ подхода к тематическому моделированию.

Научная и практическая значимость заключается в реализации и публикации в открытом доступе всех предложенных алгоритмов, которые могут быть использованы в различных областях, включающих анализ текстовой информации (такие как категоризация документов, информационный поиск, анализ банковских транзакций и другие).

Степень достоверности полученных результатов обеспечивается проведёнными экспериментами и публикациями.

Методология и методы исследования. Разработка программного кода производится на Python с использованием библиотеки BigARTM. Эксперименты удовлетворяют принципам воспроизводимости результатов.

Апробация работы. Основные результаты работы докладывались на следующих публичных выступлениях:

1. Внутритекстовая когерентность как мера интерпретируемости тематических моделей текстовых коллекций — 60-я Научная конференция МФТИ. 2017.
2. Intra-Text Coherence as a Measure of Topic Models' Interpretability — 24-я Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог». 2018.
3. Topic Modelling for Extracting Behavioral Patterns from Transactions Data — IC-AIAI 2019: International Conference on Artificial Intelligence: Applications and Innovations. 2019.
4. Банк тем: сбор интерпретируемых тем с помощью множественного обучения тематических моделей и их дальнейшее использо-

вание для оценки качества тематических моделей — 64-я научная конференция МФТИ. 2021.

5. Банк тем: сбор интерпретируемых тем с помощью множественного обучения тематических моделей и их дальнейшее использование для оценки качества тематических моделей — Математические методы распознавания образов (ММРО-2021).
6. Determination of the Number of Topics Intrinsically: Is It Possible? — The 11th International Conference on Analysis of Images, Social Networks and Texts (AIST 2023).
7. TopicBank: Collection of Coherent Topics Using Multiple Model Training with Their Further Use for Topic Model Validation — The 5th International Conference on Machine Learning and Intelligent Systems (MLIS 2023).
8. Determination of the Number of Topics Intrinsically: Is It Possible? — The 66th MIPT All-Russian Scientific Conference. 2024.
9. Итеративное улучшение аддитивно регуляризованной тематической модели — 66-я Всероссийская научная конференция МФТИ. 2024.
10. Iterative Improvement of an Additively Regularized Topic Model — The 12th International Conference on Analysis of Images, Social Networks and Texts (AIST 2024).

Публикации. Основные результаты по теме диссертации изложены в 8 печатных работах, 5 из которых изданы в журналах, рекомендованных ВАК [6—10], 3 работы — в тезисах докладов конференций [11—13]. Помимо этого, ещё 3 работы в данный момент находятся в печати [14—16]. Получены два свидетельства о государственной регистрации программы для ЭВМ [17; 18].

Личный вклад. В работе [6] автором предложены и реализованы функции внутритекстовой когерентности, создание же полусинтетических датасетов, вообще проектирование методики сравнения когерентностей проводились совместно с Булатовым В. Г. В работе [7] автор наравне с коллегами участвовал в подготовке данных, предложении и проверке гипотез, проведении экспериментов и анализе их результатов. В работе [8] автор отвечал за всё, связанное с когерентностью (скоры, ре-

цепты), также принимал то или иное участие в разработке и поддержке всех других частей библиотеки на зрелых и поздних этапах её существования, участвовал в подготовке и публикации датасетов. В работе [9] автор принимал ключевое участие. В работе [10] автором реализован метод поиска определения оптимального числа тем по устойчивости, проведены эксперименты по устойчивости, реализация же и исследование внутренних критериев качества тематических моделей проводились совместно с Булатовым В. Г. В работе [15] автором проведены все представленные эксперименты, математический же вывод и реализация регуляризаторов, их тестирование, "обкатка" сделаны совместно с Горбулевым А. И. Вклад автора во все основные положения, выносимые на защиту, является решающим.

Объём и структура работы. Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объём диссертации составляет 172 страницы с 32 рисунками и 14 таблицами. Список литературы содержит 176 наименований.

Краткое содержание по главам. Во *введении* обосновывается актуальность исследований по тематическому моделированию, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, сформулированы научная новизна и практическая значимость представляемой работы.

Первая глава посвящена обзору по тематическому моделированию, введению основных понятий. В *разделе 1.2* приводится пример применения тематического моделирования для анализа транзакций клиентов банка как пример использования тематического моделирования для решения прикладных задач. Сообщается об успешном применении тематического моделирования для выявления моделей поведения клиентов по этим данных об оплатах. Модели построены с помощью библиотеки `BigARTM`. Результаты демонстрируют способность подхода агрегировать информацию о моделях поведения различных групп потребителей. Анализ результатов позволяет увидеть тематические кластеры людей — например, путешественников или владельцев ипотечных кредитов. Кроме того, были изучены низкоразмерные эмбеддинги (векторные представле-

ния) клиентов, полученные с помощью тематической модели. Показано, что эти векторные представления содержат, кроме информации о покупках, также и демографическую информацию. В разделе также приводится описание лучшего способа предобработки клиентских данных перед моделированием. Глава завершается *разделом 1.3*, где приводится описание библиотеки тематического моделирования `TopicNet`, в частности, её преимуществ по сравнению с библиотекой `BigARTM`. Этот пакет на Питоне, распространяемый под лицензией MIT, нацелен на то, чтобы с помощью языка высокого уровня сделать тематическое моделирование с аддитивной регуляризацией более доступным для "неспециалистов". Возможности библиотеки включают в себя мощные методы визуализации моделей, различные стратегии обучения, полуавтоматический выбор модели, поддержку задания пользователем собственных метрик качества, модульный подход к обучению тематических моделей.

Вторая глава посвящена исследованию внутренних критериев качества тематических моделей вообще и когерентности в частности. *Раздел 2.2* посвящён задаче измерения интерпретируемости и когерентности (меры согласованности) тематических моделей. Предлагается новый, внутритекстовый, подход к оценке меры согласованности темы. Вычислительные эксперименты проводятся на коллекции научно-популярного контента "ПостНаука". В *разделе 2.3* отмечается, что число тем — один из самых важных параметров тематической модели. Представленно исследование по изучению большого числа способов определения числа тем, в применении к разным тематическим моделям, на нескольких общедоступных наборах данных. В результате продемонстрировано, что внутренние критерии качества тематических моделей далеко не всегда являются надёжными и точными способами оценки "оптимального" числа тем. Показано, что количество тем в датасете зависит и от метода поиска числа тем, и от используемой тематической модели — а не является абсолютным свойством конкретного корпуса текстов. Делается вывод о необходимости разработки других методов для решения проблемы о неизвестном исходном числе тем. Предлагается несколько перспективных направлений для дальнейших исследований.

Третья глава посвящена исследованию возможности решения проблем неустойчивости и неполноты тематических моделей с помощью множественного обучения тематических моделей. Среди недостатков тематических моделей в *разделе 3.2* отмечается их нестабильность в том смысле, что итоговые темы могут зависеть от случайной начальной инициализации модели, и неполнота в том смысле, что новые запуски тематических моделей на одной и той же коллекции могут давать новые темы. Это приводит к тому, что анализ данных с помощью тематического моделирования обычно требует очень большого числа экспериментов, включающих оценку качества множества тематических моделей, просмотр их тем, настройку гиперпараметров — в поисках модели, которая бы описывала данные наилучшим образом. Как способ преодолеть нестабильность и неполноту тематических моделей, предлагается постепенно (в процессе множественного обучения тематических моделей) накапливать интерпретируемые темы в "банке тем". При добавлении новых тем в банк используется двухуровневая тематическая модель, затем анализируется связь дочерних тем (кандидатов на добавление в банк) с родительскими (темами банка), с тем чтобы исключить нерелевантные или дублирующие темы, а не добавлять их в банк. Вводится новый способ оценки качества тематической модели, путём сравнения тем, найденных моделью, с темами, которые были предварительно собраны в банке тем для данного датасета. Эксперименты с несколькими коллекциями документов и тематическими моделями показывают, что предложенный метод помогает в поиске модели с наибольшим числом интерпретируемых тем. В *разделе 3.3* представлен метод тематического моделирования с использованием отложенных тем. В (полу-)автоматическом режиме определяется принадлежность темы, найденной тематической моделью, к одной из трёх категорий: хорошая (интерпретируемая), плохая (неинтерпретируемая), или "никакая" (интерпретируемая, но нерелевантная на данном этапе исследования). Основная задача состоит в улучшении базовой модели в процессе нескольких переобучений — улучшении, которое заключается в выделении новых хороших тем при сохранении всех уже найденных ранее отложенных хороших тем и уменьшении числа плохих и нерелевантных тем. Предлагаемое решение основано на применении новых регуляри-

заторов типа сглаживания и декоррелирования в рамках подхода АРТМ. Вычислительный эксперимент проводится на ряде текстовых коллекций естественного языка.

Благодарности. Автор хотел бы выразить благодарность некоторым людям. Не только за помощь непосредственно в научной работе. Но также и за “просто так”: за какую-то роль в жизни, участие, которое он помнит и которое наверняка повлияло на то, кем он в итоге стал, что ценит и как смотрит на вещи.

Семье — за всё, и, в частности, за то, что помогали быть в тонусе при написании диссертации.

Тае — за возможность понаблюдать, поразмышлять и за бесконечные прогулки, часто по местам, куда нормальным людям путь закрыт.

Воронцову Константину Вячеславовичу — за интересные идеи, увлечённость, терпение, профессионализм и способность укладываться в сроки, несмотря на загруженность другими делами (и на порой опасную близость “внезапно” возникающего дедлайна).

Белоусовой Любови Владимировне — за уроки в стенах-академиках вообще и за “тетрадки для правил” в частности.

Бабаян Ларисе Семёновне — за то, что это она была (и остаётся для автора до сих пор) “лучиком света” в школе.

Попову Владимиру Анатольевичу — за энергию, юмор и за то, что всё успевал.

Михайловой Юлии Викторовне — за один из лучших уроков физики (даже несмотря на то, что данных для точного расчёта полёта ракеты оказалось недостаточно).

Тер-Крикорову Александру Мартыновичу — за то, что провёл их через тёмный интегральный лес по тропинке, заросшей одуванчиками.

Артемьевой Ольге Андреевне — за то, что столько времени хранила у себя его “чёрточки”.

Бажиной Таисии Александровне — за то, что помогла ему вспомнить ту часть себя, о которой забыл.

Янковской Екатерине Александровне — за то, что учебники были не для них, и за “тет-а-тет”.

И, наконец, группе Heart и “ГлубокП”-у — за то, что помогли автору пережить один из самых трудных, по его ощущениям, периодов в подготовке диссертации.

Глава 1. Вероятностное тематическое моделирование

Тематическое моделирование — это область статистического анализа текстов [19]. Точнее, это совокупность автоматических способов анализа коллекции текстовых документов, которые направлены на обнаружение *скрытой тематической структуры* в больших коллекциях текстов.

Тематические модели используются в информационном поиске [20], категоризации документов [1], анализе данных социальных сетей [21; 22], рекомендательных системах [20; 23], разведочном поиске [24] и других областях. В результате обработки коллекции документов тематическая модель выдаёт набор тем, затрагиваемых в документах, распределение этих тем в документах и слова, характеризующие каждую тему [19].

При обработке данных естественного языка исследователь имеет дело с коллекцией *документов*, каждый из которых состоит из последовательности слов, или *токенов*. При работе с большими коллекциями документов полезно группировать их по смыслу. Такой тип кластеризации называется *тематическим моделированием* и строит скрытое измерение *тем*, которые обеспечивают *краткое* описание для каждого документа в коллекции.

Таким образом, тематическое моделирование — это метод извлечения из корпусов текстов скрытых вероятностных распределений на словах, называемых темами. Изначально тематическое моделирование было разработано для работы с большими коллекциями документов. Однако, будучи изначально созданными для поиска скрытых тем в текстовых документах, тематические модели доказали свою актуальность в широком спектре задач [25]. Фактически, они были применены к любым последовательным данным: будь то корзины с покупками в продуктовом магазине, поведение в Интернете, банковские транзакции. Какими бы далекими друг от друга ни казались на первый взгляд эти области, все они могут быть представлены в терминах тематического моделирования.

Но одно дело — как смотреть на данные. Другое — собственно *причина* использования тематического моделирования. В течение долгого времени тематические модели использовались в основном для двух целей:

либо для кластеризации данных в целях их последующего анализа и извлечения информации о структуре коллекции, либо для создания эмбеддингов слов по коллекциям текстов для, например, проведения рекомендаций и поиска. В последнем от исследователя требуется на размеченных наборах данных обучить рекомендательные или поисковые модели, и задача может быть сведена к классификации, где в качестве векторов признаков используются эмбеддинги документов, полученные с помощью тематической модели. Таким образом, тематические модели в данном контексте конкурируют с, например, способами получения эмбеддингов слов и документов с помощью глубокого обучения.

Тематическое моделирование не обязательно показывает лучшие результаты с точки зрения метрик классификации, но оно обладает тем преимуществом, что полученные с его помощью результаты являются *интерпретируемыми*. Каждая тема может быть представлена как распределение вероятностей на множестве слов, а каждая компонента в эмбеддинге документа имеет значение вероятности того, что документ относится к определённой теме. Это свойство тематических моделей делает их более подходящими для тех областей анализа данных, где важна ясность предсказания, или где необходимо вручную корректировать нежелательные смещения, вносимые данными в модель (например, занизить вероятности определённых слов, или, наоборот, отметить, что некоторые документы должны относиться к одной теме).

Следует отметить, что тематическое моделирование по-прежнему является ценным инструментом в арсенале исследователей в области обработки естественного языка (Natural Language Processing, NLP). В настоящее время ведётся работа по расширению современных моделей глубокого обучения для учета тематической важности слов при, например суммаризации текстов [26—28] и понимании текстов песен [29—31]. Видно, что тематическое моделирование и его вариации продолжают играть важную роль во многих задачах по обработке естественного языка.

В отличие от получения эмбеддингов слов, сочетаний слов и документов, задача по выявлению скрытой тематической структуры коллекции в некотором роде уникальна для области тематического моделирования. При этом исследователь не обязательно хочет оптимизировать

какую-нибудь метрику классификации, а просто ищет ответы на вопросы о структуре и природе коллекции. Такой подход, например, используется в биологии [3; 32] и гуманитарных науках [25; 33]. Он позволяет получить бесценное представление о больших данных, которые в противном случае могли бы остаться незамеченными исследователем. Тематическое моделирование в данном случае предоставляет уникальную возможность по-особому взглянуть на данные, увидеть то, что иначе могло бы остаться незамеченным для исследователя.

Итак, тематическое моделирование помогает находить темы. Однако само *понятие темы* может быть определено по-разному в зависимости от поставленной задачи. В статистическом тематическом моделировании (о котором и идёт речь в работе) каждая тема рассматривается как вероятностное распределение по словам в словаре. В других областях понятие темы может определяться не с помощью словаря частот слов в тексте, а на основе понятия *связности* текста. Например, в теории дискурса тема — это главный участник или главная мысль на протяжении всего связного дискурса или диалога [34]. Возвращаясь к статистической точке зрения, тематическая модель помогает получить сжатое представление каждого документа набором его тем, представляя каждый документ как распределение вероятностей на множестве тем, а каждую тему — как распределение вероятностей на множестве слов. Это может быть полезно в различных областях, например, при категоризации документов [1], для рекомендательных систем [20], разведочного поиска [2], анализа данных социальных сетей [21]. Больше идей и возможных применений тематического моделирования представлено в обзоре [35].

Представим подробнее *нотацию*, которая используется в работе, и с её помощью сформулируем задачу статистического тематического моделирования.

Пусть D означает коллекцию (множество) текстовых документов, а W — *словарь*, то есть множество всех слов, или терминов, или токенов, которые встречаются в документах коллекции. Иногда, в зависимости от контекста, мы также будем обозначать и количество документов также через D, а размер словаря через W. Термин из W может быть отдельным словом или фразой. После согласования того, какие сущности в коллекции

следует рассматривать как токены, каждый документ $d \in D$ представляется как упорядоченная последовательность n_d терминов $W_d \subseteq W$. Пусть n_{dw} означает количество раз, сколько термин $w \in W$ встречается в документе $d \in D$.

Теперь предположим, что каждый термин в каждом документе соответствует некоторой *скрытой* (латентной) *теме* из конечного множества тем T. (Опять же, далее, в зависимости от контекста, T может означать и количество тем). Назовём это предположение *гипотезой о существовании тем*. В тематическом моделировании темы (как правило) заранее не известны. И цель тематического моделирования — найти это множество T (вообще, разобраться с тем, что же вообще такое темы, которые ищем — это отдельный вопрос, но в статистическом тематическом моделировании за тему принимается вероятностное распределение на словах, поэтому о другом каком бы то ни было "смысле" за понятием тема более здесь задумываться не будем). Каждая тема $t \in T$ описывается частотами слов и распределением этих слов в документах коллекции. Коллекцию текстов можно рассматривать как выборку троек $\{(d_i, t_i, w_i)\}_{i=1}^{n}$, полученных независимо из дискретного распределения $p(d, t, w)$ на пространстве $D \times T \times W$. (Причём документы D и слова W — наблюдаемые переменные, а темы T — скрытые.)

В тематическом моделировании распространена практика рассматривать каждый документ как *неупорядоченный* набор слов. Такой формат представления данных называется "мешком слов", или *Bag-of-Words*. Согласно гипотезе "мешка слов", порядок расположения токенов в документе не влияет на темы и процесс поиска тем. Хотя такое предположение и кажется сильным упрощением действительности, в нём есть смысл, потому что, например, обе последовательности слов "Джек Лондон американский писатель автор известных приключенческих романов" и "автор Лондон известных романов Джек писатель приключенческих американский" — хоть и отличаются друг от друга последовательностью слов (причём вторая очевидно представляет из себя что-то слабо связное, невразумительное), но основная тема — литература — в обоих примерах угадывается верно. В качестве продолжения этой гипотезы можно также принять гипотезу "мешка документов", которая гласит, что темы не должны

меняться от в зависимости того, в каком порядке мы подаем документы в модель. (Однако если коллекция документов динамическая, то есть меняется со временем, например, если в неё поступают новые документы, то дообучение тематической модели на такой коллекции уже будет зависеть от порядка приходящих вновь документов.)

Также в тематическом моделировании принимается гипотеза *условной независимости*, которая гласит, что вероятность принадлежности слова теме не зависит от того, в каком документе это слово расположено:

$$p(w \mid d, t) \equiv p(w \mid t) \tag{1}$$

Наконец, предполагается, что наблюдаемая коллекция порождается распределениями $p(w \mid t)$ и $p(t \mid d)$. Согласно закону полной вероятности и предположению об условной независимости (1):

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d) = \sum_{t} \phi_{wt} \theta_{td} \tag{2}$$

где $\phi_{wt} \equiv p(w \mid t)$ и $\theta_{td} \equiv p(t \mid d)$.

Таким образом, вероятностная тематическая модель (2) описывает, каким образом образуются документы коллекции D — как смесь распределений θ_{td} и ϕ_{wt}. Обучение тематической модели — это обратная задача, когда необходимо найти эти самые распределения θ_{td} и ϕ_{wt}, зная коллекцию D. Первое распределение — это распределение “тем в документе” ($\theta_{td} = p(t \mid d)$, то есть как столбец стохастической матрицы Θ *вероятностей тем в документах* размера $T \times D$):

$$\Theta \equiv (\theta(t \mid d))_{T \times D} \equiv (\theta_{td})_{T \times D} \tag{3}$$

Второе распределение — это распределение “слов в теме” ($\phi_{wt} = p(w \mid t)$, то есть как столбец стохастической матрицы Φ *вероятностей слов в темах* размера $W \times T$):

$$\Phi \equiv (\phi(w \mid t))_{W \times T} \equiv (\phi_{wt})_{W \times T} \tag{4}$$

Нахождение по коллекции D распределений θ_{td} и ϕ_{wt} является целью тематического моделирования. Эта задача эквивалентна нахождению при-

ближённого представления матрицы частот слов в документах $F = \big(\hat{p}(w \mid d)\big)_{W \times D}$ — где $\hat{p}(w \mid d) = \frac{n_{dw}}{n_d}$, а n_{dw} есть количество вхождений слова w в документ d — в виде произведения $F \approx \Phi\Theta$ двух неизвестных матриц Φ и Θ:

$$\Phi = (\phi_{wt})_{W \times T} \qquad \phi_{wt} = p(w \mid t)$$
$$\Theta = (\theta_{td})_{T \times D} \qquad \theta_{td} = p(t \mid d)$$

Все упомянутые матрицы F, Φ, и Θ являются стохастическими: их колонки f_d, ϕ_t, и θ_d, соответственно, неотрицательны и нормированы — представляют дискретные вероятностные распределения.

Под *темой* в вероятностном тематическом моделировании понимается как раз по сути распределение на множестве слов $p(w \mid t)$, $w \in W$. Но тема также характеризуется и её распределением на множестве документов $p(t \mid d)$, $d \in D$, с помощью которого можно понять, в каких документах тема затрагивается (а по этой информации можно и визуально оценить качество темы, насколько она понятна человеку). Как правило, число тем $|T|$ *выбирают* намного меньше, чем размер коллекции $|D|$ и размер словаря $|W|$. Таким образом, в тематическом моделировании решается задача *низкорангового* стохастического матричного разложения [36].

Тематическая модель выявляет скрытую тематическую структуру коллекции и находит декомпозицию каждого документа по набору представленных в нём тем (1).

Отметим отдельно несколько понятий, которые отчасти уже использовались и которые часто в дальнейшем будут использоваться в работе:

Определение 1.0.1. Везде в работе под *хорошей* темой имеется в виду интерпретируемая тема, такая, по списку самых частых слов которой человеку понятно, про что она, какую область жизни описывает.

Определение 1.0.2. Слова *коллекция текстовых документов*, *коллекция документов*, *датасет*, *корпус текстов*, *текстовая коллекция* в пределах работы считаются равнозначными.

Определение 1.0.3. Под *самыми частыми словами*, *топовыми словами*, *топ-словами* для темы t понимается одно и то же, а именно слова, соответ-

ствующие первым k вероятностям в отсортированном по убыванию списке $\{p(w \mid t)_w\}$. Как правило, число первых позиций k не уточняется, но под ним подразумевается некоторое очень небольшое число по сравнению с общим количеством слов в словаре, например 10, 20, 50.

1.1 Аддитивная регуляризация тематических моделей

В фундаментальной работе [37] представлена одна из самых первых, и в то же время одна из самых простых и понятных тематических моделей — модель *Probabilistic Latent Semantic Analysis* (PLSA), — где распределения (2) обучаются *максимизацией логарифма правдоподобия с линейными ограничениями*.

Правдоподобие — это вероятность наблюдаемых данных как функция параметров модели Φ и Θ. В силу предположения о независимости, эта вероятность эквивалентна произведению вероятностей всех слов во всех документах:

$$p(\Phi, \Theta) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}} = \prod_{d \in D} \prod_{w \in d} p(w \mid d)^{n_{dw}} p(d)^{n_{dw}} \to \max_{\Phi, \Theta}$$

Если взять логарифм, то выражение выше превратится в сумму, а члены, которые не зависят от параметров модели, можно отбросить, поскольку они не играют роли при оптимизации. Таким образом мы получим задачу максимизации логарифма правдоподобия при линейных ограничениях на неотрицательность и нормированность (суммируются в единицу) столбцов матриц Φ и Θ (так как столбцы представляют распределения вероятностей):

$$\overbrace{\ln p(\Phi, \Theta)}^{\mathscr{L}(\Phi, \Theta)} = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w \mid d) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta} \tag{5}$$

$$\sum_{w \in W} \phi_{wt} = 1, \phi_{wt} \geq 0 \quad \sum_{t \in T} \theta_{td} = 1, \theta_{td} \geq 0 \tag{6}$$

В работах [5; 8; 38; 39] предложен и продвигается подход к обучению тематических моделей под названием “Аддитивная регуляризация тематических моделей”, или APTM (Additive Regularization of Topic Models, ARTM), который с помощью *регуляризаторов* — аддитивных добавок к оптимизируемому логарифму правдоподобия — предоставляет возможность реализовать в тематических моделях различные свойства (регуляризаторы позволяют сократить возможное множество решений задачи матричного разложения до тех решений, которые удовлетворяют определенным условиям, ограничениям). Например, с помощью регуляризаторов можно потребовать от модели, чтобы её темы отличались друг от друга, или чтобы каждая тема имела лишь небольшое число наиболее вероятных слов (чтобы вероятность была распределена по словам сильно неравномерно), или, наоборот, чтобы некоторые темы имели как можно больше слов с ненулевой вероятностью — так называемые *фоновые* темы, содержащие слова общей лексики. Или, если коллекция D несбалансирована, с помощью регуляризации можно потребовать от модели, чтобы её темы были разного размера [40]. Кроме того, что регуляризация используется для получения решения с желаемыми свойствами, она также повышает устойчивость тем моделей.

Сильной стороной подхода ARTM является то, что каждый аддитивный член регуляризации приводит к простой аддитивной модификации М-шага. Таким образом, аддитивная регуляризация APTM выражается в максимизации логарифма правдоподобия (5) со взвешенной суммой регуляризаторов $R_i(\Phi,\Theta)$:

$$\mathscr{L}(\Phi,\Theta) + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta} \tag{7}$$

относительно ограничений (6), где τ_i есть неотрицательные коэффициенты регуляризации, которые позволяют изменять силу влияния отдельных регуляризаторов.

Условия Каруша – Куна – Таккера дают необходимые условия для локального экстремума задачи (7) с ограничениями в виде системы уравнений. Точка локального экстремума может быть найдена, например, мето-

дом простой итерации, который эквивалентен EM-алгоритму и находит локальный максимум задачи (5), обновляя на каждой итерации матрицы Φ и Θ [5; 38; 39]:

$$\begin{cases} p_{tdw} = \underset{t \in T}{\text{norm}}(\phi_{wt}\theta_{td}) \\ \phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

где $p_{tdw} = p(t \mid d, w)$, norm — функция нормировки вектора ($\text{norm}_{w \in W} n_{dw} = f_d$), $n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}$, $n_{td} = \sum_{w \in d} n_{dw} p_{tdw}$. Видно, что каждый новый регуляризатор (7) в конечном счёте выражается в виде аддитивной добавки на M-шаге.

При этом в начале выполнения алгоритма нужно каким-то образом *инициализировать* матрицу Φ — и способ удачной инициализации является отдельным вопросом. Обычно же используют случайную инициализацию.

Возможность наложения дополнительных функциональных ограничений (регуляризация) позволяет улучшить качество модели в различных задачах. Этот подход предоставляет возможность формулировать ограничения задачи в математической форме. Более того, в рамках ARTM могут быть реализованы и известные тематические модели PLSA и LDA.

В случае *мультимодальных данных*, APTM можно обобщить путём оптимизации взвешенной суммы лог-правдоподобий (7):

$$\sum_{m,d} \sum_{w \in W_m} \tau_m n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta} \tag{8}$$

Иерархическое тематическое моделирование — это ответвление тематического моделирование, где *родительская тема* из множества T предполагается смесью *дочерних тем* S, при $|S| > |T|$ [41]:

$$\underbrace{p(w \mid t)}_{\phi_{wt}^{parent}} = \sum_{s \in S} \underbrace{p(w \mid s)}_{\phi_{ws}^{child}} \underbrace{p(s \mid t)}_{\psi_{st}} \tag{9}$$

Другими словами, $\Phi^{parent}_{W\times T} = \Phi^{child}_{W\times S}\Psi_{S\times T}$, где $\Psi_{S\times T}$ есть матрица связи между темами двух уровней: родительскими и дочерними. Если переформулировать такое отношение между темами в нотации ARTM, то окажется, что родительские темы можно рассматривать как псевдодокументы с частотами слов, равными n_{wt} [41; 42].

Представим некоторые известные APTM резуляризаторы.

Сглаживание и разреживание Темы T тематической модели можно разделить на темы двух типов: предметные S и фоновые B, то есть $T = S \sqcup B$.

Предметные темы S — специализированные и состоят из предметных слов. Предполагается, что такие темы разреженные и слабо связанны между собой.

Фоновые темы B состоят из слов общей лексики и равномерно распределены по документам коллекции.

Для выделения тем обоих типов получения и используется разреживающий регуляризатор. Его идея в том, чтобы сделать распределения ϕ_t и θ_d далёкими (в случае разреживания) или близкими (в случае сглаживания) к равномерным распределениям β_t и α_d:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in H}\sum_{w\in W} \beta_{wt}\ln\phi_{wt} + \alpha_0 \sum_{t\in H}\sum_{d\in D} \alpha_{td}\ln\theta_{td} \to \max_{\Phi,\Theta} \tag{10}$$

где для разреживания надо положить $H \equiv S$ и взять $\beta_0, \alpha_0 < 0$; а для сглаживания $H \equiv B$ и $\beta_0, \alpha_0 > 0$ (β_0 и α_0 есть коэффициенты регуляризации, которые находятся экспериментально).

Декорреляция Желаемое свойство тем тематической модели — их непохожесть друг на друга [5]. Поэтому декоррелирующий регуляризатор увеличивает суммарное попарное расстояние между темами:

$$R(\Phi) = -\frac{\tau}{2}\sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W} \phi_{wt}\phi_{ws} \to \max_{\Phi} \tag{11}$$

Поскольку в качестве параметра регуляризатора используется только матрица Ф, регуляризатор оказывает влияние только на М-шаге при обновлении матрицы Ф.

Добавка от регуляризатора на М-шаге будет следующей:

$$\phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = -\tau\phi_{wt}\sum_{s\in T\setminus t}\phi_{ws}$$

Выбор тем Регуляризатор, представленный в [43] максимизирует KL-дивергенцию между $p(t) = \sum_d p(d)\theta_{td}$ и равномерным распределением на темах, таким образом устраняя несущественные и линейно зависимые темы:

$$R(\Theta) = \frac{n}{|T|}\sum_{t\in T}\ln\sum_{d\in D}p(d)\theta_{td} \to \max \tag{12}$$

1.2 Пример прикладного применения тематического моделирования

1.2.1 Введение

В наши дни люди совершают огромное количество транзакций в повседневной жизни. Поэтому банки ежедневно собирают массу информации об истории транзакций своих клиентов. Применение такой информации может включать, в частности, улучшение качества обслуживания, привлечение новых клиентов, увеличение доходов, разработку ценовых стратегий, прогнозирование вероятности возврата кредита и лучшее понимание потребностей клиента. За последние несколько лет алгоритмы машинного обучения преуспели в решении различных задач и потенциально могут помочь в решении других [44]. Анализ транзакций относится к числу таких задач, поскольку данные о транзакциях велики и неструктурированы, и методы машинного обучения помогают извлекать из них определенную информацию. Многие решения в этой области были раз-

работаны с их использованием [45—47]. В основном приложения в этой области сосредоточены на таких областях, как профилирование потребителей [48—50], оценка покупательских профилей и прогнозирование покупок [51], или обнаружение групп клиентов по данным для лучшего планирования продуктовой кампании [52]. Хотя профилирование клиентов может быть выполнено с помощью различных алгоритмов кластеризации [53], результат подобных усилий обычно довольно грубый и плохо интерпретируемый для такой чувствительной области, как банковское дело. Покупательные профили и прогнозирование покупок — широко обсуждаемая тема из-за ее непосредственного применения. Было предпринято несколько попыток связать психологические профили пользователей и данные об их транзакциях с помощью модели LDA [54] или помочь маркетинговым кампаниям в прогнозировании потребления клиентов с помощью нейронных сетей и методов случайного леса [55].

Подход, решающий все эти проблемы одновременно, предполагает построение векторного пространства клиентских эмбеддингов. Некоторые авторы пошли по этому пути [50], используя автокодировщики. В данном разделе мы сосредоточимся на построении низкоразмерного векторного пространства с помощью тематического моделирования. Тематическое моделирование является мощным инструментом, который успешно применялся для решения различных задач машинного обучения [56]. Изначально оно было предложено как метод обработки естественного языка [37; 57]. В настоящее же время тематическое моделирование применяется и к нетривиальным данным, таким как анализ логов [48; 58], определение модели поведения по видео [59]. И далее в разделе будут приведены результаты применения тематического моделирования для анализа данных банковских транзакций.

Эксперименты будут проводиться с использованием библиотеки APTM [5]. Она включает в себя такие популярные тематические модели, как PLSA [37], LDA [57], а также обобщения. APTM подход к тематическому моделированию реализован в быстрой и эффективной библиотеке BigARTM [39], которая была успешно применена для решения различных задач [60]. Еще одной особенностью ARTM является возможность использования мультимодальных данных [2]. В нашем случае это очень важ-

но, поскольку информация о транзакциях состоит из различных данных о клиенте: потраченной суммы денег, кода категории торговца (merchant category code, MCC), пола, возраста.

В целом использование именно тематических моделей для решения задачи анализа банковских транзакций обусловлено рядом причин: тематическая модель не является “черным ящиком” и легко интерпретируется; APTM позволяет удовлетворить множество критериев, предъявляемых бизнесом к ML-решениям; сама тематическая модель быстро применяется и обучается, что подходит для практических приложений.

Далее изложение материала раздела построено по следующему плану. В следующем подразделе рассматривается связанная теорию тематического моделирования. Далее приводится информация об используемом наборе данных и обработке данных. Затем идут экспериментальные результаты. И в конце подводятся итоги и обсуждается вклад работы.

При обработке естественного языка исследователь имеет дело с коллекцией *документов*, состоящей из последовательностей слов, или *токенов*. В банковских транзакциях мы имеем дело с коллекцией *историй транзакций клиентов*, состоящей из последовательностей транзакций, описываемых датой, MCC-кодом и суммой, потраченной по этому коду. Таким образом, можно читать клиентов как книгу (см. рис. 2), применяя тематическую модель к их истории транзакций. Результатом этого стало бы латентное *пространство эмбеддингов*, которое представляет типы потребления, полученные из статистики данных о транзакциях. *Темы* предоставляют характеристику любого клиента разбивкой его на типы потребления, описывающих его **интерпретируемым** образом.

Если рассматривать MCC-код транзакции как токен, то частоту токена можно положить равной сумме, потраченной на этот код в документе, представляющим историю транзакций клиента.

Или, более формально, обозначая коллекцию транзакций клиентов как D, индивидуальную историю транзакций клиента как $d \in D$, и коллекцию всех возможных в D транзакций как W, с каждой транзакцией, обозначаемой как $w \in W$, мы можем следующим образом интерпретировать в рассматриваемой области гипотезы, принимаемые в тематическом моделировании:

– *Гипотеза мешка слов*. Порядок продуктов в счёте из продуктового магазина не изменит того, как этот счёт будет категоризован моделью.
– *Гипотеза условной независимости*. Каждый токен, описывающий клиента, появляется в его истории транзакций благодаря некоторому шаблону поведения, общему для многих клиентов, а не благодаря именно этому конкретному клиенту.

Классификация на основе профиля клиента Определённое ранее множество $D = \{d_i\}_{i=1}^n$ — это множество клиентов банка, а $Y = \{y_i\}_{i=1}^n$ — множество их целевых характеристик. Мы стремимся построить модель $f(\cdot \mid \alpha) : D \to Y$, использующую распределение $p(t \mid d)$ профилей клиентов в качестве начального пространства признаков, максимизируя точность:

$$\frac{1}{n} \sum_i \Big[f\big(p(t \mid d_i) \mid \alpha\big) = y_i \Big] \to \max_\alpha . \tag{13}$$

1.2.2 Подготовка данных

Данные для экспериментов представлены полями: идентификатор клиента, время транзакции, сумма транзакции, код транзакции. Кроме того, есть таблицы, содержащие дату рождения и пол клиента. Данные предоставлены индустриальным партнёром и потому не подлежат огласке.

В процессе работы получилось обогатить имеющиеся данные дополнительно иерархией MCC-кодов, объединив похожие MCC-коды в кластеры и присвоив им читаемые метки, такие как, например, “такси”, “АЗС” и другие. В ходе предварительной обработки было применено два метода для кодирования суммы, потраченной в каждой транзакции. Во-первых, сумма использовалась в качестве частоты терминов для каждой транзакции. Таким образом больше потрачено на MCC-код — чаще эта транзакция встречается в профиле пользователя. Однако это может привести к искажению, когда редкая, но дорогая покупка может учитываться так же, как

и множество обычных дешёвых транзакций. Поэтмоу приходим ко второму способу учёта суммы по транзакции — чтобы компенсировать описанное выше искажение и обогатить словарь, каждый MCC-код разбивается на квантили, то есть вводятся новые токены. Эти токены кодируют MCC-код транзакции и квантиль потраченной суммы: ниже среднего, средний, выше среднего.

Наконец, иерархия MCC-кодов вводится в качестве новых модальностей. В данной работе используются эмбеддинги, содержащие только MCC-коды и модальность малых групп (дискретных групп трат по MCC-кодам). Другие модальности используются для проверки адекватности полученной тематической модели — MCC-коды из действительно удаленных и некоррелированных групп не должны находиться в одной теме. Например, модель с темами, содержащими коды из групп “Путешествия” и “Ремонт дома”, будет отброшена в процессе обучения.

1.2.3 Эксперимент

Проверяется способность предлагаемого подхода создавать векторные представления клиентов банка на том же уровне, что и реальные данные. Наряду с основной целью проверяется влияние предварительной обработки данных на решение нашей задачи. Для обучения стандартной тематической модели необходимо определить важнейшие гиперпараметры, включающие: количество тем, количество шагов EM-алгоритма (количество итераций обучения), коэффициенты регуляризации — с помощью оценки когерентности модели, которая коррелирует с интерпретируемостью [6]. Для сравнения различных тематических моделей основные гиперпараметры фиксируются: количество шагов EM-алгоритма и количество тем. Коэффициенты же регуляризации настраиваются в соответствии с выбранной метрикой. В процессе исследований, было обнаружено, что наилучшие модели получаются при количестве тем около 30, и потому этот гиперпараметр был зафиксирован для всех тематических моделей в данной работе. В таблице 1 приведены примеры темы, связан-

ных с “Отпуском”, где вместо MCC-кодов для наглядности используются соответствующие интерпретируемые группы MCC-кодов.

Таблица 1 — Темы про “Отпуск” (Vacation).

Группа трат	**Вероятность**
Plane tickets	0.575
Duty-free	0.177
Theatres	0.0094
Hotels	0.049
Attractions	0.0038
Drug stores	0.0022
Car sharing	0.009
Пол	**Вероятность**
Мужчина	0.393
Женщина	0.607
Возрастная группа	**Вероятность**
17-23	0.138
24-35	0.442
36-54	0.376
55+	0.043

Внимательный читатель мог бы заметить, что вероятность расходов в таблице 1 не равна единице. Это связано с длинным хвостом в распределении, который не отражает основной темы, и потому не показан. Итоговую кластеризацию клиентов, выполненную первой тематической моделью, можно увидеть на рисунке 3.

Эксперимент построен следующим образом: создаётся базовое векторное представление клиентов на основе данных об их транзакциях, и векторное представление от тематической модели. Затем оценивается точность предсказания пола и возраста с помощью модели CatBoost [61]. Эмбеддинги тематической модели корректируются на обучающем наборе данных. Результаты для различных эмбеддингов представлены в таблице. 2.

В таблице 2 модели с тегом “DemoOpt” оптимизированы для лучшей производительности в предсказании пола и возраста, а модели с тегом “Coherence” были оптимизированы по показателю когерентности, что сделало их более интерпретируемыми [6]. Как видно, удалось полу-

Таблица 2 — Сравнение моделей для разных типов предобработки.

Тип модели	***Точность по полу***	***Точность по возрасту***	**Микро f1-скор**
Plain MCC	0.719	0.452	0.424
ARTM Coherence	0.658	0.405	0.377
ARTM DemoOpt	**0.728**	**0.457**	**0.435**
Categorized MCC	**0.730**	**0.465**	**0.439**
ARTM Coherence	0.665	0.439	0.414
ARTM DemoOpt	0.685	0.45	0.427

чить эмбеддинги, которые оказались лучше, чем исходные предварительно обработанные данные. При этом не удалось повторить это для данных, разделённых по категориям расходов, однако при этом, тем не менее, удалось получить результаты лучше, чем у стандартной модели.

1.2.4 Заключение

Описан подход к построению тематических моделей, которые могут дать лучшее понимание транзакционных данных. Этот подход решает проблему полезности и интерпретируемости векторных эмбеддингов клиентов в банковском деле. Полученные гиперпараметры модели были выбраны в соответствии с внешними метриками, что даёт возможность настроить полученные эмбеддинги так, чтобы они также содержали информацию, не представленную в реальных данных. Показано, что векторные представления, предоставляемые моделью, сохраняют все важные особенности, позволяющие рассказать пользователю историю о клиенте, при этом сохраняя высокую интерпретируемость, недоступную для большинства более эффективных ML-моделей.

Как видно из таблицы 2, не только тематическое моделирование, но и простая предварительная обработка, описанная в разделе, может дать толчок в нужном направлении. Правильно настроенная тематическая модель показала лучшие результаты по сравнению с некатегоризированны-

ми данными, но не смогла превзойти базовый уровень для категоризированного кода MCC. Это может быть связано с одним из двух факторов. Во-первых, ограничение на количество тем в модели могло оказаться решающим, так как словарь увеличивался в результате предварительной обработки. Во-вторых, возможно, архитектура модели не позволяет учесть некоторые особенности данных, которые стали очевидны для классификатора CatBoost на категоризированных данных.

В будущем можно попытаться улучшить модель так, чтобы она учитывала совстречаемости транзакций, также важно будет справиться с сильным дисбалансом, который наблюдался в использованном наборе данных.

1.3 TopicNet

Jordan Boyd-Graber, Yuening Hu и David Mimno в своей монографии [25] показывают, что разработка новой тематической модели является сложным предприятием, и определяют доступность как наиболее важную нерешённую проблему тематического моделирования: “The primary research challenge of topic models is… to make them more accessible”.

В том же году в другой публикации [62] утверждалось, что существующие тематические модели не предоставляют обычным пользователям (не экспертам) прямых средств для изменения темы, если она их чем-то не устраивает. В исследовании предложено несколько улучшений существующего пользовательского опыта в использовании тематического моделирования, сконцентрированных на визуализации тем для пользователей.

Статья другой группы [63] демонстрирует, что популярные тематические модели не дают хороших результатов при использовании “как есть”, со значениями параметров по умолчанию.

С тем чтобы облегчить исследования по поиску скрытых тем в текстовых коллекциях, представляется программный пакет TopicNet[1,2].

[1]Исходный код: github.com/machine-intelligence-laboratory/TopicNet.

[2]Документация: machine-intelligence-laboratory.github.io/TopicNet.

TopicNet способствует повышению гибкости проектирования тематических моделей и предоставляет мощный опыт работы “из коробки”, повышая доступность тематического моделирования для широкой аудитории.

1.3.1 Связанные работы

Существующие алгоритмы тематического моделирования берут свое начало от латентного семантического анализа — LSA [64], — который, по сути, представляет собой разложение по сингулярным значениям с наименьшим возможным рангом для матрицы размера $D \times W$. Позднее основное место в области тематического моделирования занял вероятностный подход. Вероятностную тематическую модель можно рассматривать как “чёрный ящик”, который получает на вход коллекцию текстовых документов, а на выходе выдает два семейства распределений: вероятности терминов для каждой темы $\phi_{wt} = p(w \mid t)$ и вероятности тем для каждого документа $\theta_{td} = p(t \mid d)$. Матрица Φ размера $W \times T$ и матрица Θ размера $T \times D$ являются параметрами модели, которые должны быть найдены в процессе обучения.

Развитие вероятностного моделирования тем началось с двух фундаментальных моделей: вероятностного латентного семантического анализа — pLSA [37] — и латентного распределения Дирихле — LDA [57]. LDA есть по существу pLSA с предположением о принадлежности вероятностных распределений слов в темах и тем в документах распределению Дирихле, каждое со своим параметром концентрации. Обычно в каждом из документов реальной текстовой коллекции затрагивается всего несколько тем. А каждая тема обычно сконцентрирована в относительно небольшом множестве слов. Это свойство разреженности и отражается в параметре концентрации распределения Дирихле. Таким образом, разреженность распределений Дирихле является вероятностным инструментом, который воплощает эту интуицию о свойствах текстовых коллекций. Однако, согласно статье [65], LDA требует обширной оптимизации гиперпараметров для получения хороших результатов.

За последние годы появилось много расширений моделей pLSA и LDA, каждое из которых учитывает различные особенности данных, характерные для той или иной задачи, и обеспечивает желаемые свойства решения. Начиная с модели LDA, байесовское обучение является стандартом де-факто в тематическом моделировании. В этом подходе сначала описывается вероятностная генеративная модель данных, задаются априорные распределения параметров модели, а затем с помощью байесовского вывода получаются апостериорные распределения параметров. Байесовский вывод в каждом случае уникален, и, следовательно, уникальна реализация каждой новой модели.

Аддитивная регуляризация для тематического моделирования — ARTM [66; 67] — это, напротив, не байесовский многоцелевой подход. Он основан на максимизации логарифма правдоподобия коллекции вместе со взвешенной суммой критериев регуляризации. Многие известные байесовские тематические модели могут быть переформулированы как регуляризованные в рамках APTM модели PLSA. После такой переформулировки они обычно становятся гораздо проще для понимания и реализации. Подход ARTM позволяет комбинировать тематические модели, просто добавляя или убирая регуляризаторы. Это привело к появлению модульной технологии тематического моделирования, реализованной в программном обеспечении с открытым исходным кодом BigARTM [68].[3]

При наличии генеративной модели и данных необходимо выполнить вывод для получения вероятностных распределений, зависящих от темы. Существует много алгоритмов вывода: EM-алгоритм (expectation-maximization algorithm), семплирование по Гиббсу (Gibbs sampling), вариационный вывод (variational inference), градиентный спуск (gradient descent) и передача сообщений (message passing). В ARTM для обучения параметров модели используется регуляризованный EM-алгоритм. Сходство между каждым из упомянутых алгоритмов было отмечено в [69].

Gensim [70] — один из самых популярных NLP фреймворков для тематического моделирования. Он реализует несколько популярных моделей, таких как LSA, pLSA, LDA, Hierarchical Latent Dirichlet Allocation (HLDA) и их производные. Также в Gensim реализовано вычисление ко-

[3] bigartm.org.

герентности для улучшения тем из ранее упомянутых моделей, как это сделано в статье [71]. Фреймворк написан на языке Python и оптимизирован для работы с большими массивами документов.

Stanford Topic Modelling Toolbox — TMT [72], основанный на Scala, содержит модели LDA, Labelled LDA и PLDA, доступные для обучения. Stanford TMT также включает в себя удобное взаимодействие с Excel, позволяя загружать данные из ячеек Excel и генерировать богатый вывод для отслеживания использования слов по темам, по времени и другим критериям.

MALLET [73] — этот известный фреймворк по тематическому моделированию реализован на Java. Содержит эффективные, основанные на семплировании по Гиббсу реализации моделей LDA, Pachinko Allocation [74] и HLDA. MALLET предоставляет поддержку улучшенных моделей LDA и часто используется для онлайн-сервисов [75].

Библиотека моделирования тем коротких текстов (Short Text Topic Modelling) — STTM [76] — это Java-фреймворк, который расширяет диапазон доступных методов моделирования с открытым исходным кодом и объединяет самые современные модели алгоритмов моделирования тем коротких текстов. В первую очередь он нацелен на выделение значимых тем из коротких текстов, поэтому многие высокопроизводительные модели, такие как мультиномиальная смесь Дирихле (Dirichlet Multinomial Mixture, DMM) [77], Word Network Topic Model (WNTM) [78], Pseudo-Document-Based Topic Model (PTM) [79] и Self-Aggregation-Based Topic Model (SATM) [80] представлены в этом фреймворке. Некоторые тематические модели для длинных текстов, такие как LDA и Latent Feature Model with LDA [81], также есть в STTM.

Familia [82] — это фреймворк, реализующий различные тематические модели, включая, но не ограничиваясь, LDA и Supervised LDA: Topics Over Time (TOT) [83], Bilingual Topic Model [84], Location-Aware Topic Model (LATM) [85] и некоторые другие, используя семплирование по Гиббсу в качестве “математического движка”. Несмотря на заявления авторов о том, что Familia предоставляет возможность “конструировать собственные тематические модели”, на момент проведения настоящей работы в репозитории проекта [86] не было найдено таких свидетельств.

Как уже было сказано, одной из основных проблем в области тематического моделирования является доступность моделей для широкой аудитории. Каждому из ранее упомянутых фреймворков удалось устранить этот пробел на момент их выпуска, но в силу природы байесовского подхода наиболее популярные фреймворки (Gensim, MALLET) предоставляют устаревшие модели.

Другая проблема заключается в построении сложных, легко реализуемых и многоцелевых тематических моделей с нуля. Хотя базовая модель может быть реализована за разумное время, ее улучшение остается как трудоемкой, так и подверженной ошибкам задачей [82].

В результате данной работы хочется не только сократить разрыв между новыми и популярными моделями, как это уже в своё время делали предшественники, но и предоставить инструмент, который позволит всем желающим с лёгкостью строить свои новые типы тематических моделей. Благодаря формализму ARTM, TopicNet предлагает сообществу специалистов по обработке естественного языка доступ к мультимодальному тематическому моделированию на основе Python, которому доступна обработка и больших документов, и огромных коллекций документов.

1.3.2 Основа проекта

После внимательного рассмотрения всех перечисленные подходов к тематическому моделированию, было принято решение построить фреймворк, который позволит улучшить работу пользователей с существующей библиотекой BigARTM. Ниже будут рассмотрены плюсы Python API библиотеки BigARTM, а также её минусы, с которыми как раз будет цель разобраться в рамках работы над TopicNet.

Сильные стороны BigARTM BigARTM — это быстрая и гибкая библиотека для тематического моделирования [87], основанная на ARTM формализме [66]. Идея ARTM заключается в замене оптимизируемого правдоподобия на регуляризованное правдоподобие и оптимизации этого ново-

го функционала с помощью модифицированного EM-алгоритма. Регуляризация служит двум целям. Во-первых, она обеспечивает устойчивость и ограничивает область решений. Во-вторых, каждый регуляризационный член используется для достижения различных характеристик решения, таких как разреженность, разнообразие, связность, или же для учёта при обучении модели какой-то дополнительной информации. В качестве примера работы со вспомогательной информацией можно привести учёт метаданных документа (например, авторов, временны́х меток, тэгов и n-грамм). Это связано с тем, что правдоподобие каждой дополнительной модальности можно рассматривать как регуляризатор, применяемый к тематической модели над словами соответствующей модальности.

Некоторые работы, использующие преимущества ARTM и BigARTM, включают: улучшение качества разведочного поиска [88], обучение интерпретируемых эмбеддингов слов с помощью WNTM [89], иерархическое тематическое моделирование [90], категоризация текста с несколькими метками [91], улучшение тем для векторного представления документов в задачах регрессии [92], поиск редких этнически релевантных тем в социальных медиа [93; 94], включение языковых особенностей [60], выбор тем с помощью регуляризации энтропии [95], улучшение тем через сегментацию текста [96], прямое улучшение когерентности тем [97], оставление гипотезы “мешка слов” для использования новой меры внутритекстовой связности [6]. В обзоре [67] показано, как байесовские тематические модели могут быть переформулированы гораздо проще с точки зрения ARTM, включая мультимодальные, многоязычные, темпоральные, иерархические, основанные на графах и тематические модели на коротких текстах.

Разработка теории ARTM и программного обеспечения BigARTM все ещё продолжается. Многие из существующих широко используемых регуляризаторов были предложены сообществом.

На момент проведения работы, среди других фреймворков для тематического моделирования только Familia предлагает сопоставимую гибкость. Однако возможность построения собственной тематической модели или даже её обучения на заданном корпусе отсутствует в открытом исходном коде Familia [86].

Слабые стороны BigARTM При наличии точной спецификации регуляризованной модели BigARTM может обучить её быстрым, масштабируемым и эффективным способом. Однако неясно, откуда берётся сама спецификация. В то время как выставление количества тем является нерешённым вопросом во многих фреймворках тематического моделирования, проблема в BigARTM усложняется возможностью сочетания множества различных регуляризаторов, каждый из которых имеет неизвестный индивидуальный коэффициент регуляризации. И BigARTM не предлагает практически никаких рекомендаций по выбору регуляризаторов и их структурных параметров.

Дополнительным фактором, обуславливающим высокий входной барьер, является несколько неудобный и непоследовательный API в библиотеке BigARTM. Это естественный результат внедрения новой функциональности до того, как сложилась “лучшая практика” ее использования, и невозможности изменить API впоследствии из-за соображений обратной совместимости. Поэтому по мере того, как приложения BigARTM становились все более разнообразными, а алгоритмы постепенно совершенствовались, высокоуровневый интерфейс BigARTM становился всё менее приспособленным для “лучших практик”.

Ещё одним недостатком библиотеки BigARTM является сложность её расширения. С технической точки зрения, библиотека BigARTM состоит из ядра, написанного на C++, и нескольких классов-обёрток для Python. Низкоуровневые процедуры C++ многопоточны и хорошо оптимизированы, что даёт BigARTM существенное преимущество в производительности. В то же время, это затрудняет любую модификацию низкоуровневой функциональности. Между тем, высокоуровневый API не всегда обеспечивает достаточную гибкость, чтобы, например, экспериментировать с новыми пользовательскими регуляризаторами.

1.3.3 Видение проекта

Основная цель TopicNet — свести к минимуму разрыв по части пользования библиотекой между неспециалистами и опытными пользователями. Это не означает, что обе группы будут использовать библиотеку одинаково. Скорее, это означает, что обе группы должны иметь возможность общаться друг с другом, понимать друг друга. Сформулированы следующие требования, необходимые для достижения этой цели:

- Модульность: должна быть возможность использовать лишь небольшую часть функциональности TopicNet в качестве “плагина” в независимо существующем проекте. Причина за этим требованием двоякая. Первая заключается в том, что это должно облегчить внедрение TopicNet: опытные пользователи не будут вынуждены кардинально менять свои существующие проекты, чтобы начать получать преимущества от TopicNet. Второй момент связан с созданием сообщества разработчиков вокруг TopicNet: модульные проекты с открытым исходным кодом более приветливы к соавторам.
- Инструменты визуализации: библиотека должна предоставлять готовые к использованию мощные инструменты визуализации. Такие инструменты играют важную роль в анализе ошибок и могут быть полезны для последующих задач (например, исследовательского поиска). В соответствии с предыдущим требованием, этот модуль должен быть как можно более автономным; в идеале, он должен позволить сообществу включить лучшие практики в TopicNet.
- Краткость: библиотека должна избавить пользователя от низкоуровневых деталей, освободив время для решения серьёзных (и более интересных) задач. Причин для такого требования было три. Первая заключается в необходимости создать лучшие условия для в каком-то смысле художественного процесса построения целевой функции под конкретную задачу. Второй момент — следование философии “соглашение по конфигурации” (“convention

over configuration”): уменьшая количество явно объявленных вещей и предоставляя разумные значения по умолчанию, можно внедрить “лучшие практики”, такие как автоматическое сохранение обученных моделей или раздельное хранение батчей данных для разных наборов данных. Последнее — читабельность: при работе с кратким и единообразным кодом пользователь видит на экране больше строк кода, что способствует более быстрому и в целом лучшему прочтению и пониманию эксперимента. В результате гораздо проще делиться готовыми экспериментами, просматривать и отлаживать их.

– Работа “из коробки”: библиотека должна включать в себя несколько готовых к использованию и дающих хорошие результаты пайплайнов (готовых схем, заготовок) обучения. Более того, эти пайплайны должны включать в себя самые известные подходы для как можно большего числа задач моделирования. Положительный опыт “из коробки” очень важен для привлечения новых пользователей, в то время как опытные пользователи могут поделиться своим опытом с помощью подготовленных пайплайнов.

1.3.4 Архитектура

TopicNet состоит из двух больших модулей: `Viewers` и `Cooking Machine`.

Цель модуля `Viewers` — предоставить хорошие инструменты визуализации. Дизайн модуля соответствует философии Unix: каждый Viewer имеет ограниченную область ответственности, и по умолчанию возвращает результат в виде JSON-конвертируемого объекта. Таким образом, модуль `Viewers` обладает высокой степенью компонуемости (composability) и модульности, но при этом имеет удобные методы, возвращающие `pandas.DataFrame` или HTML.

Модуль `Cooking Machine` содержит весь инструментарий моделирования, воплощенный в полуиерархической структуре основных классов

моделирования. Эти классы отвечают за построение и обучение тематической модели заданной структуры, за выбор моделей в соответствии с различными ограничениями, а также за сохранение, загрузку и ведение журнала сообщений в процессе моделирования.

Из принципа “соглашение по конфигурации”, сделаны некоторые предположения о том, какие эксперименты должен поддерживать TopicNet. Считается, что пайплайн обучения моделей может быть представлен в виде дерева. Каждый узел которого — это тематическая модель, а направленные рёбра представляют собой отношения между родительскими и дочерними моделями, например, “модель Y была получена из модели X с помощью преобразования T_{XY}”. Допустимые преобразования ограничиваются, связываются с их положением (глубиной) в дереве экспериментов: налагается требование, чтобы каждое ребро одного уровня описывало одно и то же преобразование, за исключением набора индивидуальных параметров.

Неполный список таких преобразований:

- Применение регуляризатора с произвольным коэффициентом регуляризации или изменение параметров существующего регуляризатора.
- Обучение модели в течение нескольких итераций.
- Добавление в модель тем, найденных на другом корпусе.

Класс `Experiment` отвечает за хранение, протоколирование и поддержание всей структуры эксперимента.

Преобразования между моделями разных уровней Эксперимента связаны с экземплярами класса `Cube`. Каждый Куб выступает в качестве плана для всех преобразований модели на текущем этапе эксперимента. Таким образом, о пайплайне обучения тематической модели можно думать как о нескольких последовательно сложенных друг за другом кубах.

Куб отвечает за две важнейшие функции. Первая — *спецификация*: на этапе инициализации куб преобразует заданные пользователем параметры в многомерное пространство поиска. Вторая — *изменение*: имея точку в пространстве поиска и тематическую модель, куб изменяет один или несколько гиперпараметров модели. Таким образом, Куб действует как “инкубатор” для моделей, что отражено в названии класса. Схему процес-

са обучения с двумя кубами, применяемыми к модели, можно увидеть на рисунке 4.

Вместе взятые, классы `Experiment` и `Cube` делают более чёткими, краткими и доступными сложные пайплайны обучения тематических моделей и протоколирование эксперимента. С тем чтобы получить большее от такого решения, реализован модуль `config_parser`, который позволяет задавать сложные обучающие пайплайны обучения в виде обычного текстового конфигурационного файла в формате YAML.

Ещё одна важная область, которую обычно описывают очень многословно и которая может представляться довольно запутанной — это выбор модели. В реальных экспериментах не каждая модель имеет потомков; большинство моделей отбрасываются на основе некоторых критериев. Помимо дорогостоящей ручной проверки топ-слов и топ-документов тем, другие традиционные критерии включают автоматически вычисляемые перплексию и когерентность. Библиотека BigARTM предоставляет ещё несколько других метрик, таких как разреженность, чистота, контрастность [98]. С тем чтобы облегчить бремя ручной проверки моделей и их тем по нескольким критериям, был реализован простой язык выбора модели (для примера см. рисунок 5). Он упрощает задачу выбора модели за счёт удобного использования и возможности комбинирования различных метрик в одном критерии отбора.

1.3.5 Заключение

В данном разделе предложен конфигурируемый и быстрый фреймворк TopicNet для тематического моделирования, демонстрируются его преимущества перед конкурентами. TopicNet предоставляет широкие функциональные возможности, такие как построение моделей с нуля, богатая настройка моделей и возможность дообучения.

Библиотека предоставляет задокументированные и готовые к использования пайплайны (“рецепты”) тематического моделирования, отражающие наиболее известные практики построения ARTM-моделей для

определённой задачи. ARTM-модель способна превзойти большинство популярных тематических моделей в плане предоставления связных и разнообразных тем. С помощью TopicNet исследователи-инженеры могут получить от BigARTM ещё больше пользы, лучше настроить её, применяя собственные регуляризаторы и находя лучшие значения гиперпараметров в случае многокритериального сценария обучения. В совокупности это позволяет пользователю библиотеки построить свою собственную тематическую модель (укладывающуюся в АРТМ идеологию).

Помимо инструментов для гибкой и быстрой разработки моделей, библиотекой предоставляется возможность контролировать качество модели. TopicNet обеспечивает оценку модели с помощью множества встроенных критериев и позволяет добавлять пользовательские критерии, которые будут использоваться для отслеживания характеристик модели в процессе обучения. Библиотека поддерживает множество инструментов визуализации, как традиционных, таких как просмотр топ-токенов и топ-документов, так и новых, экспериментальных.

Более того, представляемый фреймворк поставляется в виде пакета с открытым исходным кодом и имеет потенциал для дальнейшего расширения. Философия дизайна модулей `Viewers` и `Cooking Machine` позволяет сообществу включать новые разработки в библиотеку TopicNet.

Автор работы надеется, что этот фреймворк будет одинаково ценен и для инженеров-программистов, и для исследователей в области цифровых гуманитарных наук.

Рисунок 1 — Пример фрагмента текста — статьи научно-популярного ресурса “ПостНаука”. Можно проследить процесс тематизации этого фрагмента тематической моделью: после 10, 20 и 50 итераций обучения модели (обновлений матрицы Ф).

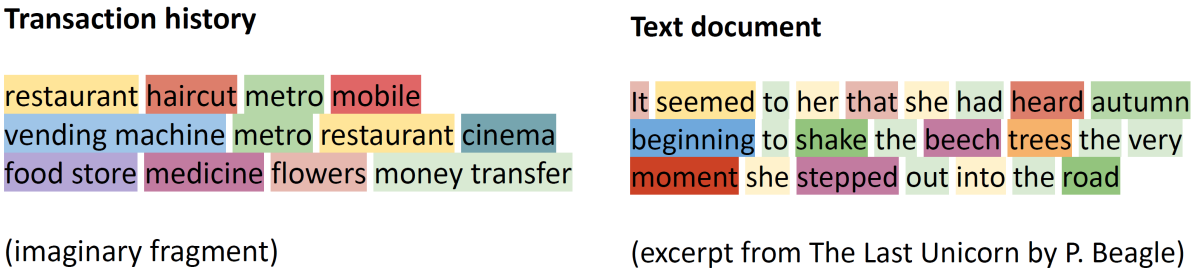


Рисунок 2 — История транзакций — как текстовый документ! Поэтому можно использовать подходы NLP для анализа транзакций.

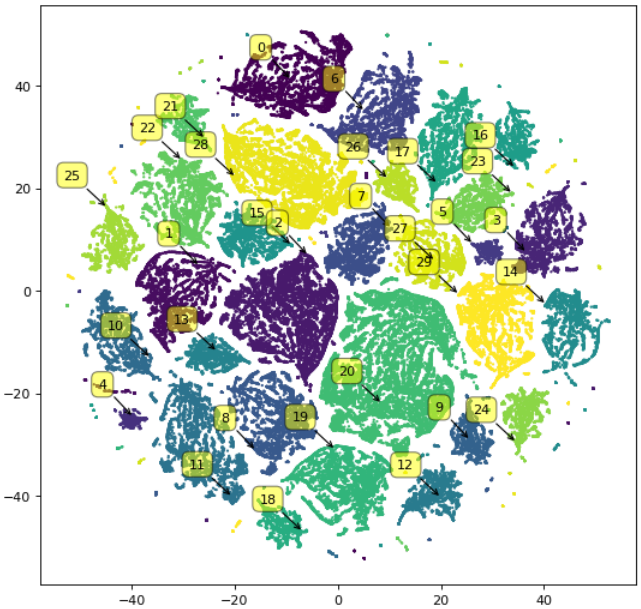


Рисунок 3 — Визуальное представление профилей потребления.

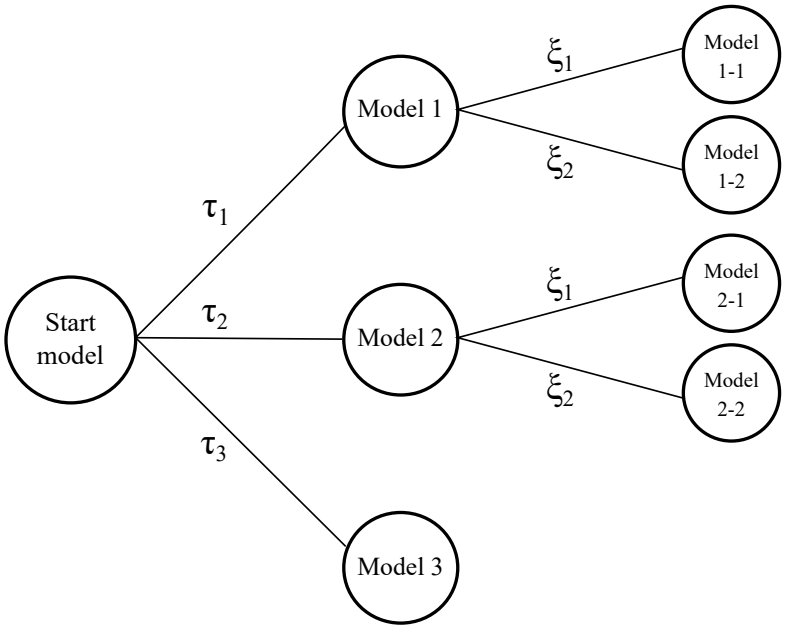


Рисунок 4 — Пример схемы двухэтапного эксперимента в рамках фреймворка тематического моделирования TopicNet. На первом этапе применяется регуляризатор с параметром τ, принимающим значения в некотором диапазоне $\{\tau_1, \tau_2, \tau_3\}$. Лучшими моделями после первого этапа являются *Модель 1* и *Модель 2* - так что *Модель 3* больше не участвует в процессе обучения. Второй этап связан с другим регуляризатором с параметром ξ, принимающим значения в диапазоне $\{\xi_1, \xi_2\}$. В результате этого этапа получаются две модели-потомка *Модели 1* и две модели-потомка *Модели 2*.

```
TopicKernel@word.average_contrast > 0.95 * MAXIMUM(TopicKernel@word.average_contrast)
   and PerplexityScore@all < 1.1 * MINIMUM(PerplexityScore@all)
   and SparsityPhiScore@word -> max
   COLLECT 3
```

Рисунок 5 — Приведённое выражение возвращает три модели, которые находятся в 5% лучших по контрастности, имеют приемлемый уровень перплексии и ещё максимально разрежены. `SparsityPhiScore` обозначает долю нулей в распределении $\phi_{wt} = p(w \mid t)$.

```
from topicnet.cooking_machine.recipes import (
    ARTM_baseline as config_string
)

dataset_path = '/data/datasets/NIPS/dataset.csv'

specific_topics   = [f'spc_topic_{i}' for i in range(19)]
background_topics = [f'bcg_topic_{i}' for i in range( 1)]

config_string = config_string.format(
    dataset_path=dataset_path,
    modality_list=['@word'],
    main_modality='@word',
    specific_topics=specific_topics,
    background_topics=background_topics
)
experiment, dataset = (
    build_experiment_environment_from_yaml_config(
        yaml_string=config_string,
        experiment_id='sample_config',
        save_path='sample_save_path'
    )
)
experiment.run(dataset)
```

Рисунок 6 — Пример базового эксперимента в рамках TopicNet.

Глава 2. Оценка качества тематических моделей

Интерпретируемость — это важное свойство хорошей тематической модели [99]. Считается, что тема хорошо интерпретируется, если она соответствует интересующей человека концепции реального мира. Однако темы, полученные с помощью тематических моделей, далеко не всегда могут быть чёткими и понятными — например, они могут включать слова из разных слабо связанных областей [100] (не соответствуя, таким образом, никакому понятному человеку понятию).

Относительно недавно была представлена автоматически вычисляемая процедура оценки интерпретируемости темы — *когерентность* [101; 102]. Она оценивает тему по списку её наиболее часто встречающихся слов (точнее, по тому, как эти частые слова расположены друг относительно друга в тексте), и, по результатам исследований, согласуется с оценкой интерпретируемости темы по тому же списку её самых частых слов экспертами.

Однако этот подход страдает от нескольких фундаментальных ограничений. Утверждается, что эти ограничения ставят под сомнение общепринятую практику рассмотрения когерентности и интерпретируемости как эквивалентных.

Цель данного раздела работы двояка. Первая — очертить круг проблем, присущих традиционным представлениям о когерентности. Ключевая проблема существующего подходя заключается в том, что при сведении тематической модели к короткому списку слов теряется слишком большой объём информации о теме. Таким образом, оценка качества темы по когерентности не может считаться надёжной. В предыдущих исследованиях, связывающих когерентность и интерпретируемость, это не учитывалось. Доля текста, занимаемая наиболее частыми словами, никак не контролируется. В разделе показывается, что на практике эта доля слишком мала, чтобы считать когерентность надёжным эквивалентом интерпретируемости.

Вторая цель представленного в текущем разделе исследования — продемонстрировать осуществимость альтернативного подхода к оценке

интерпретируемости, который будем звать *внутритекстовой когерентностью*, определяемой как среднее тематическое сходство токенов темы (не только самых частых), близко расположенных в тексте. Для обоснования этой новой меры будет адаптирована процедура, используемая в [103], [100] и [101].

2.1 Внутренние критерии качества

Оценка качества тематической модели без внешней метрики (связанной с какой-либо другой задачей) — это нетривиальный вопрос. Часто в таких случаях качество каждой тематической модели оценивается человеческими экспертами. Однако для сценария, когда обучается много тематических моделей, оценка асессорами становится дорогой, а также трудно осуществимой в короткие сроки, и потому не может быть частью последовательного процесса эксперимента. В качестве решения этой проблемы исследователи обращаются к внутренним критериям качества тематических моделей, которые в идеале должны коррелировать с человеческими оценками качества. “Внутренние” означает, что они вычисляются на той же коллекции, на которой происходило обучение тематической модели, без привлечения какой бы то ни было дополнительной информации. Далее перечислим некоторые внутренние критерии качества, которые будут использоваться в экспериментах.

Перплексия Как способ оценки качества всей тематической модели в целом исследователи часто используют *перплексию* (“недоумение”) [39; 104]. Перплексия связана с правдоподобием коллекции $\mathscr{L}(\Phi, \Theta)$ следующим образом:

$$\text{ppl}(\Phi, \Theta) = e^{-\mathscr{L}(\Phi,\Theta)} \tag{14}$$

то есть перплексия модели тем ниже, чем выше правдоподобие $\mathscr{L}(\Phi, \Theta)$ — чем лучше модель “подстроилась” под данные, под видимое распределение слов в текстах.

Чистота и контраст В отличие от перплексии, которое дает некоторую общую оценку качества модели, можно также задаться целью оценить какие-либо качества отдельных тем, составляющих модель. Приведём два возможных способа оценить качество темы на основе матрицы “слов в темах” Ф [5] — чистота и контраст темы:

$$\text{purity}(t) = \sum_{w \in W_t} p(w \mid t)$$

$$\text{contrast}(t) = \frac{1}{|W_t|} \sum_{w \in W_t} p(t \mid w)$$

где $W_t = \{w \in W \mid p(w \mid t) > \text{threshold}\}$ — есть так называемое *ядро темы* — слова, на которых вероятность темы больше некоторого порога threshold.

Обе метрики оценивают качество темы по информации о ней в Ф матрице [5]. Естественный способ выбрать параметр threshold в формулах чистоты и контрастности — задать его равным $1/|W|$: то есть так, чтобы ядро состояло только из тех слов, вероятность которых выше вероятности при равномерном распределении [105].

Кластеризация При оценке качества тематической модели возможно также использовать ряд метрик, обычно связанных с сетевым и кластерным анализом, в частности, коэффициент Силуэта (Silhouette Coefficient, *SilhC*) и индекс Калинского – Харабаша (Calinski – Harabasz Index, *CHI*) [106].

Информационно-теоретические Байесовский информационный критерий (Bayesian Information Criterion, *BIC*) позволяет определить баланс между степенью обученности модели под данные и её сложности. Исследователями также используется принцип минимальной длины описания (Minimum Description Length, *MDL*) [107] и принцип минимальной длины сообщения (Minimum Message Length, *MML*) [108].

Вычисление этих метрик проходит следующим образом. Во-первых, мы находим $\mathscr{L}(\Phi,\Theta)$ — правдоподобие модели. Во-вторых, нам нужно узнать количество свободных параметров N_p, и оно может быть вычислено двумя различными способами: по размерности Ф как $(W-1)\cdot T$ или

по количеству ненулевых записей Φ, которое мы обозначим через #Φ (заметим, что в [109] утверждается, что количество свободных параметров в LDA и разреженных моделях должно рассматриваться по-разному). Следующие выражения позволяют вычислять как разреженные, так и неразреженные варианты метрик:

$$\begin{aligned} \text{AIC} &= 2N_p - 2\mathscr{L} \\ \text{BIC} &= N_p \log(D) - 2\mathscr{L} \\ \text{MDL} &= N_p \log(TD) - 2\mathscr{L} \end{aligned} \tag{15}$$

Когерентность В вероятностном тематическом моделировании *тема* определяется как вероятностное распределение на словах. Например, тема “Театр” может представлять собой распределение вероятностей, сосредоточенное на таких словах, как “актёр”, “спектакль”, “премьера”, “партер” и “зритель” (напротив, вероятность таких слов, как, например, “кредит” и “беспозвоночные”, будет крайне мала или даже равна нулю).

Тематическая модель может быть описана с помощью двух распределений: $\phi_{wt} = p(w \mid t)$, вероятности принадлежности слова w теме t; и $\theta_{td} = p(t \mid d)$, вероятности найти тему t в документе d.

В ранних работах по тематическому моделированию его рассматривали как промежуточный этап информационно-поискового пайплайна. Возможность осмысленной интерпретации тем рассматривалась уже после. Для автоматического измерения качества тем, так чтобы успешно имитировалась оценка её качества человеком, было предложено несколько метрик.

В настоящее время исследователи сходятся во мнении, что оценка интерпретируемости темы должна соответствовать следующей схеме:

1) Выбор некоторого небольшого набора слов для каждой темы (обычно это список из десяти наиболее часто встречающихся её слов, но возможны и более сложные подходы [110]). Для обозначения этого подмножества слов используется термин *топ-слова* (*top tokens*).

2a) Предоставление этого набора человеку-эксперту для получения его оценки качества (понятно ли по этим топ-словам, о чём тема).

2b) или Сбор по тексту статистик совстречаемости этих топ-слов, и выполнение вычислений с этими статистиками.

Эта схема была представлен в фундаментальных работах [100; 101; 103] и затем значительно развита сообществом тематического моделирования. Мы будем называть эту обширную категорию вычисляемых по топ-словам и их совстречаемостям метрик — метриками на основе топ-токенов. Основная их привлекательность заключается в простоте: вместо того чтобы оценивать всё распределение вероятностей темы, исследователю достаточно просмотреть короткий список её наиболее “репрезентативных” слов.

Однако в этом же кроется и внутреннее ограничение такого подхода к оценке интерпретируемости. Список из пяти-десяти-двадцати лучших слов отражает лишь *часть* всего распределения вероятностей и плохо отражает, насколько хорошо тематическая модель представляет конкретный корпус текстов.

Автором утверждается, что список наиболее частых слов не может считаться вполне адекватным инструментом для обоснования качества тематической модели и её тем независимо от метода анализа этого списка слов. Это в равной степени относится как к оценкам экспертов-людей, так и к автоматизированным процедурам, основанным на подсчёте совстречаемостей слов (когерентностям).

В работах [100—102] предложен способ оценки качества тем, названный *когерентностью* тем: когда решение о качестве темы выносится на основе того, как часто пары слов из самых частых слов темы (например, из десяти или двадцати слов) встречаются в тексте рядом друг с другом (по сравнению с количеством раз, когда в тексте встречаются одно и другое слово не обязательно вблизи другого). Математическое выражение для

описываемого понятия:

$$\begin{aligned} \text{coh}(t) &= \underset{w_i, w_j \in \text{top}_k(t)}{\text{mean}} \text{PMI}(w_i, w_j) \\ \text{PMI}(w_i, w_j) &= \log_2 \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \end{aligned} \tag{16}$$

где $\text{top}_k(t)$ есть множество топ k слов темы t, а $P(w_i)$ и $P(w_i, w_j)$ — вероятности встретить в тексте одно слово w_i или два слова w_i, w_j вместе (например, в окне слов небольшого размера, или в одном предложении, в одном абзаце, или просто в одном документе, или в небольшом окне в одном документе). Вероятности оцениваются с помощью известного распределения слов в документах (при этом важно, чтобы текст был с естественным порядком слов, а не как “мешок слов”). В качестве текста для оценки таких вероятностей может выступать как исследуемая коллекция текстов, так и сторонняя (например, Википедия).

Когерентность по своей природе — это попытка автоматизировать человеческий способ оценки интерпретируемости темы. Как человек понимает, интерпретируемая тема или нет? Если по составляющим её словам (и по распределению этих слов в документах коллекции) можно понять, что это за тема, описывает ли она какую-то реальную известную человеку область жизни.

Ньюманом, помимо (16), было предложено ещё несколько вариантов когерентностей [102]. Но в результате сравнения лучшей была выбрана именно (16) — оказалось, что она хорошо коррелирует с человеческими оценками интерпретируемости. Схема сравнения когерентностей использовалась следующая (см. рисунок 7). Выбиралась коллекция текстов D_0 для калибровки внутреннего критерия. Далее, строилась тематическая модель Φ_0, Θ_0. Эксперты-асессоры оценивали качество тем (рейтингами или интрузиями). И среди всех когерентностей выбиралась та, значения качества тем по которой лучше всего коррелировали с экспертными оценками интерпретируемости. В итоге когерентность (16) дала максимальную корреляцию Спирмена с оценками экспертов. Более того, эта корреляция оказалась близка к “золотому стандарту” — средней корреляции Спирмена между оценками разных экспертов.

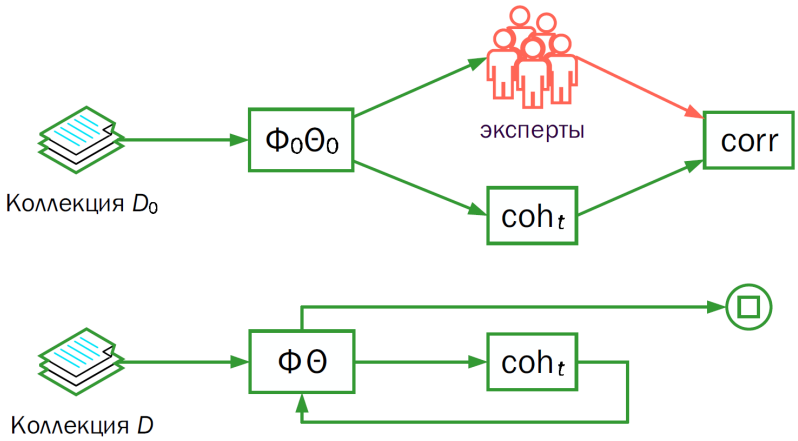


Рисунок 7 — Схема эксперимента Ньюмана. Сверху: поиск лучшей когерентности (по корреляции с экспертными оценками). Снизу: применение найденной когерентности для автоматической оценки качества тем модели.

2.2 Внутритекстовая когерентность

2.2.1 На пути к лучшей оценке интерпретируемости

Как уже отмечалось, традиционные метрики когерентности состоят из двух этапов: во-первых, они используют информацию из распределения $p(w \mid t)$; во-вторых, они извлекают статистику совстречаемостей.

Идея автоматических мер когерентности заключается в том, чтобы выяснить, как часто определённые слова появляются вместе в скользящем контекстном окне, и сравнить это число с частотой, предсказанной чистым совпадением. Тема считается когерентной, если расположение слов в ней имеет тенденцию к группировке и не выглядит случайным. Это напоминает лингвистический феномен текстовой связности [111]: предложения текстов на естественном языке связаны друг с другом с помощью синтаксических и лексических средств, таких как повторы слов, синонимы/почти-синонимы, гипонимы и так далее.

Выдвигается предположение, что тексты на естественном языке разбиты на связные сегменты, которые содержат лишь небольшое количество латентных тем. Согласно этому предположению, цель тематического

моделирования следует понимать как адекватную сегментацию исходного текста на тематически однородные фрагменты, состоящие из небольшого числа тем.

Отметим, что частые совстречаемости топовых слов — это лишь косвенный признак того, что тема представлена в коллекции текстов в виде целостных фрагментов.

Поэтому автором текущей работы утверждается, что интерпретируемость темы должна оцениваться не только по согласованности появления в тексте лишь её топовых слов, но и по согласованности появления *всех* слов темы во всём тексте (см. рисунок 8). И тогда можно получить автоматическую меру интерпретируемости модели, исследуя степень нарушения её словами описанной согласованности, исследуя степень отклонения распределения слов темы в тексте от гипотезы о сегментной структуре текста.

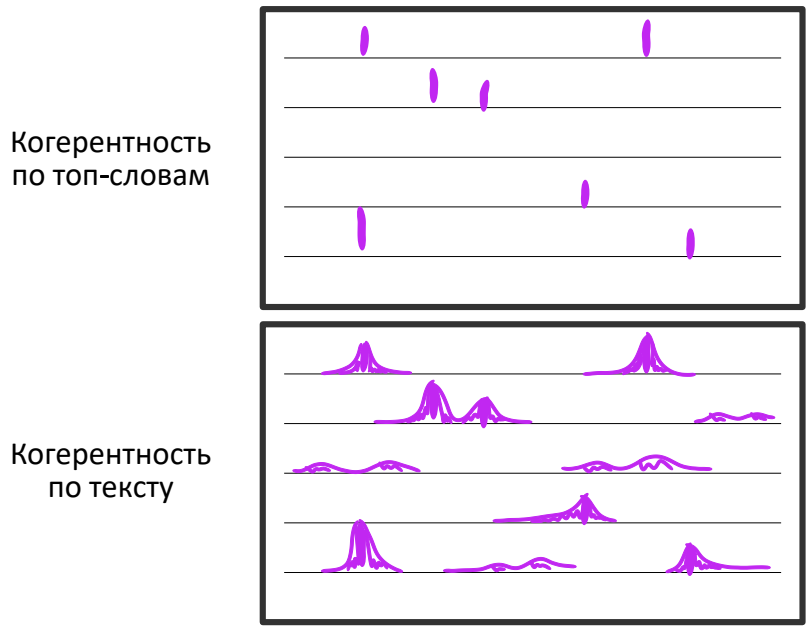


Рисунок 8 — Классическая когерентность смотрит на совстречаемости небольшого числа “избранных” слов (топ-слов) темы в тексте. Но это лишь довольно грубая попытка оценить, как тема распределена по тексту (случайно встречаются её слова в тексте или нет). Внутритекстовая же когерентность буквально проходит по тексту и смотрит на распределение темы, то есть, помимо топ-слов, принимает во внимание ещё и детали распределения. Таким образом, это попытка сделать лучше и правильнее, чем когерентность по топ-словам.

Вместо того чтобы делать выводы обо всей теме на основе совстречаемостей небольшого списка из десяти или двадцати её наиболее частых слов, следует начать с изучения слов, встречающихся в тексте вместе, а затем сравнить вероятности темы на них как $p(t \mid w)$ или $p(t \mid w, d)$.

Более подробно эта процедура будет рассмотрена в следующем подразделе.

2.2.2 Предлагаемые функции когерентности

Представляется несколько автоматических мер когерентности, отличающихся от традиционных подходов к вычислению когерентности, основанных на использовании лишь топовых слов.

Первый метод — *SemantiC* (Semantic Closeness) — оценивает семантическую близость близко расположенных в тексте слов как векторов с компонентами $p(t \mid w)$. Для оценки близости между словами в таком случае можно вычислить l_2-расстояние между соответствующими векторами:

$$\text{SemantiC}_{L_2} = -\langle [\rho(w_i, w_j) \leq \text{window}] \| \boldsymbol{w}_i - \boldsymbol{w}_j \|_2 \rangle \tag{17}$$

где $\rho(w_i, w_j)$ — расстояние по тексту между словами (количество других слов между ними), window — окно слов, в котором w_i и w_j считаются близкими по тексту. Знак минус имеет тот смысл, что делает когерентность выше, если векторы слов близки. В дополнение к величине, противоположной евклидову расстоянию, можно использовать, например, косинусную меру сходства:

$$\text{SemantiC}_{Cos} = +\langle [\rho(w_i, w_j) \leq \text{window}] \cos(\boldsymbol{w}_i, \boldsymbol{w}_j) \rangle \tag{18}$$

Третий предложенный способ оценки семантической близости слов по теме t заключается в вычислении дисперсии между компонентами векторов слов в близком окне — компонентами, соответствующими этой теме:

$$\text{SemantiC}_{Var}\big|_t = -\operatorname{Var}\big(\boldsymbol{w}_i(t), \boldsymbol{w}_{i+1}(t), \ldots, \boldsymbol{w}_{i+\text{window}}(t)\big) \tag{19}$$

Перед вычислениями все векторы слов были умножены на 1000, с тем чтобы увеличить значение абсолютной величины когерентности.

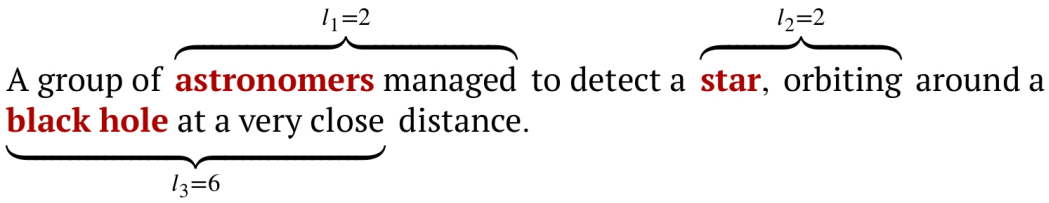


Рисунок 9 — Пример, иллюстрирующий идею когерентности TopLen. Пока в тексте встречаются слова интересующей нас темы t, продолжается счёт слов. Если же встречается какое-то не связанное с темой t слово, то оно даёт отрицательный штраф. Когда абсолютное значение общего штрафа оказывается достаточно большим, процесс останавливается, и количество подсчитанных слов даёт одно значение длины темы (длины найденного сегмента темы).

Другой метод — *TopLen* (Topic Length) — рассчитывает среднюю длину темы t в тексте, для каждого слова вычисляя балл, или скор — разницу между компонентой вектора слова $\boldsymbol{w}(t)$, соответствующей теме t, и максимальной компонентой среди остальных тем (20). В качестве вектора слова w, где компонентам соответствуют принадлежностям темам, можно взять, например, $p(w \mid t)$, $p(t \mid w)$ или $p(t \mid d, w)$. Неотрицательный параметр threshold сглаживает эффект, когда TopLen при подсчёте длины темы встречает слова не из темы: процесс подсчёта продолжается до тех пор, пока порог (который в данной работе был выбран равным 0.01) плюс суммы скоров по словам неотрицательны (см. рисунок 9 для примера).

Algorithm 1. TopLen

function score(w_j, t)
    w_j is scored

    return $\boldsymbol{w}_j[t] - \boldsymbol{w}_j[\arg\max_{\substack{1\le\tau\le|T|\\ \tau\neq t}} \boldsymbol{w}_j[\tau]]$

series $\leftarrow$ []

for $d \in D$ **do**
    for $w_i \in W_d$ **do**
        if $w_i \in W_t$ and (w_i is not scored) **then**
            series $\leftarrow \max\left\{ n \geq 0 \,:\, \text{threshold} + \sum_{j=i}^{i+n} \text{score}(w_j, t) \geq 0 \right\}$

$\text{TopLen}\Big|_t \leftarrow \Big\langle \text{series} \Big\rangle$

(20)

Последний из предложенных методов — *FoCon* (Focus Consistency) — оценивает, насколько сильно различаются соседние слова во всем тексте, суммируя для каждой пары слов две разности между компонентами их векторов $p(t \mid w)$: пара компонент для вычисления разностей состоит из компонент, максимальных в соответствующих векторах. Знак минус выполняет ту же роль, что и в случае с SemantiC (17): когерентность возрастает, когда слова различаются меньше.

$$\begin{cases} \text{FoCon} = -\sum\limits_{d\in D} \sum\limits_{\substack{w_i, w_j \in W_d \\ j-i=1}} |\boldsymbol{w}_i(t) - \boldsymbol{w}_j(t)| + |\boldsymbol{w}_i(\tau) - \boldsymbol{w}_j(\tau)| \\ t = \arg\max\limits_s \boldsymbol{w}_i(s),\ \tau = \arg\max\limits_s \boldsymbol{w}_j(s) \end{cases} \tag{21}$$

2.2.3 Эксперименты

Интерпретация и представление Автоматизированные меры когерентности основаны на подсчёте совстречаемости слов. Если топ-слова часто встречаются вместе в контекстном окне, то этот набор слов счита-

ется связным, или *когерентным*, т. е. эти слова сочетаются друг с другом естественным и разумным образом.

Неявно предполагается, что если набор топ-слов когерентный, то и вся тема тоже когерентная. На подобное допущение обращали внимание и критиковали и раньше [112], но хочется разобраться в этом вопросе количественно. Какая доля коллекции задействуется при подсчёте совстречаемостей для данного набора топовых слов?

Пусть Q — некоторое множество слов. Будем называть позицию слова $w \in Q$ *представленной*, если она имеет ненулевой вклад при подсчёте совстречаемостей слов Q (см. рисунок 10). Таким образом, оценить долю текста, участвующего при подсчёте совстречаемостей — значит оценить *представленную частоту* топ-слов всех тем модели.

Напротив, если предположить существование суперсимметрии, то введение новых **частиц** приводит как раз к такому объединению. Оказывается, что суперсимметрия не только обеспечивает объединение взаимодействий, но и стабилизирует объединённую теорию, в которой присутствуют два совершенно разных масштаба: масштаб масс обычных **частиц** (порядка 100 масс протона) и масштаб великого объединения (порядка 10^{16} масс протона). Последний масштаб уже близок к так называемому планковскому масштабу, равному обратной ньютоновской константе тяготения, что составляет порядка 10^{19} масс протона. На этом масштабе мы ожидаем проявление эффектов квантовой гравитации. В этом моменте нас ожидает приятный сюрприз. Дело в том, что гравитация всегда стояла несколько особняком по отношению к остальным взаимодействиям. Переносчик гравитации, гравитон, имеет спин 2, в то время как переносчики остальных взаимодействий имеют спин 1. Однако суперсимметрия перемешивает спины.

first top words of *topic 3: физика* with **top 10** in bold: **частица**, **электрон**, **кварк**, **атом**, **энергия**, **вселенная**, **фотон**, **физика**, **физик**, **эксперимент**, масса, теория, свет, симметрия, протон, эйнштейн, нейтрино, вещество, квантовый, ускоритель, детектор, волна, эффект, свойство, спин, гравитация, материя, адрон, поль, частота

Рисунок 10 — Слова, используемые моделью при вычислении когерентности. В представленном фрагменте тексте всего один топ-токен (“частиц”) из первых десяти, а также широкий спектр менее ярко выраженных тематических слов, которые, однако, игнорируются при расчёте когерентности методами, основанными на топ-токенах.

Основной массив данных, который будет использоваться — это корпус статей, опубликованных в “ПостНауке” — популярном российском научном интернет-журнале. Будем работать с тематической моделью для этого датасета, состоящей из 19 предметных и одной фоновой темы (см. таблицу 3).

Таблица 3 — Темы датасета “ПостНаука”, каждая из которых представлена тремя топ-словами.

Topic	First Top-Word	Second Top-Word	Third Top-Word
1: математика	математика (0.016)	задача (0.008)	декарт (0.008)
2: технологии	технология (0.015)	робот (0.012)	сеть (0.010)
3: физика	частица (0.027)	электрон (0.015)	кварк (0.015)
4: химия	химия (0.021)	молекула (0.019)	материал (0.016)
5: земля	земля (0.029)	планета (0.028)	атмосфера (0.012)
6: астрономия	звезда (0.039)	галактика (0.031)	вселенная (0.019)
7: биология	клетка (0.027)	организм (0.011)	мозг (0.010)
8: медицина	пациент (0.016)	препарат (0.012)	заболевание (0.012)
9: психология	психология (0.009)	мозг (0.009)	психолог (0.008)
10: экономика	экономика (0.016)	страна (0.010)	цена (0.008)
11: история	история (0.010)	историк (0.007)	власть (0.006)
12: политика	государство (0.014)	политика (0.012)	политический (0.011)
13: социология	социология (0.013)	социолог (0.009)	социальный (0.008)
14: культура	культура (0.015)	фильм (0.007)	искусство (0.006)
15: образование	университет (0.021)	образование (0.014)	школа (0.013)
16: язык	язык (0.077)	слово (0.037)	словарь (0.011)
17: философия	философия (0.018)	философ (0.013)	философский (0.008)
18: религия	святилище (0.010)	религия (0.007)	царь (0.006)
19: россия	россия (0.028)	страна (0.009)	русский (0.009)

Далее обратим внимание на тематическую модель, представленную в [103], которая использует выборку статей Википедии. На основе оценки 10 топ-слов экспертами эта модель была признана лучшей. Модель состояла из 50 тем. Как видно из таблицы 4, топ-токены, использовавшиеся для оценки модели, охватывают лишь незначительную часть корпуса. Другими словами, когерентности, основанные на совстречаемостях топ-токенов, могут игнорировать более 98% коллекции!

Таблица 4 — Доля корпуса, покрываемая совстречаемостями 10 наиболее частых слов каждой темы.

	PostNauka, %	Wikipedia, %
Minimum	0.016	0.0065
Median	0.048	0.029
Mean	0.062	0.036
Maximum	0.28	0.11
Total	1.2	1.7

Истинная разметка Оценка интерпретируемости темы является трудоёмкой. Сильной стороной когерентностей по топ-словам является их

способность сводить темы модели к компактному, обозримому списку слов. Но даже в этом случае сбор человеческих оценок для большого числа тем является непростой задачей.

Таким образом, возникает следующая проблема. С одной стороны, мы пытаемся построить меру когерентности, учитывающую целиком матрицы "слов в темах" Φ и "тем в документах" Θ и весь корпус. С другой стороны, для проверки такой когерентности необходимо будет провалидировать её с помощью людей, сравнить с оценками интерпретируемости тем асессорами. Поэтому необходимо собрать человеческие оценки для всего корпуса и всего распределения вероятностей.

Предлагается следующий способ обойти эту процедуру: вместо того чтобы просить экспертов-людей предоставить разметку, мы сгенерируем *полусинтетический набор данных с известной разметкой*. В этой связи большую помощь окажет структура самого корпуса данных "ПостНаука". Так, темы статей достаточно общие и разнообразные (темы известны), причём можно считать большинство документов *монотематическими*: т. е. каждое слово такого документа может быть отнесено либо к одной конкретной теме из известного заранее набора тем, либо к фоновой теме.

Итак, будем использовать эти монотематические документы для создания полусинтетического набора данных. Идея создания полусинтетического датасета заключается в том, чтобы "разрезать" монотематические документы на более небольшие монотематические сегменты, а затем "сшить" эти сегменты вместе в случайном порядке. Цель полусинтетического набора данных — служить в качестве истинной разметки (ground truth), по которой можно будет оценивать тематические модели.

Процедура генерации датасета гарантирует, что известны истинные темы для каждого слова документа. С помощью этой информации можно оценить качество сегментации текста любой тематической моделью. Предлагается два способа сделать это:

- *soft*: для каждой темы t вычисляется *сумма* $p(t \mid d, w)$ по всем парам (d, w), $d \in D$, $w \in W_d$, при этом общее значение качества темы равно сумме этих сумм для всех тем (см. рисунок 11);
- *strict*: для каждой темы t для всех сегментов темы t считается *количество* совпадений темы, предсказанной моделью для слова w в

документе d, которая равна $\arg\max_\tau p(\tau \mid d, w)$, с темой t сегмента, к которому принадлежит это слово w.

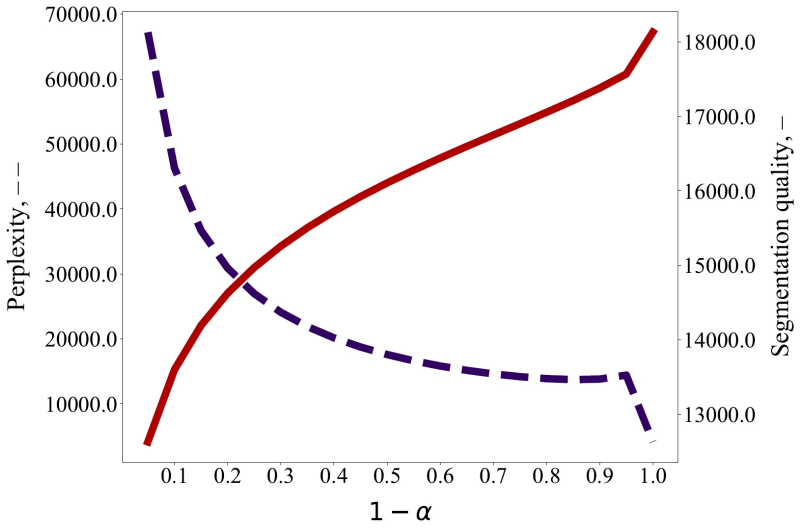


Рисунок 11 — Связь между качеством сегментации (soft) и перплексией тематической модели. По оси X отложена доля хороших матриц Φ: единица минус α (степень деградации Φ). Тот факт, что качество сегментации монотонно возрастает при уменьшении перплексии, говорит о том, что предложенный метод оценки качества сегментации может быть использован в качестве меры качества тематических моделей (точнее, что нет оснований его как таковой не использовать).

Разобравшись с истинной разметкой, мы можем оценить по ней различные меры когерентности. Качество каждой когерентности-кандидата будет равно коэффициенту корреляции Спирмена между значениями соответствующей когерентности и значениями качества сегментации (оцененными по известной разметке).

Для получения диапазона тематических моделей, дающих различные качества сегментации, генерируется несколько различных матриц Φ, каждая из которых получается как взвешенная комбинация “хорошей” матрицы Φ_{good} (которая есть известная тематическая модель датасета “ПостНаука”) и “плохой” матрицы Φ_{bad} (которая есть просто набор случайных столбцов, взятых из распределения $\text{Dirichlet}(0.01^{|W|})$):

$$\Phi(\alpha) = \alpha \cdot \Phi_{bad} + (1 - \alpha) \cdot \Phi_{good} \tag{22}$$

На рисунке 12 представлены примеры текстов, сегментированных хорошей и плохой тематическими моделями.

Для каждого α, взятого из полуинтервала $[0, 1)$ с некоторым шагом, вычисляется качество сегментации и все исследуемые метрики когерентности. На четырёх полусинтетических наборах данных, с размерами сегментов 50, 100, 200 и 400 слов, было проведено четыре таких серии экспериментов, с разными матрицами Φ_{bad}. Результаты экспериментов можно увидеть в таблице 5 и на рисунке 13.

Таблица 5 — Корреляции Спирмена между когерентностью и качеством сегментации (soft) для датасетов с различными размерами сегментов: 50, 100, 200 и 400 слов — и с 5 темами в каждом полученном полусинтетическом документе.

Coh	Corr	Coh	Corr	Coh	Corr	Coh	Corr
Newman	0.75	Newman	0.94	Newman	0.80	Newman	0.85
Mimno	0.96	Mimno	0.96	Mimno	0.94	Mimno	0.97
SC L2	0.92	SC L2	0.91	SC L2	0.70	SC L2	0.59
SC Cos	–0.97	SC Cos	–0.96	SC Cos	–0.97	SC Cos	–0.96
SC Var	**1.00**	SC Var	**1.00**	SC Var	**1.00**	SC Var	**1.00**
TopLen	**1.00**	TopLen	**1.00**	TopLen	**1.00**	TopLen	**1.00**
FoCon	**1.00**	FoCon	**1.00**	FoCon	**1.00**	FoCon	**1.00**

Все эксперименты с данными проводятся с использованием библиотеки BigARTM [39; 67].

2.2.4 Результаты

Представлены три новых метода оценки интерпретируемости тематической модели: когерентности SemantiC, TopLen и FoCon, — которые реализуют стремление учесть *все* слова текста при вычислении когерентности, с тем чтобы таким образом преодолеть недостатки когерентностей, основанных на совстречаемостях лишь небольшого подмножества слов тем (самых частых, топовых слов).

Для анализа введённых и существующих когерентностей были проведены эксперименты на полусинтетических наборах данных, состоящих

Good Topic Model

topic 16: язык

Категория будущего времени в большинстве языков Африки отсутствует. Есть много способов говорить о будущем, но это более сложные способы, касающиеся предположения, желания. Нормальный африканский грамматический приём — не говорить "я это сделаю"или "это будет а сказать "это возможно"или "я хочу это сделать"они говорят о будущем, но "попадают"в будущее непрямым путём.

topic 12: политика

И я посылаю деньги борцам за независимость Курдистана, участвую в акциях поддержки курдских повстанцев и так далее. Вот такое наложение друг на друга разных членств, разных "гражданств". В литературе последних десяти лет бытуют такие выражения, как "гендерное гражданство "экономическое гражданство". Первое указывает на членство в воображаемом сообществе женщин, приверженных идеям феминизма.

SQ (S)	SQ (H)	N	M	SC L2	SC Cos	SC Var	TL	FC
16.0e3	**3.76e4**	**-3.65**	**-2.69**	**-3.70**	0.700	**-8.12e3**	**3.45**	**-5.44e4**

Bad Topic Model

topic 16: язык

Категория будущего времени в большинстве языков Африки отсутствует. Есть много способов говорить о будущем, но это более сложные способы, касающиеся предположения, желания. Нормальный африканский грамматический приём — не говорить "я это сделаю"или "это будет а сказать "это возможно"или "я хочу это сделать"они говорят о будущем, но "попадают"в будущее непрямым путём.

topic 12: политика

И я посылаю деньги борцам за независимость Курдистана, участвую в акциях поддержки курдских повстанцев и так далее. Вот такое наложение друг на друга разных членств, разных "гражданств". В литературе последних десяти лет бытуют такие выражения, как "гендерное гражданство "экономическое гражданство". Первое указывает на членство в воображаемом сообществе женщин, приверженных идеям феминизма.

SQ (S)	SQ (H)	N	M	SC L2	SC Cos	SC Var	TL	FC
5.54e3	1.10e4	-4.83	-3.12	-12.9	0.947	-37.0e3	2.87	-13.9e4

Рисунок 12 — Иллюстрация процесса сегментации тематической моделью полусинтетического текста. На рисунке показаны два сегмента из разных тем размером 50 слов каждый после обработки “плохой” тематической моделью (*Bad Topic Model*) и “хорошей” (*Good Topic Model*) (соответственно при $\alpha = 1$ и $\alpha = 0$ (22)). Представленные сегменты были извлечены из одного из сгенерированных полусинтетических документов, в котором они были соседними. Словам, которые не помечены, были присвоены темы, отличные от тем двух представленных сегментов. Под сегментами указаны значения когерентности. SQ (S) — обозначает soft качество сегментации, SQ (H) — strict качество сегментации, N — Newman, M — Mimno, SC — SemantiC, TL — TopLen, FC — FoCon. Значения, выделенные жирным шрифтом, указывают на то, что функция когерентности возрастает по мере увеличения качества модели.

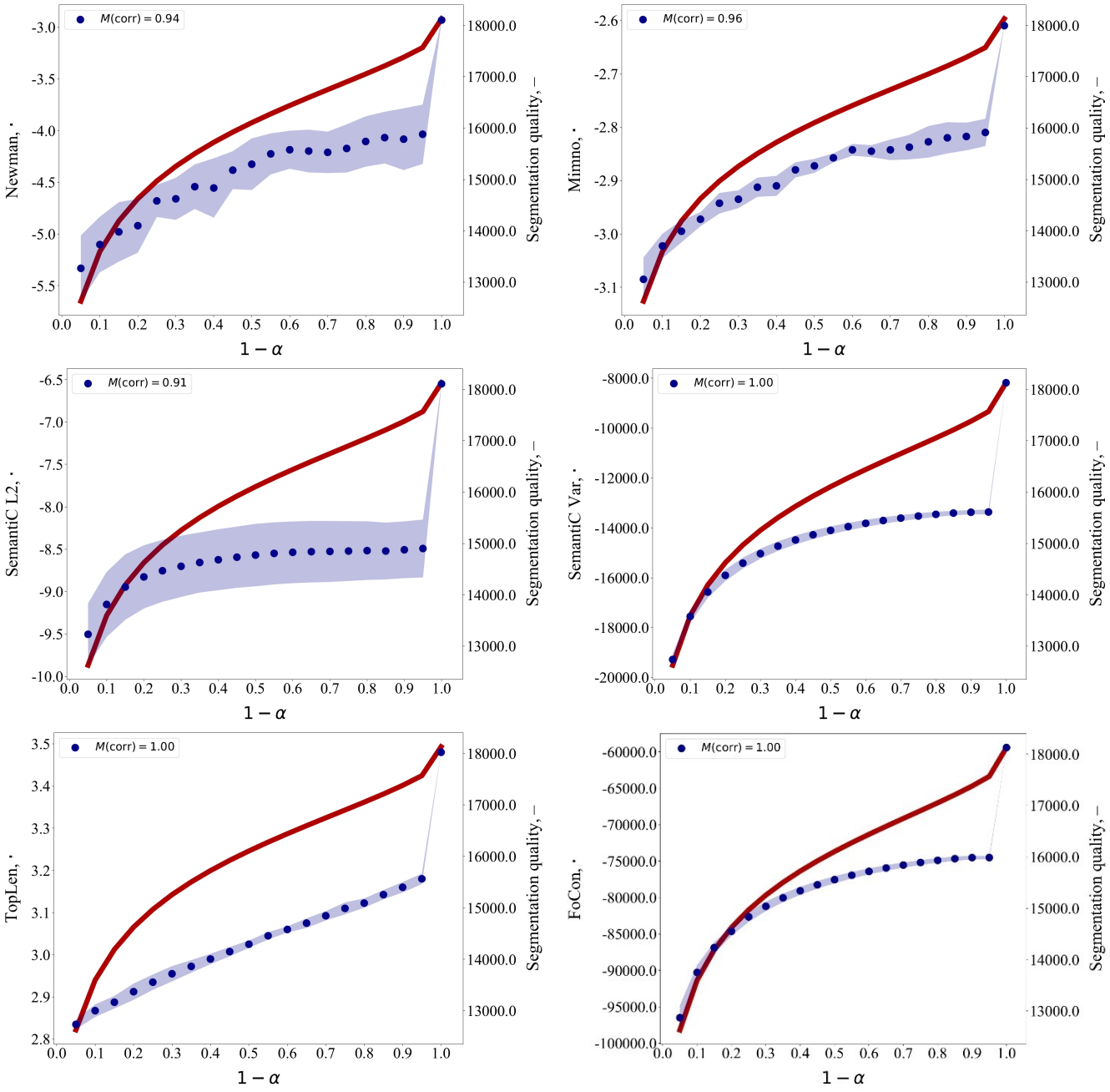


Рисунок 13 — Сравнение различных мер когерентности с качеством сегментации в зависимости от α (параметр “деградации” тематической модели: чем больше, тем ближе модель к заведомо “плохой” Φ_{bad}). Значения когерентности на графиках — это усреднённые значения по серии из четырёх экспериментов с различными матрицами Φ_{bad}.

из одинакового размера сегментов различных тем. Предложенные методы демонстрируют высокую корреляцию с качеством сегментации (оцененным на полусинтетических датасетах). Наилучшие результаты корреляции показали когерентности SemantiC$_{Var}$ и TopLen.

2.2.5 Возможные направления дальнейших исследований

Внутритекстовые когерентности, в отличие от когерентности по топ-словам, опираются на полную информацию о распределении темы в тексте при оценке её качества. Однако вместе с тем предложенные когерентности обладают и заметным недостатком. Который заключается в том, что они не сохранили преемственность идеи Ньюмана об оценке неслучайности совместной встречи слов в тексте с помощью PMI. Кажется же, что PMI важно сохранить, потому что PMI оценка неслучайности встречи слов неплохо себя зарекомендовала и часто используется, не только при вычислении когерентностей [78]. Кроме того, сохранение PMI позволило бы обобщить когерентность Ньюмана: предложить такую внутритекстовую, которая бы включала когерентность Ньюмана как частный случай.

И в качестве такой обобщающей когерентности можно предложить следующую — *средневзвешенную внутритекстовую когерентность*:

$$\operatorname{coh}(t) = \frac{\sum_{u,v} \operatorname{rel}_t(u,v) \operatorname{coh}(u,v)}{\sum_{u,v} \operatorname{rel}_t(u,v)} \tag{23}$$

где $\operatorname{coh}(u,v)$ — *сочетаемость* пары слов u, v в текстах (оценка неслучаемости совместной встречи); $\operatorname{rel}_t(u,v)$ — *релевантность* слов u и v теме t.

В частности, выбор $\operatorname{coh}(u,v) = \operatorname{PMI}(u,v)$ и $\operatorname{rel}_t(u,v) = [u,v \in \operatorname{top}_k(t)]$ приводит к когерентности Ньюмана по топ-словам (16).

Учесть большее число слов (перейти к внутритекстовой когерентности) можно, положив, например, $\operatorname{rel}_t(u,v) = \phi_{ut} + \phi_{vt}$, или $\sqrt{\phi_{ut}\phi_{vt}}$, или $[\phi_{ut}\phi_{vt} \geq \varepsilon]$ (где можно взять $\varepsilon = (1/W)^2$).

В качестве $\mathrm{coh}(u, v)$ можно оставить PMI. Однако можно также попробовать ещё и другие варианты, также оценивающие неслучаемость совместной встречи слов u и v. Например, $\mathrm{coh}(u, v) = (\mathrm{PMI}(u, v) - \delta)_+$, или $\mu\left(\frac{P(u,v)}{P(u)P(v)}\right)$, или $\frac{P(u,v)-P(u)P(v)}{\sqrt{P(u,v)}}$.

Суммирование в (23) предполагается по всем парам W^2 слов словаря (у которых ненулевая $\mathrm{rel}_t(u, v)$). Но это чрезмерно большой объём вычислений. Этого можно избежать, если учитывать лишь пары слов, находящиеся в общих *контекстах*, например, на расстоянии не более 10 слов друг от друга, или в одном предложении, в одном абзаце. В таком случае все $\mathrm{rel}_t(u, v)$ можно брать ненулевыми. Более того, появляется следующая возможность вычисления релевантности пары слов u, v теме t:

$$\mathrm{rel}_t(u, v) = \sqrt{p(t \mid d, u)p(t \mid d, v)}$$

В итоге за один проход по коллекции для каждой темы t аккумулируются суммы в числителе и знаменателе (23).

Отметим, что значение $\mathrm{rel}_t(u, v)$ зависит только от слов u, v и темы t. Однако тут ещё играет роль, как мы смотрим на слова. Обобщённая внутритекстовая когерентность — внутритекстовой её делает то, что она проходит по всему тексту и смотрит на все слова, а не только на топ-слова, но учитывает их с разным весом (который как раз определяется как rel_t от слов). Это можно считать “первым уровнем внутритекстовости”. Но можно пойти дальше. Можно при оценке релевантности слов теме смотреть не только на “статичные” вероятности типа $p(w \mid t)$, но, скажем, можно смотреть на вероятности $p(t \mid d, w)$ — вероятности, зависящие от контекста (от документа). Получается, в таком случае мы будем смотреть не просто на слова, а на слова в документе. Это можно считать “вторым уровнем внутритекстовости”, который не про то, сколько слов мы учитываем, а про то, как мы их учитываем, как воспринимаем слова, как понимаем их тематику (через контекст).

В этом свете можно предложить ещё одну функцию $\mathrm{rel}_t(u, v)$, которая представляет из себя нечто среднее между когерентностью по топ-словам и внутритекстовой когерентностью. Пусть есть тема t. Далее для каждого документа d можно взять top_k слов по вероятностям $p(t \mid d, w)$ и смотреть

на их совстречаемости в этом документе, чтобы посчитать когерентность по топ-словам в рамках одного этого документа d. Последующее усреднение по документам D дало бы значение когерентности темы t. Получается когерентность, с одной стороны, похожая на когерентность Ньюмана по топ-словам (потому что также опирается лишь на топ-слова и их совстречаемости), однако топ-слова эти находятся не один раз по вероятностям $p(w \mid t)$, а для каждого документа по вероятностям $p(t \mid d, w)$ — получается как бы внутритекстовый поиск топ-слов. Подобный взгляд на топ-слова темы через $p(t \mid d, w)$ вместо $p(w \mid t)$ напоминает последний опциональный слой в модели BERTopic [113] —Representation layer, — когда по выстроенной по вероятностям $p(w \mid t)$ последовательности топ-слов темы некоторым алгоритмом получают *другую* последовательность топ-слов для той же темы, которая должна быть более интерпретируемой (репрезентативной), чем изначальная.

Показать состоятельность предлагаемого подхода (и также продолжить идею с преемственностью) можно повторением эксперимента по схеме Ньюмана (7). Однако в эксперименте Ньюмана эксперты оценивали качество тем так же лишь по небольшому числу топ-слов... Таким образом, схема Ньюмана сравнения когерентностей обладает тем же недостатком, что и сама когерентность по топ-словам! Поэтому мало представить обобщающую формулу когерентности — вместе с ней надо ещё представить новую схему эксперимента (отталкивающуюся также от использованной Ньюманом). И схема предлагается следующая (см. рисунок 14). Вместо того, чтобы предлагать экспертам оценивать качество тем по неполной информации (по топ-словам), можно с помощью экспертов разметить текст (выбранные заранее из коллекции D_0 фрагменты текста), выделив в нём *тематические цепочки слов*. И далее можно будет сравнить найденные моделью темы (их распределение в тексте) с экспертной разметкой на цепочки, получив оценку *согласованности* тем разметке. Тогда лучшей будет та когерентность, которая лучше всего коррелирует с оценками согласованности между темами и размеченными тематическими цепочками.

Покажем, как можно оценить согласованность темы и разметки текста на тематические цепочки. Тематика цепочки C как подмножества слов

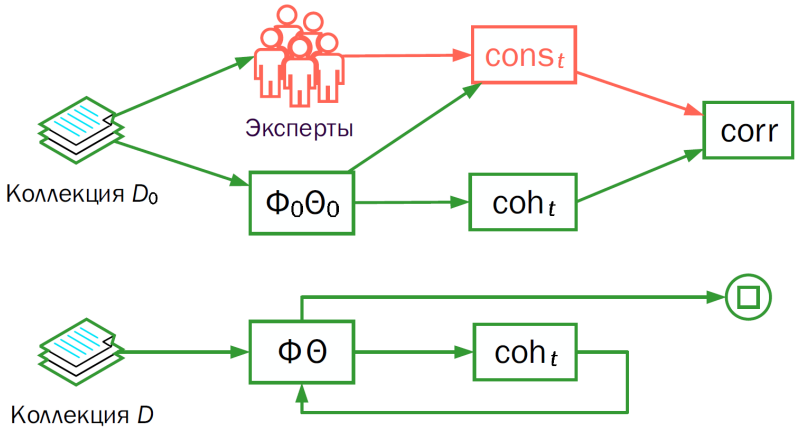


Рисунок 14 — Схема эксперимента, предлагаемая вместо Ньюмановской (7), для сравнения обобщённых взвешенных внутритекстовых когерентностей. Сверху: эксперты выделяют в тексте тематические цепочки слов, оценивается согласованность тем модели с этими цепочками, и по корреляции значений когерентностей и согласованностей выбирается лучшая когерентность. Снизу: применение когерентности для автоматической оценки качества тем — так же, как в (7).

(из формулы полной вероятности):

$$p(t \mid C) = \sum_{w \in C} p(t \mid w) p(w \mid C) = \operatorname*{mean}_{w \in C} p(t \mid w)$$

Множество цепочек, *согласованных* с темой:

$$C(t) = \{C \mid t = \arg\max_{\tau} p(\tau \mid C)\}$$

И тогда мера согласованности темы с размеченными цепочками может быть определена как:

$$\text{cons}(t) = \operatorname*{mean}_{C \in C(t)} p(t \mid C)$$

Вообще, на выделение тематических цепочек слов можно смотреть как на получение идеальной тематической модели (см. рисунок 15). Поэтому, сравнивая с такой разметкой распределение найденной моделью темы, мы можем понять, хорошая тема или нет (согласуется с заведомо хорошей темой или нет). Внутритекстовая когерентность же как раз

и стремится оценить качество найденной темы тоже по её распределению в тексте. (Поэтому стоит ожидать, что когерентность по тексту должна лучше коррелировать с согласованностью разметке, чем когерентность по топ-словам.)

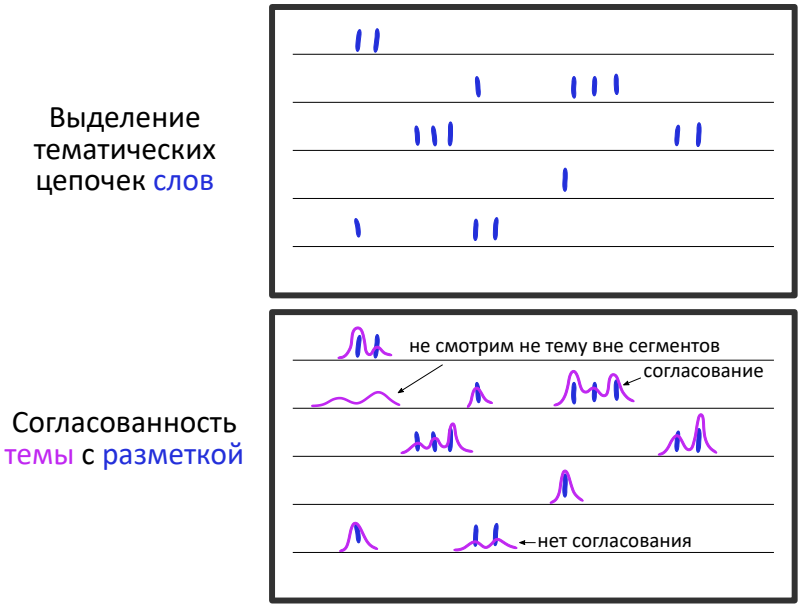


Рисунок 15 — В результате выделения в тексте асессорами тематичных слов получается размеченный документ. Но такой документ — это по сути то, что мы хотим получить от тематической модели на выходе (помимо информации о вероятностях слов в темах). Выделив цепочки, мы получаем “идеальную” разметку и можем потом сравнить с ней ту разметку, которую получаем с помощью тематической модели: для каждой цепочки оценивая, как хорошо выражена в ней та или иная тема. Если вероятности темы на размеченных словах большие, то это означает, что ответ модели согласуется с найденными цепочками и тему можно считать “хорошей”.

Отдельного внимания заслуживает вопрос о том, что мы считаем тематической цепочкой (определение) и как человеку (асессору) находить её в тексте. Что хочется от цепочки — чтобы это было одно или несколько слов, идущих в тексте одно за другим и связанных одной темой. Таким образом, в самом простом и понятном смысле, мы можем считать цепочками просто одиночные слова, 2-граммы, 3-граммы и так далее. То есть по сути просто тематичные термы. Однако понятие “тематическая цепочка” можно рассматривать и в более широком смысле. Так, надо ли требовать

от цепочки какой-то “смысловой цельности”, помимо монотематичности (см. рисунок 16)?

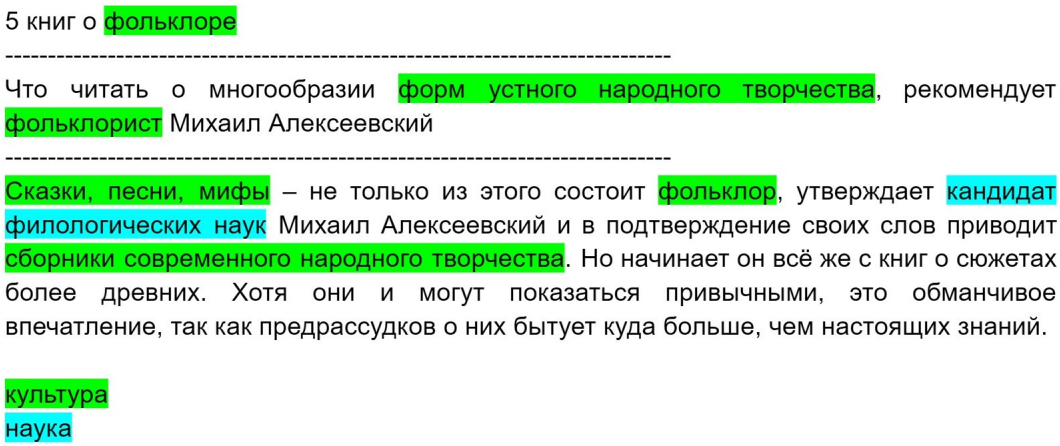


Рисунок 16 — Пример выделения тематических цепочек в тексте. В первом предложении идут слова “сказки, песни, мифы” — то отдельные тематически окрашенные токены, но их можно объединить в одну цепочку, так как они идут подряд друг за другом. Но с идущей далее фразой “сборники современного народного творчества” уже получается не так однозначно. С одной стороны, можно бы было выделить как цепочку только “народного творчества”, потому что “сборники” и уже тем более “современные” можно отнести к словам общей лексики. С другой стороны, можно выделить как цепочку и всю фразу целиком, потому что это “единая”, оформленная фраза — “сборники народного творчества” — единая по теме, завершённая по смыслу. Да, при этом внутри цепочки будет стоять ещё слово “современного” — хоть оно и общее, но вместе с ним в качестве тематической цепочки получается цельная фраза. Похожая ситуация и с цепочкой “форм устного народного творчества” выше в заголовке: “форм” можно было и не включать в цепочку, получилось бы “устного народного творчества” (что уже есть популярное сочетание слов), но “формы” можно и включить в цепочку, чтобы дооформить фразу.

В более общем смысле, о тематической цепочке можно думать как о расширении понятия “сочетание слов”. Сюда входят и 2-граммы, и 3-граммы, и “скип-граммы” (когда допускается, чтоб интересуемые слова

были отделены друг от друга другими словами).[1,2,3] Таким образом может получиться так, что, например, слова по отдельности будут относиться к разным темам (“войны”, “звёздные”), но при выделении цепочек важно смотреть на слова в контексте — и тогда они вместе как коллокация будут принадлежать другой теме (“звёздные войны”). Из сказанного вытекает следующий вариант определить тематическую цепочку — одна или несколько подряд идущих коллокаций слов (вместе с, возможно, словами общей лексики, идущими между образующими коллокацию словами), относящихся к одной теме. Однако, имея лишь “сырой” неразмеченный текст, без тем, не понятно, что вообще означает “относящихся к одной теме”. Поэтому вместо этого можно сказать “относящихся к родственным темам” (где “родственные” — потомки одной, как, например, “алгебра” и “геометрия” — потомки “математики”). Однако определение и через “родственные” темы не идеально, потому теперь не вполне понятно, что считать “родственными” темами. Потомки одной (частные случаи более общей), или такие, у которых просто есть какой-то общий предок (“математика” + “история” = “история математики”; “математика” + “симпсоны” = “математические сюжеты в симпсонах” — примеры двух тем с общим предком “математика”). Возможно, лучше тогда будет думать про цепочку как про максимальный по длине фрагмент текста из нескольких подряд идущих сочетаний слов, которые в рамках контекста можно считать относящимися к одной теме (и тогда слова про “алгебру” и “геометрию” можно считать относящимися к одной теме “математика”, если это поможет получить более длинную цепочку). (На примере “тематической цепочки” видно, что самым базовым понятиям бывает не просто дать формальные, точные и ясные определения. Как с тем же основополагающим понятием “темы” в тематическом моделировании, которую вообще можно определять по-разному (и которая для человека и для модели — уже не одно и то же).)

Кроме определения понятия тематической цепочки, важным кажется вопрос о размере датасета с размеченными цепочками. Или, иными словами, каким должен быть этот датасет, чтобы полученным с его по-

[1] nltk.org/howto/collocations.html.
[2] nltk.org/api/nltk.util.html#nltk.util.skipgrams.
[3] radimrehurek.com/gensim/models/phrases.html.

мощью результатам можно было доверять. Идеальный датасет цепочек — это просто весь размеченный датасет документов. Но подготовить датасет цепочек такого размера (сравнимого по размеру с датасетом, к которому применяется тематическое моделирование) с помощью асессоров — это очень масштабная работа. Кроме того, кажется, что весь датасет документов и нет смысла размечать на цепочки. Что хочется от датасета цепочек? Во-первых, чтобы он всё-таки был в некотором смысле не маленьким (чтобы при создании этого датасета использовалось сколько-то документов и чтобы в результате получилось сколько-то цепочек). Кроме этого, хотелось бы ещё, чтобы этот датасет цепочек был разнородным, то есть чтобы это был пусть и не большой, но полноценный срез всей коллекции документов, чтобы при накоплении цепочек использовались разные документы, по разным темам. Так чтобы цепочки в совокупности давали хорошее представление всего датасета.

Одним из способов проверки качества датасета цепочек также может быть следующий. При оценке согласованности тем модели множеству цепочек важно убедиться, что *различимость* выделенных цепочек близка к единице:

$$\text{diff} = \frac{\sum_C \#\{t : C \in C(t)\}}{\#\{C\}} \geq 1$$

Для получения датасета из тематических цепочек можно, как и Ньюман, привлекать асессоров. Но помимо этого, перспективной кажется возможность использования для этой цели *больших языковых моделей* [114].

2.3 Определение оптимального числа тем

Тематические модели — это статистические модели, которые обычно используются в области анализа текстов без учителя (unsupervised text analysis). Тематическое моделирование предполагает, что в тексте коллекции существует сколько-то *скрытых* тем, которые её объясняют.

Обучение тематической модели сводится к нахождению двух вероятностных распределений: “слов в темах” $(\phi_{wt})_{w \in W}$ для каждой темы $t \in T$ (таким образом, получается стохастическая матрица Φ размера $W \times T$) и

“тем в документах” $(\theta_{td})_{t \in T}$ для каждого документа $d \in D$ (таким образом, получается стохастическая матрица Θ размера $T \times D$). Мы ограничиваемся матрицами параметров Φ и Θ, поскольку они присутствуют в каждой тематической модели. Некоторые тематические модели вводят дополнительные параметры, помимо этих — исследование внутренних мер качества, связанных с этими параметрами, выходит за рамки данной работы.

Число тем T является ключевым гиперпараметром большинства тематических моделей. Поэтому естественно, что существует ряд влиятельных публикаций, предлагающих способ выбора этого гиперпараметра [115—120]. Однако общепринятого консенсуса по этому вопросу не существует. В частности, в литературе нет согласия по поводу последовательности шагов, которые необходимо выполнить для определения оптимального количества тем. Также, по-видимому, существуют разногласия по поводу того, какие методы являются подходящими [121] (точнее, в самом ли деле являются подходящими некоторые методы).

Нас больше всего интересуют *внутренние* критерии качества тематических моделей — которые не используют никаких внешних ресурсов, меток или человеческих оценок. Подход по определению числа тем с помощью внутренних критериев обычно основан на получении различных моделей с разным количеством тем, вычислении для этих моделей определённой метрики качества (возможно, с использованием кросс-валидации на отложенной подвыборке документов) и выборе количества тем, соответствующего модели с лучшим значением метрики. В данной работе не рассматриваются внешние подходы, направленные на оптимизацию какого-либо внешнего критерия, например, классификация с использованием размеченного валидационного датасета, поскольку очевидно, что явная оптимизация для какой-то конкретной задачи даёт лучший результат, измеряемый производительностью модели на этой задаче.

Отметим, что из рассмотрения исключаются модели, которые получают необходимое количество тем автоматически (например, иерархический процесс Дирихле (hierarchical Dirichlet process) [122]). Исключаются по двум причинам. Во-первых, они, как правило, добавляют новый набор гиперпараметров, требующих оптимизации. Во-вторых, они не универсальны: любая тематическая модель, содержащая распределения Φ и Θ,

может быть оценена по любой внутренней метрике, в то время как сложные байесовские тематические модели обучаются оптимизацией некоторой функции потерь, специфичной для их параметров.[4]

В данной работе исследуется несколько функций качества, предложенных в литературе в связи с исследованием различных датасетов и тематических моделей, и предпринимается попытка сформулировать набор полезных рекомендаций для практиков тематического моделирования по возможности определения числа тем в корпусе текстов.

Далее в разделе изложение организовано следующим образом.

В подразделе 2.3.1 рассматриваются работы, связанные с предлагаемой методологией в целом, а в подразделе 2.3.2 рассматриваются и оцениваются различные методы, используемые для выбора "оптимального" числа тем. Подраздел 2.3.3 описывает дизайн эксперимента, а именно, тематические модели и используемые датасеты. В подразделе 2.3.4 приводятся экспериментальные доказательства, связанные с проблемой определения числа тем T, обсуждаются некоторые моменты. Наконец, в подразделе 2.3.5 подводятся итоги.

2.3.1 Связанные работы

Многие исследователи предлагают способы определения числа тем T (см. раздел 2.3.2), но они, как правило, фокусируют свои эксперименты вокруг определённых семейств тематических моделей или привязываются к датасетам определённых типов, размеров, а также проводят очень ограниченный обзор других существующих подходов. В этом разделе мы обсуждаем некоторые предыдущие работы, в которых рассматривается целый ряд методов для определения числа тем.

Отметим работу [123], которая предлагает новый процесс оценки количества кластеров в датасете и сравнивает его с большим количеством традиционных метрик, реализованных в пакете `NbClust`. Сравнение проводится на корпусе 20NG и нескольких подмножествах корпуса

[4]Тем не менее, такие модели иногда используют независящие от самой модели метрики, которые неявно в процессе оптимизируются. Эти метрики качества мы по возможности *будем* использовать.

WikiRef220; большинство традиционных метрик не дают хороших результатов.

В связанной работе [121] перечисляется и рассматривается ряд других методов.

Автору известны только два программных пакета, которые реализуют ряд метрик и работают по похожей идее: они позволяют пользователю исследовать значения нескольких метрик качества тематических моделей в некотором диапазоне числа тем T и выбрать то значение числа тем, которое кажется наиболее подходящим.

Пакет `ldatuning`[5] для R [124] позволяет применять следующие методы к модели LDA: D-Spectral [115], D-avg-COS [116], D-avg-JS [117], и holdPerp [119]. В недавней работе [125] рассматривается производительность этих методов на нескольких сгенерированных наборах данных с известным значением T.

Пакет `TOM`[6] для Python [126] реализует несколько методов для оценки T. Такие как тот же D-Spectral [115], также toptokens-ssample-stab [118]. Поддерживаются модели LDA и модели, основанные на неотрицательном матричном разложении (Non-Negative Matrix Factorization).

Пакет `OCTIS`[7] также реализует ряд метрик качества тематических моделей, но не использует их для выбора T.

2.3.2 Внутренние критерии качества для выбора числа тем

Перплексия Классический внутренний подход к оценке качества, вычисляется либо на отложенной выборке [119] (*holdPerp*), либо на той же самой, на которой проходило обучение тематической модели. В работе [120] этот метод улучшается: вместо “сырой” перплексии рассматривается *скорость изменения перплексии* (*RPC*) (то есть, по сути, учитывается наклон кривой перплексии, а не её абсолютные значения).

[5] github.com/nikita-moor/ldatuning.
[6] github.com/AdrienGuille/TOM.
[7] github.com/mind-Lab/octis.

Кластеризация Интересно, что Краснов и другие [121] не смогли воспроизвести результаты по определению числа тем с помощью коэффициента Силуэта *SilhC* и индекса Калинского – Харабаша *CHI* на конкретном наборе данных.

Разнообразие и полнота Кажется интуитивно понятным, что когда количество выставляемых в тематической модели тем слишком велико, модель порождает множество мелких тем, похожих друг на друга. И наоборот, если тем выставлено мало, то повышается риск какие-то темы потерять (не считая того, что какие-то темы при этом могут получиться плохо интерпретируемыми как смесь нескольких более маленьких тем).

Наиболее влиятельная работа в этом направлении [116] предлагает использовать среднее косинусное расстояние между темами (*D-avg-COS*) в качестве критерия для выбора модели. Эта идея продолжается в работе [117] путём рассмотрения дивергенции Йенсена – Шеннона (*D-avg-JS*) вместо косинусного расстояния. В работе [127] используется другой способ оценки разнообразия тем — среднее евклидово расстояние (*D-avg-L2*).

В данной работе существующая методика будет ещё более расширена: путём рассмотрения среднего расстояния до ближайшей темы (вместо среднего расстояния между парами тем). В результате получаем функции качества *D-cls-COS*, *D-cls-JS* и *D-cls-L2*. Кроме того, представляются *D-avg-H* и *D-cls-H*, которые основаны на расстоянии Хеллингера.

Другая метрика (*D-Spectral*) была предложена в [115], она интегрирует информацию, находящуюся в матрицах Φ и Θ, учитывая спектральные значения Φ и строки ненормированной Θ. Предложенная функция качества тематической модели *Spectral Divergence Measure* отражает степень ортогональности между векторами тем.

Подход, предложенный в [43], начинает с чрезмерно большого количества тем, а затем использует регуляризацию, чтобы свести большинство из них к нулю. Примечательно, что предложенный регуляризатор способен удалять из модели линейные комбинации существующих тем. Функция, максимизируемая регуляризатором, представляет собой дивергенцию Кульбака – Лейблера (KL) между равномерным распределением $u(t) = \frac{1}{T}$ и распределением, неявно определяемым тематической моделью

$p(t)$:

$$\mathrm{KL}(u(t) \parallel p(t)) = \mathrm{KL}\left(\frac{1}{T} \middle\| \sum_d \theta_{td} \frac{n_d}{n}\right) \to \max_\Theta$$

Будем интерпретировать эту величину как критерий качества, именуемый *uni-theta-divergence*.

Авторы [127] используют этот процесс в противоположном направлении. Они начинают с небольшого количества тем и итеративно добавляют новые темы, описывающие документы, которые плохо объясняются моделью. Правда, в этом подходе вместо неизвестной T получаем неизвестный порог ϵ. В статье предлагается использовать общее разнообразие тем в качестве критерия для определения ϵ.

Лифт Эта мера качества (*lift-score*) была введена в относительно недавней работе [128], где было замечено, что модели LDA с более "продвинутыми" информативными априорными распределениями соответствуют более высоким показателям *lift-score*. Следовательно, *lift-score* может быть полезен для настройки гиперпараметров модели. Это ставит интересный вопрос о том, можно ли использовать *lift-score* для определения числа тем T.

Информационно-теоретические Другим методом является использование байесовского информационного критерия *BIC*, принципа минимальной длины описания *MDL* или принципа минимальной длины сообщения *MML* (15). Наиболее заметной из последних работ, использующих *BIC*, является [129], где *BIC* значительно расширен, и выведена форма *BIC*, учитывающая дополнительные параметры новой предложенной модели. В работе [109] исследуется информационный критерий Акаике (Akaike Information Criterion, *AIC*) и *BIC* как функция от количества тем (хотя это и не является основным направлением указанной работы).

Энтропия В работе [105] проводится аналогия между моделями тем и неравновесными сложными системами, где количество тем эквивалентно количеству состояний, которые может занимать каждая частица (слово). Предполагается, что "правильное" число тем должно соответствовать

равновесному состоянию, которое характеризуется минимумом энтропии. Таким образом, проблема сводится к нахождению минимума определённой функции.

С тем чтобы вычислить энтропию, нужно определить множество $S = \{(w,t) \mid \phi_{wt} > \varepsilon_0\} \subset \Phi$ для некоторого фиксированного ε_0. После этого энергия определяется как $E = -\log \sum_{(w,t)\in S} \phi_{wt}$, свободная энергия как $E_f = E - T \log\big(|S|/(WT)\big)$. Наконец, энтропия Реньи вычисляется как $-E_f/(T-1)$.

В работе [105] используется значение $\varepsilon_0 = (W)^{-1}$, однако в процессе экспериментов, проводимых по настоящей работе, было обнаружено, что это не всегда приводит к хорошим результатам. Поэтому решено было рассмотреть также случаи $\varepsilon_0 = 2(W)^{-1}$ и $\varepsilon_0 = 0.5(W)^{-1}$. Все эти критерии обозначим соответственно *renyi-1*, *renyi-2* и *renyi-0.5*.

Устойчивость Авторы [118] используют анализ устойчивости (стабильности) для выбора гиперпараметров моделирования (в частности, количества тем). Этот подход часто используется при анализе моделей кластеризации, таких как k-means или неотрицательное матричное разложение (Non-Negative Matrix Factorization). Интуитивно кажется понятным, что решения с “неправильным” количеством кластеров неустойчивы, так как они вынуждены либо объединять кластеры произвольным образом, либо создавать случайные кластеры данных.

Устойчивость (*toptokens-ssample-stab*) измеряется путём многократного создания перетасованной подвыборки данных, обучения на ней тематической модели и последующего сравнения полученных в результате топ-токенов — с топ-токенами эталонной модели, построенной на всём наборе данных. Численное значение устойчивости рассчитывается как коэффициент сходства Жаккара (для сопоставления тем вновь обученной модели и эталонной, полученного с помощью Венгерского алгоритма).

Другой важной работой в том же направлении является [130], где авторы сосредоточились на воспроизводимости назначений классов. Так, каждому документу d присваивается его наиболее вероятная тема. Затем для каждой пары (d_1, d_2) фиксируется, принадлежат ли d_1 и d_2 к одному кластеру. В результате получается матрица связности $D \times D$. Эта матрица

усредняется по нескольким различным тематическим моделям. В итоге предлагаемая мера устойчивости определяется как коэффициент кофенетической корреляции (cophenetic correlation coefficient) средней матрицы связности. Интересно отметить, что подвыборки авторами не используются. Рандомизация происходит только за счёт различных инициализаций модели. Библиотека `TOM` по умолчанию использует десять случайных запусков для оценки устойчивости. (В данной работе описанный подход проверить не получилось — из-за больших вычислительных затрат, связанных с обучением дополнительных тематических моделей на больших текстовых коллекциях.)

Когерентность Хотя это и не отражено в научной литературе, другой разумный подход к определению числа тем состоит в том, чтобы построить множество моделей с разными значениями T и выбрать ту, которая дает наибольшее значение когерентности (*Coherence*) [131]. Когерентность — это широко используемая метрика качества для тематических моделей, которая вычисляется с помощью подсчетов совстречаемостей небольшого количества (например, 10) наиболее вероятных слов каждой темы (топ-слов).

Краснов и другие [121] предлагают интересную вариацию: 10 топ-слов заменяются их плотными эмбеддингами (dense embeddings, использовались GloVe), а количество тем выбирается в соответствии с индексом Дэвиса – Болдина (Davies – Bouldin Index), показателем качества кластеризации

Повторимся, что в данной работе фокус сделан на возможность определения числа тем в датасете с помощью *внутренних* метрик качества тематических моделей (2.1, 2.3.2). Исходный код вычисления метрик на Питоне находится в открытом доступе.[8]

[8] github.com/machine-intelligence-laboratory/OptimalNumberOfTopics.

2.3.3 Методология

Вопрос, на который хочется ответить: хорошо ли определено понятие “оптимального числа тем”? Другими словами, есть ли согласие между различными подходами по поиску числа тем, предложенными в литературе? Кроме того, достаточно ли устойчивы предлагаемые подходы к изменению параметров?

Методология заключается в следующем. Обучается несколько разных тематических моделей (например, PLSA, LDA, ARTM и TARTM; см. раздел 2.3.3), в течение количества итераций, достаточного для сходимости, с гиперпараметром T, варьирующимся от T_{min} до T_{max}. Процесс обучения запускается три раза, каждый раз с новой случайной инициализацией, и для каждого прогона измеряется несколько показателей качества. Если метод подсчёта качества требует задаваемого пользователем параметра, исследуются различные значения-кандидаты.

Этот процесс повторяется для нескольких различных корпусов. Для того чтобы можно было вычислить held-out перплексию, каждый корпус случайным образом разбивается на обучающий и тестовый (80% обучающий, 20% тестовый). Мы не перемешиваем документы, поскольку используемый алгоритм вывода (BigARTM [87]) не зависит от порядка документов.

В качестве отдельного эксперимента, для определения T применяется анализ устойчивости. Реализация оценки по устойчивости адаптирована из библиотеки `TOM` [126], вдохновлена [63; 118]. Вся процедура заключается в оценке разнообразия тем на подвыборках данных. Создаётся 5 подвыборок оригинального датасета (без повторов документов, without replacement). Размер всех подвыборок $D_i \subseteq D$, $i \in [0, S)$ фиксирован и равен 0.5 от размера исходного набора данных. Тематические модели обучаются для диапазона номеров тем $t \in [T_{\text{min}}, T_{\text{max}}]$ на каждой из полученных подвыборок документов. Случайное начальное значение (random seed), определяющее начальные веса в матрице Φ_t, фиксировано и равно 0 для всех тематических моделей. После подготовки подвыборки данных и обучения моделей, последним шагом является сравнение моделей с оди-

наковым количеством тем t, но обученных на разных подвыборках данных D_i, D_j, $i \neq j$. Значение расхождения между такими тематическими моделями считается путём их попарного сравнения по функции расстояния Жаккара. Далее, имея $|t| \times |t|$ матрицу расстояний между темами и получив решение этой линейной задачи о назначениях, вычисляем расстояние между тематическими моделями $\rho_{stab}\big(\Phi_t(D_i), \Phi_t(D_j)\big)$, $i, j \in [0, S)$. Эти расстояния затем усредняются по количеству сравнений тематических моделей (которых по результатам проведённых экспериментов всего $\binom{5}{2} = 10$):

$$\frac{1}{\binom{S}{2}} \sum_{0 \leq i < j < S} \rho_{stab}\big(\Phi_t(D_i), \Phi_t(D_j)\big)$$

Чем меньше это число, тем лучше. Но эта логика отличается от [118], где предлагаемый показатель стабильности, напротив, является оценкой сходства моделей, и поэтому чем он был выше, тем лучше. Другим заметным отклонением является расчёт, включающий $\binom{S}{2}$ попарных сравнений — вместо обучения начальной эталонной модели S_0 и проведения лишь S сравнений с её использованием в качестве эталона. Отметим, однако, что подобные сравнения с эталонной неявно основаны на предположении, что эта эталонная модель действительно хороша — что может быть трудно гарантировать (особенно учитывая, что оптимальное значение T неизвестно, поэтому каждое число-кандидат t должно быть проверено).

В идеале, значение числа тем, “рекомендуемое” тем или иным методом, должно соответствовать ярко выраженному минимуму/максимуму на графике зависимости соответствующей функции качества тематической модели от числа тем. Этот факт мотивирует дальнейший анализ, в котором предпринимается попытка найти и классифицировать глобальные оптимумы алгоритмически следующим образом.

Во-первых, отмечаются значения в самой высокой и самой низкой точках графика (h и l). Далее, выбираются все точки, значения которых попадают в отрезок $[h - \alpha(h - l), h]$ (или $[l, l + \alpha(h - l)]$, если оценка должна быть минимизирована). В нашем анализе $\alpha = 0.07$. Во-вторых, проверяется, являются ли эти точки соседними друг с другом (если да, то оптимум один и устойчив; в противном случае кривая либо прыгает, либо имеет

несколько выраженных локальных оптимумов). Кроме того, проверяется, был ли оптимум достигнут на границе исследуемого диапазона чисел тем.

Исходный код методов поиска оптимального числа тем, а также код проведённых экспериментов находятся в открытом доступе.[9]

Исследуемые тематические модели В этом подразделе описываются тематические модели, использованные в экспериментах.

PLSA. Вероятностный латентный семантический анализ (Probabilistic latent semantic analysis, PLSA) [37] — это простая тематическая модель без каких-либо дополнительных гиперпараметров, кроме T.

LDA. Латентное распределение Дирихле (latent Dirichlet allocation, LDA) — известная тематическая модель, с априорным распределением η для Φ, и априорным распределением α для Θ (η и α могут быть числами или векторами). В рамках текущей работы в технологическом стеке TopicNet/BigARTM было реализовано три варианта модели LDA: симметричная *double-symmetric* ($\eta = \alpha = \frac{1}{T}$); асимметричная *asymmetric* (по рекомендациям [65], используется симметричное априорное распределение для Φ и асимметричное для Θ: $\eta = \frac{1}{T}$, $\alpha_{td} = \frac{1}{\sqrt{t+T}}$, $0 \leq t \leq T$) и "эвристическая" *heuristic* (со значениями $\alpha = \frac{50}{T}$ и $\eta = 0.01$, которые были использованы в [132]).

Декоррелированные модели. Было показано, что LDA имеет тенденцию создавать коррелированные темы, когда T слишком велико или слишком мало [116]. Поэтому интересно исследовать модели, которые явно пытаются уменьшить попарные корреляции между темами. Простейшим примером является TWC-LDA [133], которая уже реализована в библиотеке BigARTM [43]. Мы рассматриваем три возможных регуляризационных коэффициента τ для декорреляции: 0.02, 0.05, 0.1 — при этом коэффициент γ равен 0 (т. е. используется так называемая "относительная регуляризация").

Разреженные модели. Другим свойством LDA является сложность получения разреженных моделей из-за сглаживающих априорных распределений [109] (точнее, из-за того, что априорное распределение Ди-

[9] github.com/machine-intelligence-laboratory/OptimalNumberOfTopics.

рихле в принципе не предполагает векторов с нулевыми компонентами). Некоторые информационные метрики качества по-разному относятся к разреженным и сглаженным тематическим моделям [109], поэтому важно включить разреженные модели в анализ. Простейшая разреженная модель делит темы на две категории: фоновые (общие, содержащие много стоп-слов, неинформативные) и предметные (доменные, специфические, содержательные), которые намного более разрежены по сравнению с фоновыми [128]. BigARTM поддерживает такие модели [134] с помощью комбинации четырех регуляризаторов: двух регуляризаторов, сглаживающих ϕ_{wt} и θ_{td} для общих тем; и двух регуляризаторов, разреживающих ϕ_{wt} и θ_{td} для доменных тем. В качестве абсолютных значений регуляризационных коэффициентов для сглаживающего и разделяющего регуляризаторов перебираются значения 0.05 и 0.1 (коэффициенты регуляризации снова относительные — абсолютные считаются по этим с поправкой на конкретный датасет). Кроме этого, в качестве примера модели, в которой разреженность является исходным, прирождённым свойством, исследуется тематическая модель без матрицы Θ (TARTM), разработанная на базе TopicNet/BigARTM и представленная в [135].

Разреженные декоррелированные модели. Для полноты картины рассматривается модель, в которой объединяется несколько ограничений — которая в результате является одновременно и разреженной, и декоррелированной. Аддитивная регуляризация тематических моделей даёт возможность обучать подобные модели путём комбинации вместе (или в нужном порядке) отвечающих за то или иное свойство модели регуляризаторов [8]. Таким образом, разреженная декоррелированная модель получается по сути из PLSA и объединения в процессе обучения двух регуляризаторов.

Используемые датасеты Краткая информация об используемых в работе датасетах представлена в таблице 6. Тексты были лемматизированы (с помощью пакета `pymorphy2` [136] для русскоязычного текста и `NLTK` для английского). Для удаления часто встречающихся неинформативных слов использовались заранее составленные списки стоп-слов.

Таблица 6 — Истинные значения чисел тем в исследуемых датасетах.

Dataset	D	W	T_{expected}	$T_{\min}$	$T_{\max}$
WikiRef220	220	4839	5	2	20
20NG	18846	2174	15–20	3	40
Reuters	10788	5074	90	5	150
Brown	500	7409	10–20	5	25
StackOverflow	895621	3430	40	5	60
PostNauka	3404	8417	15–30	5	50
RuWiki-Good	8603	236018	10/90	5	100

WikiRef220,[10] датасет который был представлен в [123], состоит из 220 новостных статей, каждая из которых ссылается на определённую статью Википедии. Документы разделены на 16 различных групп в зависимости от статьи, на которую дана ссылка, но только 5 групп содержат более 5 записей. Следуя этой линии рассуждений, авторы описывают этот датасет как имеющий 5 тем и шум.

PostNauka — это корпус из 3404 статей популярного российского научно-популярного онлайн-журнала “ПостНаука”.[11] Каждый документ помечен рядом тегов, что позволяет оценить разумное количество тем как лежащее в диапазоне $[10, 30]$. В результате предыдущего исследования этой коллекции была получена тематическая модель, состоящая из 19 предметных тем [6].

20NewsGroups (20NG), **Brown** и **Reuters** — известные в NLP коллекции. По общему мнению исследователей, 20NG состоит из 15-20 тем, Brown — из 10-20 тем, а Reuters — из 50-100 тем.

StackOverflow — известный сайт вопросов и ответов, посвящённый в основном программированию. Были проведены различные исследования по поиску хороших тем на StackOverflow [137]. И в работе используется уже предобработанная версия корпуса из [63], состоящая из 895,621 документов.

Русская Википедия. Представляется новый датасет под названием “RuWiki-Good”, доступный через библиотеку TopicNet [8].[12] Для его полу-

[10] multisensorproject.eu/achievements/datasets.

[11] postnauka.ru.

[12] А также по ссылке: huggingface.co/datasets/TopicNet/RuWiki-Good.

чения использовался дамп базы данных русской Википедии, из которого было извлечено 8603 статей, каждая из которых попадала в одну из категорий: “избранные”, “хорошие” или “добротные”. Статьи этого корпуса отличаются проработанной иерархией меток: каждая статья попадает в одну из 11 основных категорий[13], причём каждая из них подразделяется на различное количество подкатегорий (например, “История” → “История Великобритании” → “Убийства в Соединённом Королевстве”). Поэтому автор работы верить, что данный датасет может быть также ценным “полигоном” для тестирования идей, связанных с гранулярностью тем.

2.3.4 Результаты и обсуждение

Результаты проведённых экспериментов сведены в таблицу 7. Для того чтобы получить осмысленное представление о способности рассматриваемых метрик качества служить для определения числа тем, было разработано три признака, характеризующие их поведение. Хотелось оценить способность метрики давать оценку количества тем независимо от случайной инициализации модели, “читаемость” полученных графиков зависимости метрики от числа тем, и точность метрики в плане предсказания ожидаемого числа тем.

Первый столбец таблицы 7 есть метрика Жаккара, которая рассчитывается следующим образом: для каждой случайной инициализации находится оптимальное значение числа тем или диапазон значений в соответствии с особенностями метрики (например, убывает они или возрастает при улучшении модели). Затем вычисляется расстояние Жаккара между пересечением и объединением этих множеств чисел тем (по всем экспериментам для одного датасета, но с разными инициализациями модели), с исключением случаев, когда метрика указывает на границы исследуемого интервала.

[13]Биология, география, наука, искусство, история, культура и общество, личности, религия и философия, спорт и развлечения, технологии, экономика.

Таблица 7 — Сравнение метрик по применимости к определению числа тем в среднем по датасетам. Некоторые метрики, основанные на разнообразии тем, удалены для краткости; в целом их эффективность представляется неудовлетворительной.

Score	Jaccard	Informativity	Expected
AIC	0.280	0.542	0.578
AIC sparse	0.219	0.111	0.100
BIC	0.128	0.444	0.461
BIC sparse	0.274	0.164	0.128
MDL	0.096	0.488	0.414
MDL sparse	0.282	0.428	0.256
renyi-0.5	0.470	0.507	0.425
renyi-1	0.356	0.475	0.394
renyi-2	0.230	0.299	0.183
D-Spectral	0.456	0.144	0.083
D-avg-L2	0.682	0.250	0.119
D-cls-H	0.595	0.245	0.189
D-avg-JH	0.302	0.053	0.022
lift	0.383	0.123	0.033
holdout-perplexity	0.228	0.025	0.019
perplexity	0.218	0.023	0.014
CHI	0.277	0.157	0.008
SilhC	0.233	0.079	0.028
average coherence	0.780	0.472	0.208
uni-theta-divergence	0.470	0.197	0.047

Во втором столбце приведена доля, показывающая, сколько раз результаты метрики были “читаемые”, то есть попадали в одну из категорий:

- Имеют ярко выраженное минимальное/максимальное значение/значения.
- Имеют интервал/ы вокруг минимального/максимального значения.
- Имеет область чередующихся пиков.

Все остальные типы поведения метрики можно описать как не зависящие от количества тем, либо не имеющие ни одного из описанных выше типов

поведения (имеющие оптимальное значение тем вне рассматриваемого в эксперименте диапазона).

Последний столбец представляет собой среднее значение булевской величины: было ли ожидаемое число тем датасета в диапазоне оптимальных значений, полученных из графика метрики для данной модели. Результаты, приведенные в таблице 7, ставят под сомнение представление о том, что количество тем является чётко определённым свойством конкретного корпуса (или, по крайней мере, что существующие способы оценки качества тематических моделей подходят для его определения).

Наблюдения и комментарии Приведём некоторые общие наблюдения, выявленные в ходе экспериментов.

Зависимость от модели. Первое наблюдение заключается в том, что "оптимальное количество тем" зависит от используемой тематической модели. На число тем может повлиять и класс модели, и даже конфигурация гиперпараметров в используемой модели. Рисунки 17 и 18 демонстрируют это на примере датасета WikiRef220, но описанное характерно и для других коллекций.

Рандомизация. Ещё один момент, который следует отметить — это расхождение, вызванное случайной инициализацией (и этот вопрос ещё более обострится, если обучение модели зависит от порядка документов). Как уже отмечалось, эксперименты проводятся с тремя случайными инициализациями. Если рассматривать каждую кривую отдельно (а не усреднять их все три вместе), то часто оказывается, что их поведение отличается. Наиболее частый случай — частично перекрывающиеся и соседние пики (например, `seed_0` дает максимум при 15 темах, `seed_1` также дает 15 тем, но `seed_2` дает 14 тем). Общий случай разных, но соседних пиков также легко поддается анализу, но встречаются и проблемные случаи, когда пики на кривых от разных инициализаций значительно отделены друг от друга, или когда лишь некоторые траектории имеют заметные пики.

Естественным решением в таком случае кажется построить больше моделей со значениями T, расположенными в интересующей небольшой области, и выбрать то, которое является наилучшим в среднем или наилучшим "с запасом" (best in the worst case). Однако такой подход являет-

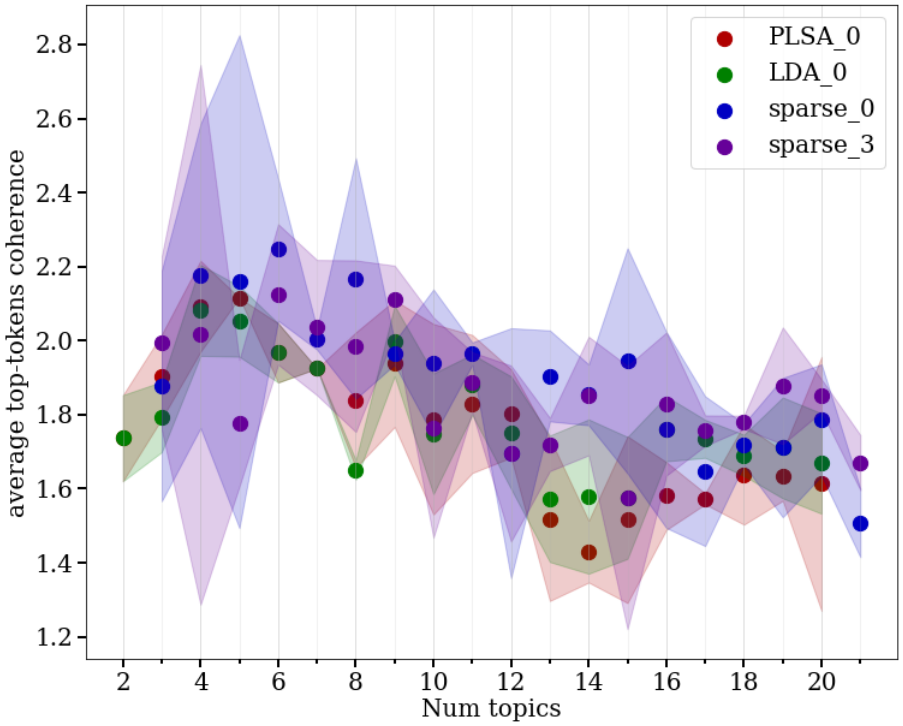


Рисунок 17 — Средняя когерентность каждой темы, $1 < T < 21$. Изображены модели LDA с симметричным априорным распределением (symmetric), LDA с эвристическим априорным распределением (heuristic), и разреженная модель `sparse_0`.

ся в некоторой степени “сам себя изобличающим”, поскольку специалисты обычно ищут единственную “лучшую” тематическую модель. И потому практика, при которой обучается статистически значимое количество различных тематических моделей, определяется подмножество моделей с “лучшим” гиперпараметром T, а затем выбирается произвольная модель из этого подмножества, считаемая “лучшей” — представляется небезупречной.

Направление, указанное, например, в исследованиях [9; 106], представляется более перспективным подходом к решению описанной проблемы расхождения результатов в зависимости от инициализации. Во-первых, необходимо создать несколько различных тематических моделей. Во-вторых, необходимо извлечь из этих тематических моделей темы, которые считаются “хорошими” (интерпретируемыми или хотя бы когерентными) или “сильными” (strong) (надёжными, воспроизводимыми). Эти темы сохраняются для последующего анализа (при этом следует также позаботиться о том, чтобы все сохранённые темы были достаточно

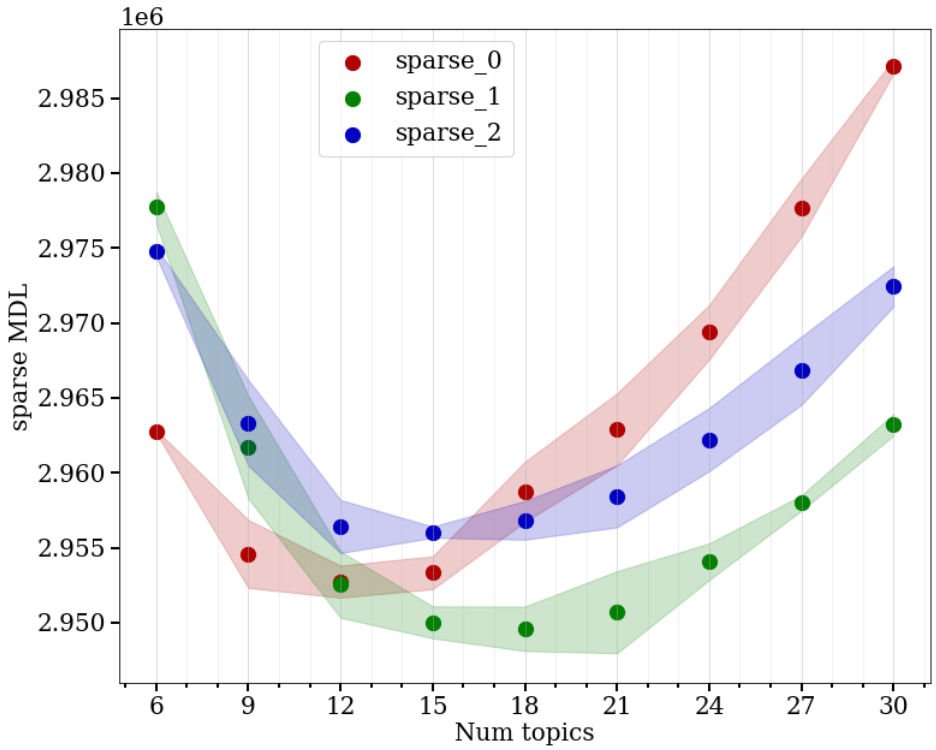


Рисунок 18 — Разреженный MDL критерий для моделей с различными значениями гиперпараметра, отвечающего за разреженность.

уникальными). Когда процесс извлечения тем сходится, набор различных “хороших” (или “сильных”) тем, найденных таким образом, образует искомую “лучшую” тематическую модель. А значение оптимального числа тем T является побочным продуктом этого процесса — как число тем в получившейся модели.

Несогласие методов. Вопрос о согласии можно разделить на два: 1) Согласуются ли различные критерии друг с другом? 2) Согласуются ли вариации одного и того же критерия?

Ответ на первый вопрос по результатам экспериментов можно дать отрицательный. Как правило, значение числа тем, определяемое методами, основанными на разнообразии, в несколько раз превышает значение, определяемое другими методами. Различия между другими критериями не столь значительны, но часто бывают существенными. Единственным исключением является датасет WikiRef220, где регионы чисел тем перекрываются, как видно из рисунка 19.

На второй же вопрос не получается дать убедительного ответа. Значения чисел тем, полученные схожими методами, как правило, отличаются незначительно. Однако часто бывает так, что какая-то одна вариа-

ция критерия даёт ответ, а несколько других — нет. Поэтому рекомендуется изучать несколько взаимосвязанных показателей. Изменение метрики среди методов, основанных на разнообразии, не влияет на расположение пиков, но влияет на их выраженность (что может и повлиять на итоговый результат, так в качестве рекомендуемого значения T выбирается самый высокий пик). При этом евклидова метрика представляется наименее информативной из всех.

Проблемы объективности. Используемый в работе подход, основанный на поиске единственного ярко выраженного минимума/максимума на графике зависимости какой-либо функции качества тематической модели от числа тем не учитывает другие, менее надёжные особенности данных. Такие, как, например, местоположение первого пика, место, где кривая выравнивается (образуя плато), угловые точки и точки перегиба — которые также могут быть полезны на практике. Однако использование этих признаков вносит слишком много субъективности и шума, поэтому они не подходят для выбора оптимального значения T как объективного, абсолютного свойства корпуса. Поэтому решено было придерживаться более простого подхода к анализу кривых зависимостей от T.

Возможно, некоторые плохо показавшие себя методы можно будет улучшить, если дать какие-то рекомендации для каждого из таких случаев. Так, например, когда кривая уплощается, достаточно вычесть некоторую линейную от T функцию, чтобы получить ярко выраженный экстремум. Однако подобные вопросы оставляем за рамками текущего исследования.

Свойства исследуемых критериев Приведём некоторые наблюдения, относящиеся к внутренними критериями качества, использовавшимся для поиска оптимального числа тем.

Разнообразие. Значение T, полученное с помощью этого метода, значительно превышает ожидаемое количество тем. Для большинства датасетов оптимум вообще находится за пределами исследуемого диапазона чисел тем. Кривая имеет тенденцию уплощаться, а не выходить на выраженный максимум (что иногда исправляется путём рассмотрения среднего расстояния до ближайшей темы вместо среднего попарного расстоя-

ния между темами). Расположение оптимума может меняться при разных случайных инициализациях тематической модели.

Информационно-теоретические. Эти методы лучше использовать в сочетании, так как любой один из них часто не справляется с предсказанием T для некоторых моделей и датасетов. Однако в совокупности они обычно дают разумные значения T (которые, однако, всё-таки несколько отличаются от "золотого стандарта"). Местоположение оптимума оказывается на удивление стабильным при разных случайных инициализациях.

Энтропия. Эти метрики дают ярко выраженные оптимумы, устойчивые к случайным инициализациям. Однако расположение минимума существенно зависит от значения гиперпараметра метода ε_0, и "дефолтное" значение $\varepsilon_0 = (W)^{-1}$ не приводит к ожидаемым значениям числа тем.

Кластеризация. Коэффициент Силуэта и индекс Калинского – Харабаша почти никогда не дают близкой к ожидаемой оценки T. Возможно, пространство признаков, обусловленное Θ, плохо подходит для кластерного анализа. Это предположение подтверждается работой [121].

Спектральная дивергенция. Этот метод очень шумный и сложный в применении. Он не может дать никакой оценки, когда применяется к разреженным моделям (кривая просто монотонно спадает).

Когерентность. Этот метод очень шумный и сложный в применении. Видно, что средняя когерентность является убывающей функцией от T, при этом часто флуктуирующая. Следовательно, обычно глобальный максимум достигается на малых T, что делает когерентность непригодной для выбора числа тем T.

Перплексия. Почти во всех случаях перплексия оказывается монотонной, без каких-либо заметных особенностей, полезных для выбора T (см. рисунок 20). Такое поведение противоречит некоторым работам других исследователей, в которых перплексия имела ярко выраженный локальный минимум при некотором числе тем. Автор предполагает, что это может быть связано с некоторыми деталями реализации, такой как, например, обработка слов, не входящих в словарь W. Скорость изменения перплексии часто даёт устойчивые пики и плато, но они не совпадают с ожидаемым значением числа тем. Числовой результат сильно зависит от модели.

Лифт. Для этой метрики оптимальное количество тем соответствует максимальному значению на графика. И на большинстве датасетов это максимальное значение оказывается далеко за пределами ожидаемого диапазона для оптимального числа тем. Однако для датасета StackOverflow наблюдаются ярко выраженные максимумы для разных семейств моделей (см. рисунок 21). Подозреваем, что так получилось за счёт агрессивной фильтрации токенов в исходном датасете.

Стабильность. Авторы этого подхода предлагают искать минимум нестабильности или один из нескольких локальных минимумов в качестве индикатора оптимального количества тем. Однако в проведённых по данной работе экспериментах не всегда наблюдалось желаемое поведение на всех текстовых коллекциях, и приходилось довольствоваться плато или просто снижением темпов роста нестабильности. Кроме этого, с подходом по стабильности обнаружились также следующие проблемы. Во-первых, она становится слишком шумной для моделей с небольшим количеством тем (менее 10–15). Во-вторых, стабильность плохо подходит для разреженных моделей: полученные результаты слишком зашумлены, чтобы сделать вывод о количестве тем. В-третьих, оценка оптимального числа тем, полученная с помощью стабильности, обычно оказывается ниже ожидаемого числа тем.

2.3.5 Заключение

Анализ показывает, что внутренние методы оценки качества тематической модели далеко не всегда являются надёжными и точными инструментами в поиске оптимального числа тем.

Из провелённых экспериментов видно, что наилучшие результаты показали самые простые подходы: AIC, BIC, MDL, renyi. Эти метрики выносят своё суждение на основе грубой оценки состояния модели, в отличие от других критериев, ищущих ответ по более тонкой структуре моделей на уровне тем. Более сложные методы (lift, coherence, diversity), кото-

рые пытаются напрямую измерить некоторые желаемые качества тематической модели, не дают удовлетворительных результатов.

Видно, что многие подходы к определению числа тем дают в качестве ответа несколько решений или даже диапазон чисел тем. Это противоречит наивному представлению об оптимальном T как о единственном фиксированном значении, присущем корпусу текстов. Такое поведение может указывать и на другую проблему в области: понятие темы плохо определено с точки зрения гранулярности. Каждая тема может быть разделена на подтемы, возможно, без ухудшения внутренних метрик, таких как различность тем или качество кластеризации. Отмечая, что не удалось найти методы, согласующиеся друг с другом на почти всех датасетах, приходим к выводу, что понятие "оптимальное число тем" может включать в себе несколько мифов, требующих более глубокого рассмотрения.

"Классический" ("школьный") взгляд на тематическую модель как на извлекающую "скрытые" (латентные) распределения слов в темах приводит к представлению о том, что темы реально существуют в коллекции и могут быть найдены вне зависимости от подхода к их поиску. Согласно этой точке зрения, текстовая коллекция содержит информацию об истинном числе порождающих её распределений. Это число не должно меняться в зависимости от типа модели, ищущей эти распределения, или от конкретной внутренней метрики, считающей или нет это число "оптимальным". Однако в проведённых экспериментах это оказалось не так. Наоборот, оказалось, что количество тем в основном зависит от модели и частично определяется реализованным подходом к определению оптимального количества тем. Если из полученных результатов можно извлечь урок, то он заключается в том, что понятие "число тем" было неверно понятно сообществом, здесь нет ничего "истинного", это — просто ещё один гиперпараметр модели, который нужно настраивать.

В свете этих соображений видно несколько способов, с помощью которых сообщество уже решает эту проблему:

- Выбор модели в соответствии со вторичной задачей.
- Построение иерархии тем и её последующий прунинг.
- Улучшение процесса человеческого (полу-)контроля. Примером такого подхода может служить предложение от [127], заключаю-

щееся в использовании слабого контроля (weak supervision) со стороны пользователей для настройки гиперпараметра порога, определяющего количество тем.

Предлагается также обратить внимание на следующие направления, поскольку они могут оказаться полезными для решения вопроса о числе тем:

- Устранение гиперпараметра T.
- Разработка тематических моделей, более устойчивых к изменению T. Например, рассмотрим гипотетическую процедуру, позволяющую построить тематическую модель для заданного количества тем, в которой все темы будут интерпретируемыми. Такая процедура сделает вопрос определения T в значительной степени неактуальным.
- Алгоритм обнаружения новых событий с последующим автоматическим изменением T.

Рассмотренные в данной работе способы оценки “оптимального числа тем” с помощью внутренних критериев качества тематических моделей не внушают доверия. Приходим к выводу, что оптимальное количество тем зависит не столько от корпуса текстов, сколько от самого метода определения количества тем и используемой для этого тематической модели, или даже от цели, для достижения которой применяется тематическое моделирование.

Пока предложенные выше методы остаются скорее фантастикой, чем реальными алгоритмами, предложим читателям рассматривать T просто как ещё один гиперпараметр модели. Полученные результаты показывают, что практикам не следует пытаться оценить “естественное” число тем, заложенных в корпусе; вместо этого следует сосредоточиться на вопросах, подобных следующим:

- Сколько документов в среднем должна содержать тема?
- Какая должна быть степень гранулярности тем?
- Если доступны внешние метки: как включить эти дополнительные знания в модель?
- Достаточно ли уникальны найденные моделью темы?

В итоге приходим к выводу, что основной целью тематического моделирования должен быть не поиск “оптимального” числа тем, а поиск такого метода обучения тематической модели, который при любом заданном заранее числе тем приводит к модели, все темы которой при отсутствии внешнего критерия являются интерпретируемыми.

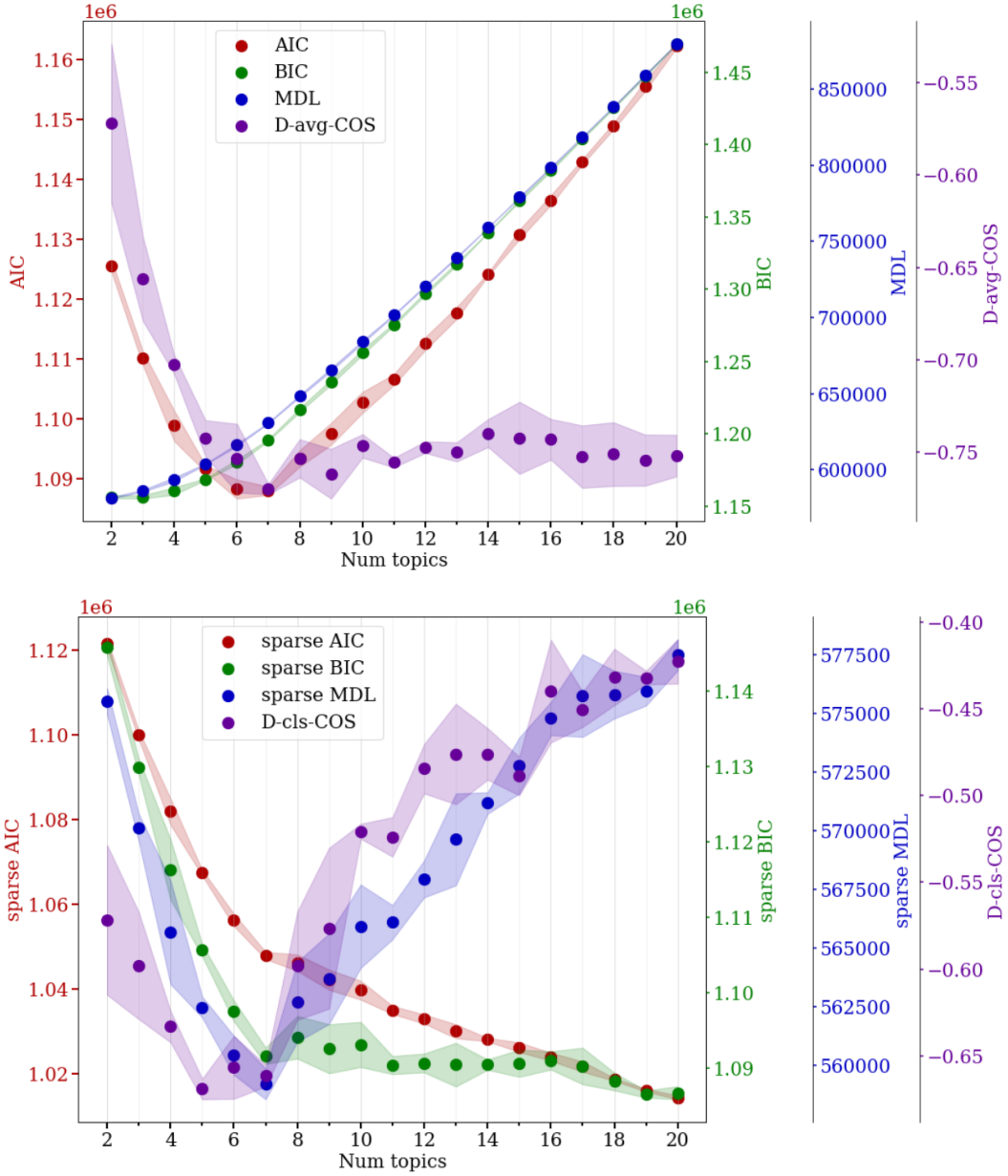


Рисунок 19 — Несколько метрик качества для исследования различных T для модели PLSA, $1 < T < 21$. Изображены метрики AIC, MDL, и косинусное расстояние (взятое с отрицательным знаком, так что минимум соответствует “лучшему” значению). Видно, что все метрики сходятся на том, что 7 является разумным значением для числа тем T.

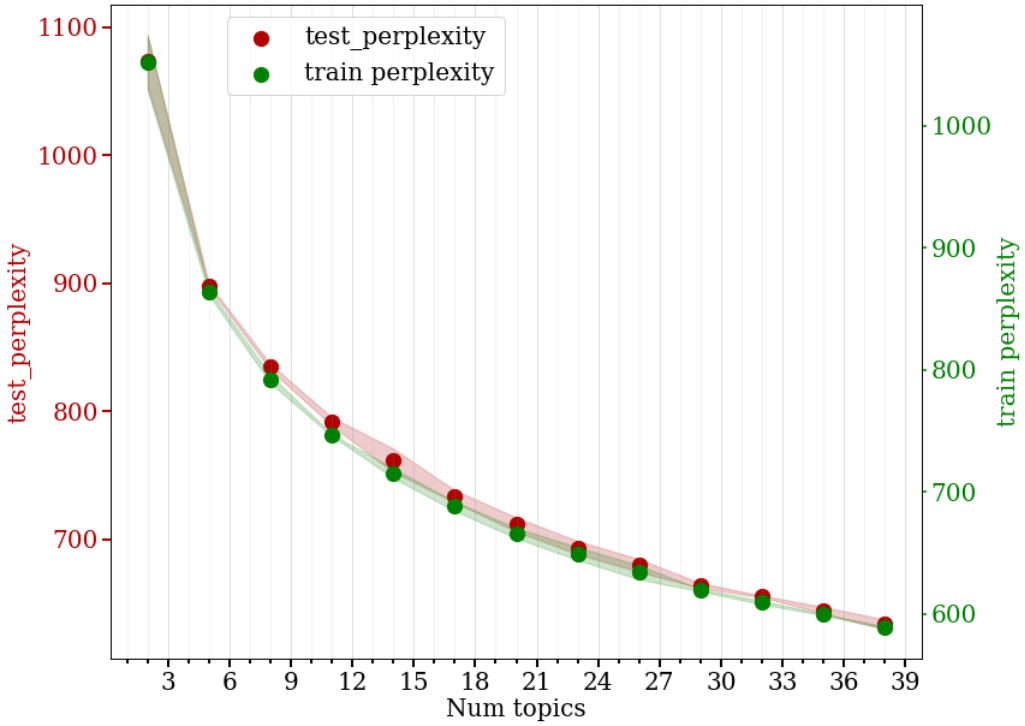


Рисунок 20 — Сравнение перплексии на отложенной выборке и на обучающей для модели LDA. Аналогичное поведение наблюдалось для всех рассматриваемых датасетов.

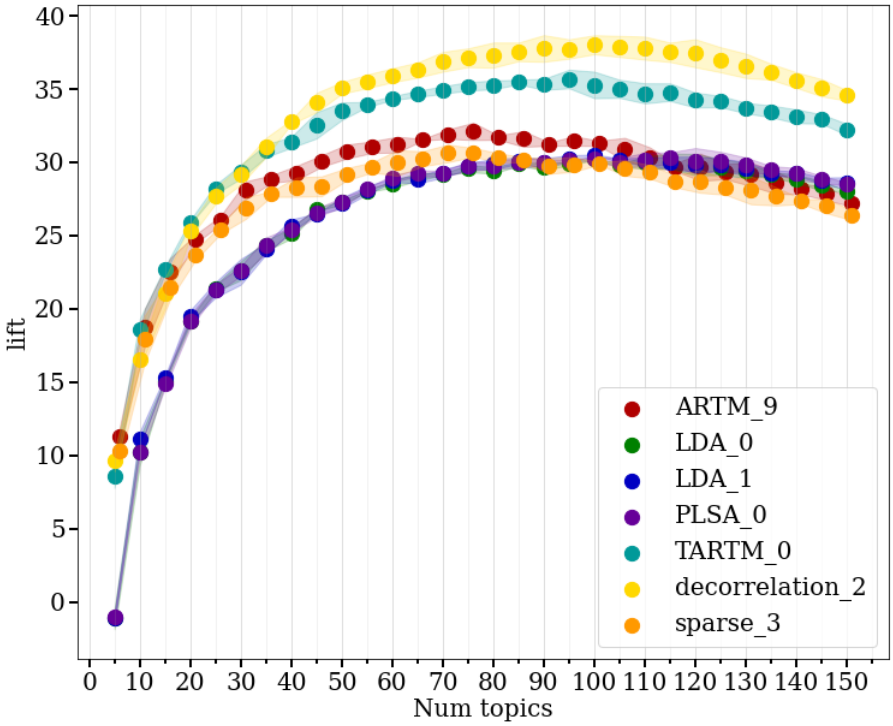


Рисунок 21 — Множественные максимумы метрики Lift для разных типов тематических моделей.

Глава 3. Неустойчивость и неполнота тематических моделей

3.1 Проблема неустойчивости и неполноты тематических моделей

Современная литература по тематическому моделированию предоставляет множество тематических моделей для различных ситуаций [35]. Самыми базовыми моделями являются вероятностный латентный семантический анализ, или PLSA [138], и латентное распределение Дирихле, или LDA [57]. Сотни тематических моделей являются расширениями PLSA и LDA. Более того, есть не только разные тематические модели, но и целые разные парадигмы построения тематических моделей. Например, байесовский подход, начало которому положила LDA, когда сначала описывается вероятностная генеративная модель данных, затем задаются априорные распределения параметров модели, а затем с помощью байесовского вывода получаются апостериорные распределения параметров. Другой подход — аддитивная регуляризация тематических моделей, или ARTM [5], в основе которой лежит максимизация логарифма правдоподобия коллекции вместе со взвешенной суммой регуляризационных критериев. Каждый регуляризатор представляет собой некоторое желаемое свойство тематической модели и, следовательно, ограничивает количество возможных решений задачи тематического моделирования — тематических моделей. Однако, какую бы тематическую модель ни выбрал исследователь, каждая такая модель по своей сути *неполна* и *неустойчива*.

Неполнота тематических моделей означает, что нет никакой гарантии, что одна тематическая модель, какая бы хорошая они ни была, сможет идеально найти скрытую тематическую структуру текстовой коллекции.

Нестабильность означает, что качество тематических моделей зависит от многих вещей. Во-первых, необходимо сказать, что идеи и гипотезы, принимаемые в тематическом моделировании, в конечном итоге позволяют свести исходную абстрактную задачу поиска тем в документах к практической задаче матричного разложения, которая решается итера-

ционным алгоритмом. Однако задача матричного разложения *некорректно поставлена*: она имеет бесконечно много решений. Одним из возможных способов преодоления этого является, например, применение регуляризации, которая накладывает дополнительные ограничения и даже приводит к лучшему итоговому решению [5]. Кроме того, результат итерационного алгоритма зависит от инициализации модели [4]: разные инициализации Ф могут приводить к разным итоговым темам. Некоторые темы могут быть одинаковыми для многих тематических моделей с разными инициализациями, некоторые же требуют определённой инициализации модели, а некоторые темы при этом вообще могут быть неинтерпретируемыми: включать слова из несвязанных областей [2; 100]. Многие исследователи работают над оценкой стабильности тематических моделей [139—141].

Выбор лучших гиперпараметров также связан с устойчивостью модели. Например, настройка весов регуляризаторов влияет на тематические модели [2]. Такой гиперпараметр, как количество тем, влияет не только на качество получаемых тем с точки зрения интерпретируемости, но и на возможность решения других задач с использованием векторных представлений слов и документов, полученных с помощью тематической модели [42].

Как уже отмечалось, тематическое моделирование имеет множество применений. Нас же будет интересовать следующее: исследование данных — когда исследователь просто хочет найти все темы, представленные в коллекции текстов. Таким образом, основным желанием является построение такой тематической модели, все темы которой были бы интерпретируемыми, различными и в совокупности бы идеально описывали данные. Но в виду упомянутых проблем неустойчивости и неполноты тематических моделей эта задача по исследованию данных может быть не решена до конца.

3.2 Банк тем: валидация тематических моделей через их множественное обучение

Вероятностное тематическое моделирование — это инструмент для выявления без учителя скрытой тематической структуры текстовой коллекции естественного языка. Получая на вход лишь текстовое содержание документов, тематическая модель стремится найти в нём темы как вероятностные распределения на словах. Недостатками тематических моделей являются их неустойчивость — в том смысле, что темы могут зависеть от, например, случайной инициализации модели — и неполнота — в том смысле, что каждый новый запуск модели на одной и той же коллекции в принципе может помочь обнаружить какие-то новые, не виденные до этого темы. Это означает, что исследование данных с помощью тематического моделирования обычно требует довольно много экспериментов по просмотру множества обученных тематических моделей и настройке их гиперпараметров — в поисках модели, которая описывала бы данные наилучшим образом. Чтобы справиться с неустойчивостью и неполнотой тематических моделей, предлагается постепенно накапливать найденные в процессе множественного обучения тематических моделей хорошие (интерпретируемые) темы в “банк тем”. Решение о добавлении или нет новой темы в банк выносится на основе изучения связи между темами вновь обученной модели и темами, уже сохранёнными в банке — связей, которые выявляются благодаря двухуровневой иерархической тематической модели, родительский уровень которой есть темы банка тем, а дочерний состоит из новых тем. Такой анализ помогает выявить и исключить из добавляемых в банк плохие и дублирующие темы. Представляется новый способ оценки качества тематической модели — путём сравнения найденных моделью тем, с темами, которые содержатся в готовом банке тем. Эксперименты с несколькими датасетами и тематическими моделями показывают, что предложенный метод валидации тематических моделей с помощью банка тем действительно помогает находить модель с более хорошими темами.

3.2.1 Введение

Основной посыл данного раздела заключается в том, что постепенный сбор интерпретируемых тем с помощью множественного обучения тематических моделей может привести к более качественному и тщательному исследованию данных методами тематического моделирования, чем просто случайные попытки обучить сразу лучшую модель путём изменения её гиперпараметров. На это можно смотреть как ка способ проведения эксперимента по тематическому моделированию. Итак, в результате многократного обучения моделей можно получить *подмножество интерпретируемых тем, содержащихся в датасете*, которое далее будем называть *банком тем*. С тем чтобы показать, что предложенный метод действительно помогает лучше исследовать данные, можно задаться вопросом: существует ли такой способ обучения тематических моделей, который бы приводил к наилучшему качеству тематической модели, *оценённому с помощью собранного банка тем*. Для исследования берутся несколько текстовых коллекций на естественном языке, далее по каждой коллекции: создаётся банк тем и обучается ряд тематических моделей. Гипотетически, поскольку рассматриваемые датасеты по сути своей похожи, должна существовать такая модель, которая бы хорошо описывала по крайней мере большинство датасетов. Это может быть не самая лучшая модель для каждой коллекции, она просто должна быть лучше других рассматриваемых моделей. Если банк тем может найти такую модель, то это значит, что банк тем — не просто интуитивно лучший способ проведения экспериментов, он помогает в оценке моделей, облегчая исследователю поиск оптимальной для рассматриваемых данных.

Резюмируя, вклад работы можно обозначить следующим образом:

- Предлагается собирать интерпретируемые темы в процессе множественного обучения тематических моделей, с тем чтобы сформировать *банк тем*.
- Предлагается алгоритм создания банка тем для данной текстовой коллекции. При сборе тем для банка обращается внимание на следующее: темы, добавляемые в банк, должны быть хорошими и

различными. Качество темы при этом оценивается по значению её когерентности.

– Представляется метод сравнения тем двух тематических моделей. Метод основан на использовании двухуровневой иерархической тематической модели, каждый уровень которой представляет темы одной из сравниваемых тематических моделей. Это позволяет не только оценить близость тем разных моделей, но и понять, можно ли, например, считать тему одной модели родительской для темы другой.
– Показывается, что собранный банк тем может быть использован для автоматической оценки качества вновь обученных тематических моделей.

В следующих подразделах более подробно описываются: понятие банка тем, как можно создать такой банк тем с помощью множественного обучения тематических моделей, как оценить качество модели с помощью банка тем, а также эксперименты на нескольких текстовых коллекциях.

3.2.2 Мотивация и связанные работы

Как уже отмечалось, из-за *неустойчивости* тематических моделей, такая задача, как *исследование данных*, поиск тем в коллекции документов, может занимать много времени. Помимо предварительной обработки данных, приходится обучать несколько тематических моделей и подбирать гиперпараметры, оценивая качество итоговых тем или сравнивая между собой темы обученных моделей. Однако даже после получения достойной модели нельзя гарантировать, что её темами представлены абсолютно все темы, которые есть в текстовой коллекции, поскольку тематические модели ещё *неполны*. Существует множество работ, посвящённых решению проблемы неустойчивости и неполноты тематических моделей. Неустойчивости обычно преодолевается путем внесения некоторых изменений в процесс инициализации или обучения модели, что кажется ра-

зумным. Другое дело с неполнотой: существует множество эвристических способов борьбы с ней, иногда не совсем очевидных и интерпретируемых, но все они упускают из вида простую идею поиска всех тем в наборе данных, а затем использования этих тем *в неизменном виде* в качестве простого способа валидации новых тематических моделей. Ниже приведены краткие описания некоторых соответствующих работ, но подчеркнём ещё раз: преимущество подхода, предлагаемого в данной работе, заключается в простоте, интерпретируемости и общей возможности вовлечения человека в процесс.

Авторы [142] предлагают обучать несколько тематических моделей с разными инициализациями, а затем кластеризовать темы всех моделей, чтобы объединить похожие. При этом центры полученных кластеров тем могут быть выбраны в качестве начального приближения тем для тематической модели. В указанной статье выдвигается гипотеза, что размер тематического кластера (то есть то, как часто тема находилась разными тематическими моделями) тем больше, чем чаще эта тема встречается в документах коллекции.

Суть предложения, озвученного в [143], заключается в том, чтобы несколько раз обучать модели, а затем опять же провести кластеризацию на множестве тем всех моделей. Если тема хорошая, то, по мнению авторов, она будет часто повторяться в разных моделях и получится чёткий кластер экземпляров именно этой темы. С другой стороны, если тема шумная, то кластер с ней будет более неоднородным. Получается, это подход для выявления интерпретируемых тем, основанный на устойчивости. Однако если тема интерпретируема, но встречается только в одной модели, то предложенный в статье метод её не обнаружит [40].

В статье [63] авторы хотят найти лучшую тематическую модель для текстовой коллекции, сравнивая темы нескольких моделей и получая таким образом *оценку качества тематической модели относительно устойчивости*. Так, из всего множества документов выбирается подмножество, на котором обучается новая модель, темы которой запоминаются. Это повторяется несколько раз (в статье — 10 раз). Затем измеряется, насколько часто темы повторяются в моделях, путём сравнения тем моделей, обученных на разных подвыборках документов. Весь этот процесс, в свою

очередь, повторяется несколько раз (снова 10), а итоговой оценкой качества модели является медиана полученных оценок.

В работе [116] авторы хотят найти лучшую *тематическую структуру* (в частности, количество тем) для тематической модели LDA, оценивая её устойчивость. Оценка устойчивости модели вычисляется как среднее косинусное расстояние между каждой парой её тем. Далее авторы предлагают идею поиска оптимальной тематической структуры для LDA — адаптивно с использованием кластеризации тем и устойчивости модели. Однако оценка устойчивости модели в целом является плохо формализованной задачей, и весь процесс, описанный в статье, кажется сложным и плохо интерпретируемым.

Поскольку неустойчивость модели отчасти вызвана случайной инициализацией, один из способов повысить устойчивость — сделать инициализацию *осмысленной*, то есть из каких-либо соображений выбрать начальное приближение для модели лучше, чем случайное.

Например, в работах [144; 145] авторы вводят понятие *якорных слов* (anchor words) — слов, по которым можно сразу определить, относится ли документ к некоторой определённой теме или нет. Иными словами, якорные слова принадлежат только одной теме, являются “маркерами” этой темы. Требование о том, чтобы темы вообще имели такие якорные слова, накладывает дополнительные ограничения на решение задачи матричного разложения, лежащей в основе вероятностного тематического моделирования. Однако было показано [146], что задача поиска якорных слов проще, чем задача матричного разложения. А уже имея якорные слова, можно достичь локального минимума оптимизационной задачи матричного разложения всего за несколько итераций. Однако стоит отметить, что не все вообще темы могут иметь якорные слова: одного слова может быть недостаточно для определения темы, особенно если речь идет о темах с большим количеством слов общей лексики или родительских и дочерних темах. Использование же модели с якорными словами в качестве *инициализации* для матрицы Φ (4) обучаемой тематической модели кажется разумным, поскольку такие матрицы будут содержать, по крайней мере, *часть искомой тематической структуры*, уже какую-то информацию о темах коллекции. Стоит также отметить, что к определению якорных

слов можно подходить по-разному. Например, можно предположить, что у каждой темы может быть только одно якорное слово или что их может быть несколько для каждой темы.

В работе [147] авторы предлагают иной подход. Вводится понятие *контекста слова* — как слов, которые часто встречаются вместе с данным словом недалеко друг от друга в тексте. В основе такого понятия лежит следующая гипотеза: слова, наиболее точно характеризующие тему, обычно встречаются в тексте вместе, их взаимное расположение не является случайным. Это позволяет искать начальное приближение матрицы Ф следующим образом: разбить документы исходной коллекции на сегменты (например, абзацы или предложения). Оценить вероятности совместных близких встреч $p(w_1 \mid w_2)$ в тексте для всех пар слов, выбрать среди всех слов те, которые встречаются достаточно часто вместе *лишь с небольшим количеством* других слов, и кластеризовать векторы вероятностей совместных встреч таких слов. Если каждая тема представлена в тексте сегментами, то в результате кластеризации выделенные слова должны быть объединены в один кластер — тему. А центр этого кластера может аппроксимировать тему как столбец в матрице Ф.

Кроме вопросов о собственно неустойчивости и неполноте тематических моделей, когда тематические модели каким-то образом сравниваются между собой, важен и способ *оценки качества* одиночных тематических моделей и составляющих их тем. Причём особый интерес в плане практического применения представляют именно автоматически вычисляемые функции качества.

В работах [100—102] авторы предложили метод оценки итнерпретируемости тем, названный *когерентностью*: когда решение о качестве темы принимается на основе того, насколько неслучайно пары наиболее частых слов темы оказываются рядом друг с другом в тексте.

В работе [6] предлагается автоматических подход к оценке итнерпретируемости тем, названный *внутритекстовой когерентностью*. Когда при оценке качества темы учитывается распределение слов во всей коллекции, а не только совстречаемости небольшого числа наиболее вероятных слов темы. Идея такого подхода основана на гипотезе о сегментной структуре текста естественного языка, которая гласит, что темы пред-

ставлены в тексте в виде сегментов, а не слов, расположенных случайным образом. Метод оценки качества тем лишь по небольшому числу топ-слов даёт преимущество в скорости вычислений, а также в простоте и понятности всей идеи для человека (ведь люди тоже часто оценивают тему лишь по списку её наиболее частых слов). Недостатком же подхода к когерентности, основанному на совсречаемостях топ-слов, является учёт лишь сильно урезанной информации о теме — таким образом, бóльшая часть темы по сути не используется при оценке её качества. Более подробно различия в подходах между когерентностями обсуждаются в [6].

3.2.3 Идея Банка тем и Полного набора тем

Задача тематического моделирования считается завершённой, когда в исследуемой коллекции текстов найдены все скрытые темы. Однако не существует универсального способа обучения хорошей тематической модели для любой коллекции текстов. Чтобы получить хорошую тематическую модель, необходимо провести серию экспериментов, в результате оценивая качество каждой обученной тематической модели. Важной частью оценки качества тематической модели при этом является просмотр тем модели глазами, что требует времени.

Кроме того, тематические модели неустойчивы и неполны. Это приводит к тому, что, во-первых, некоторые темы, найденные моделью, могут быть плохого качества. Во-вторых, при проведении экспериментов с рядом тематических моделей одна и та же тема может появляться в некоторых, но не обязательно во всех тематических моделях.

Таким образом, можно выделить следующие две проблемы, возникающие при работе с тематическими моделями (точнее, при работе с *несколькими* тематическими моделями — несколькими в том числе и тогда, когда проходит подбор гиперпараметров для одной модели):

– Некоторые уже найденные интерпретируемые темы могут быть впоследствии потеряны.

– Тратится время на анализ тем, которые похожи на те, что уже встречались до этого.

Неустойчивость и неполнота обусловлены тем, что по своей природе задача тематического моделирования есть задача кластеризации. И чтобы справиться с этими проблемами, исследователи вынуждены вводить в процесс внешние критерии, например, оценивать разнообразие найденных тем или требовать, чтобы темы были согласованы в рамках нескольких процедур обучения модели при разных её случайных инициализациях. Такие приёмы в самом деле могут помочь в повышении качества результирующей тематической модели. Однако это не есть конечное и полное решение проблем неустойчивости и неполноты.

Принимая во внимание сказанное, кажется возможным организовать процесс поиска тем полуавтоматизированным образом, используя пользовательские критерии для отбора тем. Такой подход естественным образом формирует хранилище хороших и плохих примеров тем коллекции, отсюда и название: Topic Bank, или TopicBank. Таким образом, TopicBank — это своего рода *обёртка* над тематическим моделированием, когда информация о наборе данных накапливается постепенно (см. рисунок 22). Подчеркнём, что TopicBank также *не решает* проблем неустойчивости и неполноты тематических моделей. Однако TopicBank принимает эти проблемы во внимание и помогает справиться с ними — путём множественного обучения моделей. Основные функции TopicBank заключаются в следующем:

– Сбор интерпретируемых тем.
– Использование собранных тем для автоматической валидации вновь обученных моделей.

В данной работе под *хорошими темами* будем понимать интерпретируемые темы, которые являются осмысленными для человека. Однако понятие “хорошей темы” может иметь и какое-нибудь другое значение — в зависимости от задачи, для которой применяется тематическое моделирование.

Создаваемый программный модуль Банка тем призван обеспечить возможность получения множества хороших тем со следующими свойствами: отобранные хорошие темы

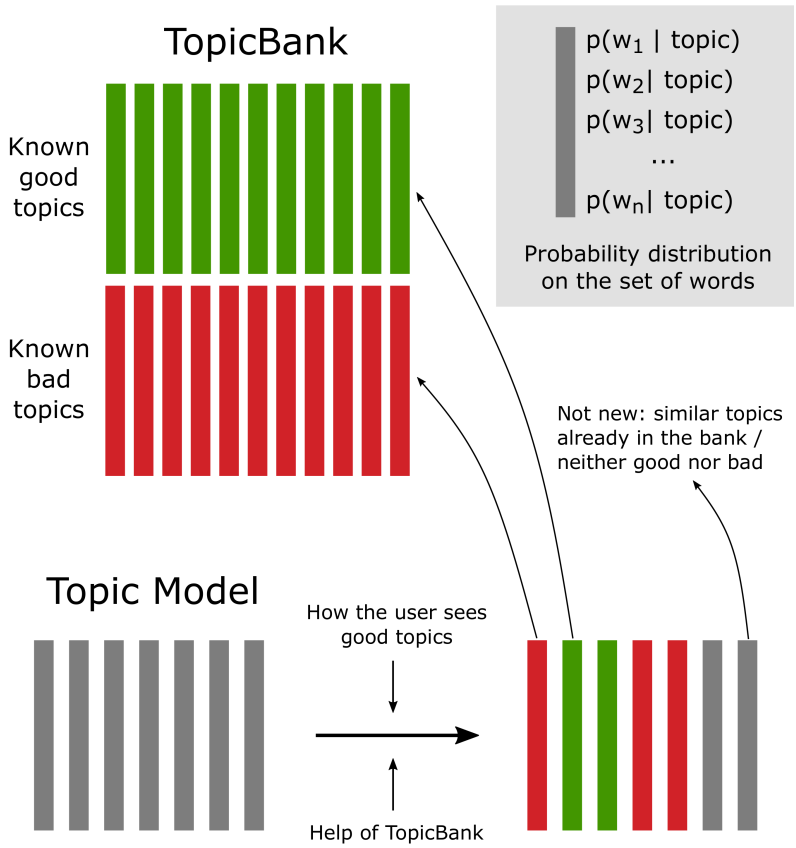


Рисунок 22 — Концепция Банка тем: в нём накапливаются хорошие и (опционально) плохие темы. Для работы Банку нужна информация от человека о том, как по теме понимать, является ли она хорошей (как определять хорошую тему). Это может быть либо автоматически вычисляющаяся функция качества, которая переводит тему в булевское значение, означающее, хорошая тема или нет; или же просто сам человек может оценивать темы.

- удовлетворяют внешней задаче
- являются разнородными, то есть не могут быть получены как линейная комбинация других тем
- являются решением задачи матричного разложения, максимизируют правдоподобие

Набор тем с такими свойствами будем называть *полным набором тем*. В дальнейших экспериментах сосредоточимся на разнообразии тем и на их интерпретируемости, определяемой когерентностью — как на внешней цели тематического моделирования.

Отметим на полях, что банк тем можно использовать как инструмент для оценки числа тем в коллекции, удовлетворяющих внешнему

критерию — прекращая добавлять темы в банк, когда разнообразие или перплексия имеющегося полного набора тем перестаёт улучшаться. Таким образом, банк тем может помочь как в поиске хороших тем, так и в определении их числа.

Дадим два определения сущности банка тем (два возможных взгляда на него).

Определение 3.2.1 (Банк тем как функциональный объект)**.** Банк тем — инструмент, способ работы с тематическими моделями, основные функции которого:

- хранение информации о просмотренных темах (вероятности слов; оценка качества темы; некоторые выявленные зависимости между темами, такие как похожесть, родственность)
- ускорение ручного анализа новых моделей (путём сортировки новых тем в порядке непохожести до ранее сохранённых в банк тем, то есть от более новых к, вероятно, похожим на ранее просмотренные темы)
- валидация в автоматическом режиме новых моделей (путём сравнения новых тем с темами, хранящимися в банке).

Определение 3.2.2 (Банк тем как математический объект)**.** О банке тем можно думать как о кортеже

$$B = \langle T, q, \rho \rangle$$

где T — множество тем, $q : T \to \{0, 1\}$ — булевская функция качества темы (“хорошая” тема или нет), $\rho : T \times T \to \mathbb{R}_+$ — функция расстояния между темами.

При этом определены операции добавления в банк новой темы:

$$\text{add} : \langle T, q, \rho \rangle, t \mapsto \langle T \cup \{t\}, q, \rho \rangle$$

И поиска для данной темы ближайшей темы из банка:

$$\text{nearest} : \langle T, q, \rho \rangle, t \mapsto \arg\min_{\tau \in T} \rho(\tau, t)$$

3.2.4 Эксперименты

Реализация банка тем была протестирована на выбранных тематических моделях и датасетах. В экспериментах хотелось продемонстрировать преимущества использования TopicBank для выбора наилучшей тематической модели. Далее мы более подробно рассмотрим аспекты эксперимента. Реализация TopicBank, а также ноутбуки с экспериментами находятся в открытом доступе[1].

Данные Для демонстрации работы TopicBank было выбрано несколько довольно известных текстовых коллекций, и несколько не очень известных. Подробности о датасетах можно найти в таблице 8.

Датасет ПостНауки[2], уже использовался в разделе про внутритекстовую когерентность, содержит научно-популярные тексты на самые разные темы — от социологии до астрофизики. Reuters [148] и Brown [149] — известные датасеты по тематическому моделированию.[3], Twenty Newsgroups (20 NG)[4] датасет хорошо известен в NLP-сообществе, благодаря своим сбалансированным категориям, которые часто рассматриваются в экспериментах как темы [150]. AG News [151] используется для классификации новостей (таким образом, классы можно считать темами). Habrahabr[5] есть популярная российская коллекция в основном технических статей, сгруппированных по темам в соответствии с существующей категоризацией сайта. Watan2004 [152] — арабский размеченный по темам датасет, используется для более широкого тестирования Банка тем, за пределами англо- и русскоязычных текстовых коллекций.

Будем создавать банк тем для каждой из указанных текстовых коллекций, чтобы в дальнейшем использовать его для оценки качества обученных на соответствующих коллекциях моделей.

[1] github.com/machine-intelligence-laboratory/OptimalNumberOfTopics.
[2] postnauka.ru.
[3] Датасеты доступны в рамках NLP библиотеки NLTK: nltk.org/book/ch02.html.
[4] Датасет доступен в рамках библиотеки Scikit-learn: scikit-learn.org/0.19/datasets/twenty_newsgroups.html.
[5] habr.com.

Таблица 8 — Датасеты, используемые в экспериментах ($|D|$ означает количество документов в датасете).

Название	$\lvert D\rvert$	Язык
PostNauka	3 446	Russian
Reuters	10 788	English
Brown	500	English
20 NG	18 846	English
AG News	127 600	English
Watan2004	20 291	Arabic
Habrahabr	133 978	Russian

Создание банка тем Банк тем возникает в результате обучения нескольких тематических моделей. В экспериментах по созданию банка общее число обучаемых моделей выставляется равным 20 моделям. Каждая модель обучается в процессе 100 обновлений матриц Φ и Θ (итераций) в рамках EM-алгоритма. На каждой итерации оценивается качество полученных тем по значению внутритекстовой когерентности [6]. Лучшие 10% тем с наибольшей внутритекстовой когерентностью считаются хорошими и помечаются как кандидаты на добавление в банк тем. Далее эти вновь полученные темы сравниваются не только между собой, но и с темами, уже имеющимися в банке тем. Банк тем помогает оценить, несут ли темы обученной модели новую информацию — а потому стоит ли их добавлять в банк тем. Решение по добавлению в банк зависит от отношений между темами вновь обученной модели и темами, уже присутствующими в банке.

Рассмотрим этот процесс более подробно. Пусть t — некоторая тема из банка, а s — некоторая новая тема. Тогда $p(s \mid t)$ — это вероятность того, что тема s является дочерней темой темы t. Эта вероятность может быть оценена с помощью иерархической тематической модели (9):

$$\underbrace{p(w \mid t)}_{\phi^{bank}_{wt}} = \sum_{s \in S} \underbrace{p(w \mid s)}_{\phi^{new}_{ws}} \underbrace{p(s \mid t)}_{\psi_{st}} \tag{24}$$

Если $p(s \mid t) > 1/|S|$, то тема t считается родительской для s. Возможны следующие случаи:

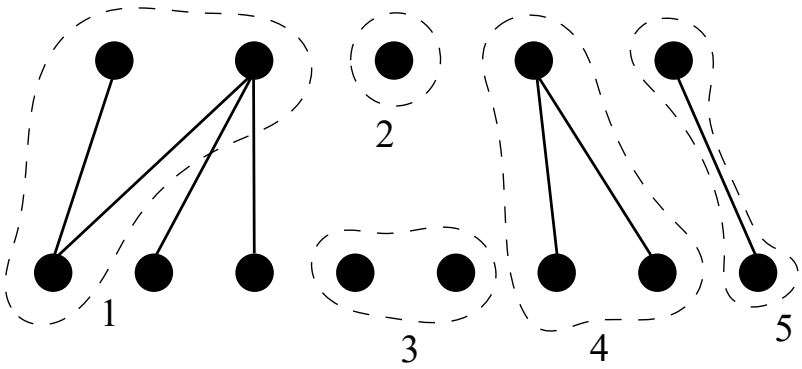


Рисунок 23 — Различия между иерархическим тематическим моделированием и банком тем. Верхний слой точек представляет собой родительские темы (темы банка), нижний слой — дочерние темы (темы новой модели). Ситуация номер 1 показывает случай соединения нескольких тем. Это нормально для иерархии, но не для банка тем: исходные темы остаются в банке, новые темы в него не добавляются. Ситуация номер 2 — когда у родительской темы нет дочерних тем. Это нормально как для иерархии, так и для банка тем. Случай номер 3 — ситуация, когда дочерние темы не имеют родительской темы. Это неприемлемо для иерархии, но допустимо для банка тем: новые темы сразу же добавляются в банк. Ситуация номер 4 показывает расщепление — это когда у одной родительской темы есть несколько дочерних тем, и эти дочерние темы имеют только одну тему в качестве родительской. Это обычная ситуация для иерархии. Что касается банка тем: родительская тема может быть заменена дочерними темами, если все они интерпретируемы. И случай 5 — когда родительская тема переходит на следующий уровень как есть. Это допустимо для иерархии. Также и для банка тем: в банке будет храниться лучшая тема из двух.

– тема s имеет более одного родителя. В этом случае тема s не будет добавлена в банк, так как она является линейной комбинацией некоторых тем, уже имеющихся в банке.
– тема s не имеет родителей. В этом случае она может быть добавлена в банк, если её когерентность высокая.
– тема s имеет единственного родителя. В этом случае родительская тема может быть заменена новой, если когерентность новой темы выше.

Среди тем с наивысшим значением когерентности проходят дальше в процессе отбора только те, которые удовлетворяют условию, связанному с отношением между темами (3.2.4) — имеют менее двух родителей.

Далее отобранные темы сравниваются попарно с темами банка тем с помощью расстояния Жаккара. Только те темы модели, для которых расстояние до ближайшей темы банка не меньше 0.5, добавляются в банк тем (см. рисунок 24). В то же время, если у вновь добавленной темы есть только один родитель в банке тем (в соответствии с одним из вариантов отношений тем (3.2.4), описанных выше), то эта родительская тема удаляется из банка. Таким образом, размер банка тем постепенно увеличивается.

График 25 демонстрирует накопление банком таких тем, у которых высокое значение когерентности. При этом наблюдаем также увеличение когерентности по топ-токенам, которая не была являлась критерием отбора тем в банк.

График 26 показывает пример того, как может эволюционировать банк тем в зависимости от количества обученных моделей. Количество тем в банке увеличивается, сначала резко, затем более плавно, достигая плато в конце. Перплексия банка тем, напротив, уменьшается, также достигая плато. Это служит индикатором того, что процесс сходится и процедура обучения больше не может принести в банк уникальных тем.

Модели На множестве разных тематических моделей хочется проверить, что банк тем может служить инструментом для валидации вновь обученных тематических моделей.

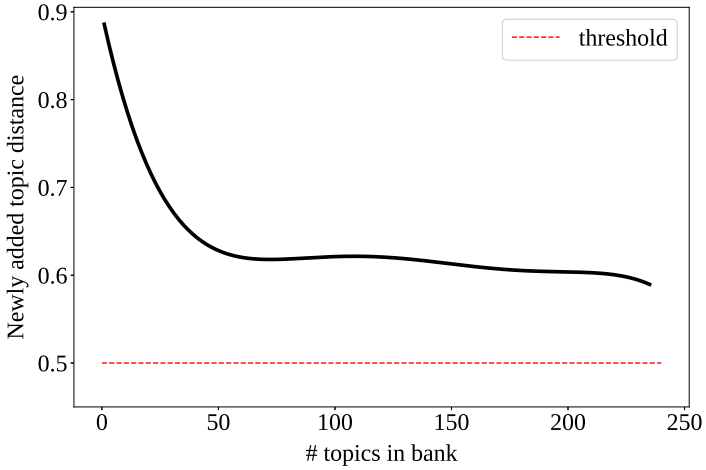
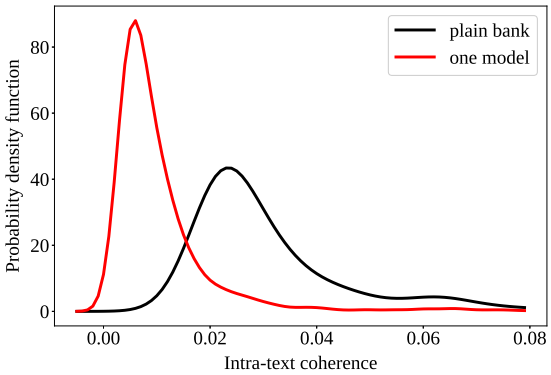
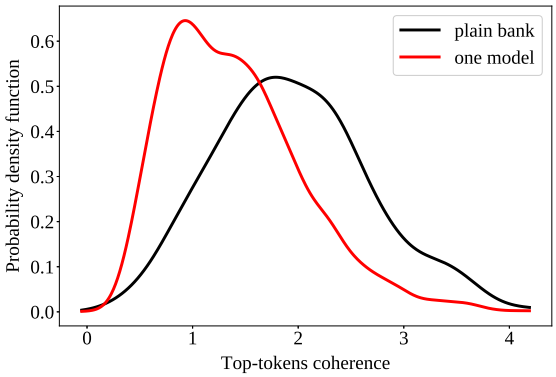


Рисунок 24 — Расстояние между вновь добавленной в банк темой и ближайшей к ней темой, среди уже находящихся в банке. Пунктирная линия обозначает порог расстояния: только темы с расстоянием выше него считаются новыми и могут быть добавлены в банк. Видно, что расстояние до ближайшей темы уменьшается, но при этом не достигает порога. Возможно, это связано с одинаковым процессом обучения для всех моделей: их темы в среднем различаются на некоторую одинаковую величину.

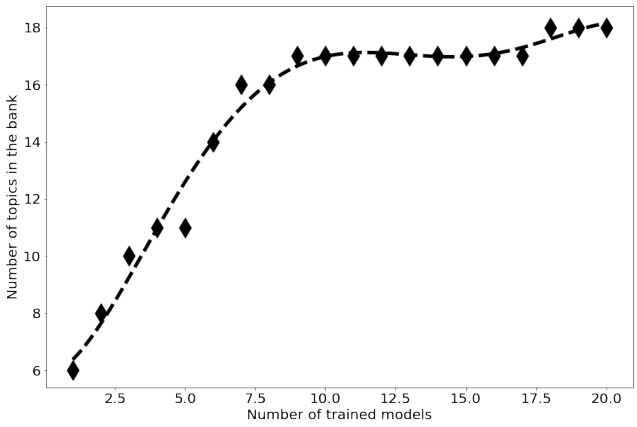


а) KDE оценка плотности значений внутритекстовой когерентности.

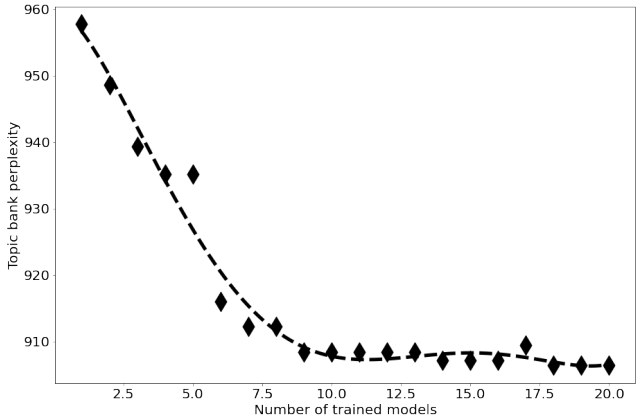


б) KDE оценка плотности значений когерентности по топ-словам.

Рисунок 25 — Два графика демонстрируют KDE оценку плотности, полученную на датасете PostNauka для значений внутритекстовой (a) и по совстречаемостям топ-слов (b) когерентностей тем в банке тем и в одиночных моделях. Видно, что пиковое значение когерентности среди тем банка выше, чем пиковое среди тем одной модели.



а) Количество тем в банке тем в зависимости от количества обученных моделей.



б) Перплексия банка тем как тематической модели в зависимости от количества обученных моделей.

Рисунок 26 — Зависимости некоторых характеристик банка тем от числа обученных моделей (на примере датасета Habrahabr).

Далее опишем тематические модели, используемые в экспериментах. Все модели имеют обучаются при одинаковом постоянном числе тем $T = 100$.

Именем *plsa* обозначаем модель PLSA [138]. Это модель без гиперпараметров, кроме числа тем T. Именем *lda* обозначаем модель LDA [57] с симметричными априорными распределениями для Φ и для Θ.

Имена *arora* и *cdc* представляют модели PLSA со специальным образом проинициализированной матрицей Φ: с помощью алгоритма Arora [144] и с помощью алгоритма CDC [147] соответственно. Для инициализации модели *arora* обучаем ровно T тем Arora. Что касается CDC, гиперпараметры алгоритма подбираются таким образом, чтобы количество тем CDC было близко к T. Если число тем CDC оказывается меньше T, остальные темы модели *cdc* инициализируются случайным образом.

Следующая группа моделей все опираются на использование регуляризаторов [39]. Во-первых, *sparse* использует разреживающий регуляризатор (10), который стремится сделать распределения T предметных тем более разреженными. Во-вторых, *decorr* представляет модель, обученную с помощью регуляризатора декоррелирования (11), который стремится сделать распределения тем более отличными друг от друга. Именем *bcg* обозначена модель со сглаживающим регуляризатором (10), применённым к 2 фоновым темам — то есть модель *bcg* фактически имеет $T+2$ тем, где T тем рассматриваются как предметные темы, а остальные 2 — как фоновые темы. Именем *regul1* обозначена модель с двумя регуляризаторами: разреживающим и декоррелирующим.

Следующая группа моделей, которую хочется упомянуть отдельно — это модели с регуляризатором выбора тем (12). Влияние этого регуляризатора на модель зависит от того, в каком он применяется сочетании с другими регуляризаторами [153]. Поэтому используется несколько моделей, которые отличаются не столько тем, какие регуляризаторы используются, сколько тем, как они применяются вместе. Так, модель *sel_aooa* использует регуляризатор декоррелирования на первой трети итераций обучения, регуляризатор выбора тем на второй трети, а затем разреживающий регуляризатор на оставшихся итерациях. Порядок регуляризаторов в модели *sel_cp* такой же, как и в модели *sel_aooa*. Однако веса регуляризаторов меняются в процессе обучения модели: τ линейно увеличивается с итерацией EM-алгоритма от минимального значения 0 до максимального — используемого для модели *sel_aooa*. Что касается мо-

дели *sel_ao*, то она использует постоянные коэффициенты регуляризации, как и модель *sel_aooa*. Однако регуляризаторы декоррелирования и выбора тем включаются только на чётных итерациях. Таким образом, на нечётных итерациях регуляризаторы не применяются. Наконец, модель *sel_aocp* имеет следующую стратегию регуляризации. Регуляризаторы декоррелирования и выбора тем чередуются: декоррелирующий применяется на нечётных итерациях, а выбора тем — только на чётных. Более того, веса регуляризации τ линейно увеличиваются с ростом числа итераций обучения, как и в случае модели *sel_cp*.

Веса всех регуляризаторов находятся по сетке из нескольких значений путём обучения ряда моделей только с одним регуляризатором для всех исследуемых коэффицинетов регуляризации — и последующего выбора модели с наименьшей перплексией (14). Вес регуляризатора, соответствующий этой модели, фиксируется и используется далее во всех описанных ранее моделях, использующих этот регуляризатор. Для регуляризатора декоррелирования оптимальный вес находился по следующей сетке значений: $[1, 10, 100, 1000, 1e4, 1e5, 1e6]$. Для регуляризатора выбора тем оптимальный коэффициент τ искался среди $[0.01, 0.1, 0.2, 0.5, 0.8, 1.0, 1.5, 2.0]$. А для разреживающего и сглаживающего регуляризаторов использовались сетки поиска $[-10, -1, -0.2, -0.1, -0.02, -0.01, -0.001]$ и $[0.001, 0.01, 0.02, 0.1, 1.0, 10.0]$ соответственно. Более подробную информацию об аддитивной регуляризации можно найти в [5].

Последняя модель, *regul2*, представляет собой модель с тремя регуляризаторами: разреживания, сглаживания и декоррелирования. Коэффициент декорреляции τ для этой модели фиксирован и равен 0.01 — и остаётся неизменным для всех рассматриваемых текстовых коллекций. Коэффициенты разреживания и сглаживания τ также фиксированы. Однако в случае модели *regul2* для разреживания и сглаживания используются *относительные* коэффициенты регуляризации [8]: -0.05 и 0.1 соответственно. Эти относительные τ веса затем преобразуются в абсолютные, которые для модели *regul2* уже будут свои на каждой текстовой коллекции.

Валидация тематических моделей с помощью банка тем Обозначим как B темы в банке тем, а как T — темы текущей тематической модели. Предлагается несколько функций качества тематической модели, основанных на банке тем:

$$\begin{aligned}
\text{recall@bank} &= \frac{\left|t \in B \mid \exists \tau \in T \;:\; \rho(t,\tau) < h\right|}{|B|} \\
\text{coherence@bank} &= \frac{\left|t \in T \mid \exists \tau \in B \;:\; \rho(t,\tau) < h\right|}{|T|} \\
\text{precision@bank} &= \frac{\left|t \in B \mid \exists \tau \in T \;:\; \rho(t,\tau) < h\right|}{|T|}
\end{aligned} \tag{25}$$

Функция recall@bank говорит, какая часть тем банка (хороших тем) были найдены моделью; coherence@bank означает долю тем модели, которые также находится и в банке тем (и потому хорошие); precision@bank есть доля найденных тем банка от числа тем модели.

С тем чтобы вычислить recall@bank, coherence@bank, и precision@bank, надо определиться со способом вычисления расстояния между темами $\rho(t_1, t_2)$. Будем использовать для этого коэффициент Жаккара:

$$\begin{aligned}
\rho(t_1,t_2) = 1 - \sum_{w \in \text{Ker}_{12}} \min_{i \in \{1,2\}} \big(p(w \mid t_i)\big) \Big/ \\
\left(\sum_{i=1}^{2} \sum_{w \in \text{Ker}_i \setminus \text{Ker}_{12}} p(w \mid t_i) \;+\; \sum_{w \in \text{Ker}_{12}} \max_{i \in \{1,2\}} \big(p(w \mid t_i)\big) \right)
\end{aligned} \tag{26}$$

где $\text{Ker}_i \equiv \text{Ker}(t_i)$, $\text{Ker}_{12} \equiv \text{Ker}(t_1) \cap \text{Ker}(t_2)$, и $\text{Ker}(t)$ есть ядро темы t:

$$\text{Ker}(t) = \big\{ w \in t \;:\; p(w \mid t) > 1/|W| \big\}$$

3.2.5 Результаты

Будем называть темы похожими, если расстояние между соответствующими им векторами $\big(p(w \mid t)\big)_{w \in W}$ меньше порога $h \in H \subseteq (0, 1]$,

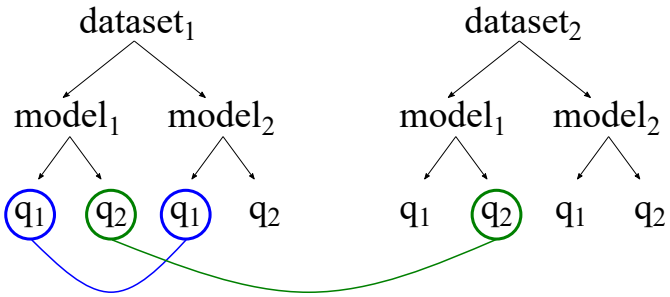


Рисунок 27 — Иллюстрация к уравнению (27): имеется много датасетов, для каждого из которых обучено много моделей, каждая из которых, в свою очередь, имеет много оценок качества q_i. Синим обозначено усреднение оценки качества по всем моделям в рамках одного датасета, зелёный же означает усреднение оценки качества в рамках одной и той же модели, но по разным датасетам.

$|H| < \infty$. Очевидно, что чем ниже порог, тем выше в среднем будет сходство между темами.

Проиллюстрируем подход к оценке качества модели на примере метрики recall@bank (другие метрики следуют аналогичной процедуре). Имеем recall@bank как функцию от набора данных d, модели m и порога h. Усредним метрику по различным значениям порога. Дальнейший поиск лучших моделей будет основан на вычислении средневзвешенного значения. Для этого нужно ввести коэффициент, который даёт преимущества моделям с темами, близкими к темам банка тем. Потому задаём отображение порога расстояния h на вес таким образом, чтобы меньшее значение порога давало б'ольший вес. Например:

$$w(h) = \begin{cases} 1/h^2, & h \le 0.8 \\ 0, & h > 0.8 \end{cases}$$

И получаем recall@bank, усреднённую по порогам H:

$$\langle \text{recall@bank}(m, d, h) \rangle_h = \frac{\sum\limits_{\eta \in H} w(\eta) \cdot \text{recall@bank}(m, d, \eta)}{\sum\limits_{\eta \in H} w(\eta)}$$

В свою очередь, усреднение данной величины ещё раз, но уже по датасетам даёт уже оценку качества одной модели (не зависящую от данных

и от порога):

$$\langle \text{recall@bank}(m, d, h)\rangle_{d,h} = \frac{1}{|\mathscr{D}|}\sum_{\delta\in\mathscr{D}}\frac{\langle \text{recall@bank}(m, \delta, h)\rangle_h}{\sum\limits_{\mu\in\mathscr{M}}\langle \text{recall@bank}(\mu, \delta, h)\rangle_h} \tag{27}$$

где $\mathscr{D}$ есть множество датасетов (см. таблицу 8), и $\mathscr{M}$ есть множество моделей (см. раздел 3.2.4). Рисунок 27 предоставляет иллюстрацию к описанному двухстадийному процессу усреднения.

Сравнение качества разных моделей на всём множестве датасетов представлено на рисунке 28. Видно, что методы *arora* и *cdc* превосходят остальные. Также видно, что определённые ранее метрики качества recall@bank, coherence@bank и precision@bank дают одинаковое упорядочивание моделей по качеству. Для каждого отдельного датасета результаты сравнения изображены на рисунке 29 (только для функции качества recall@bank).

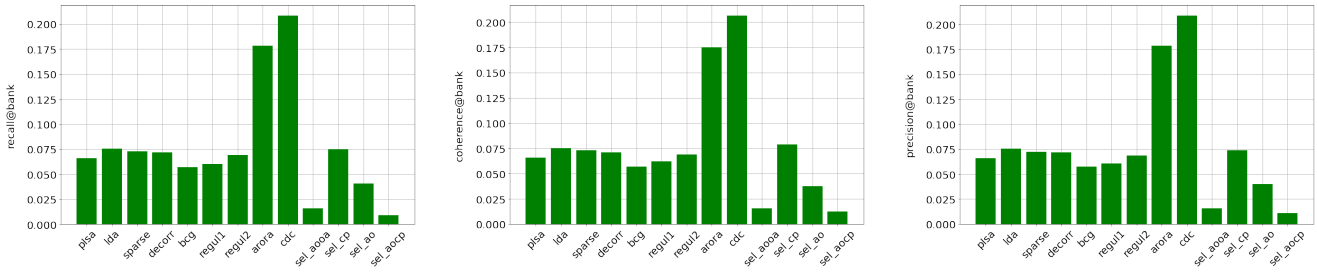


Рисунок 28 — Оценки качества моделей по банку тем, усреднённые по датасетам.

3.2.6 Обсуждение

В этом разделе описан и продемонстрирован новый подход к поиску лучшей тематической модели для заданной текстовой коллекции — подход более прямой и систематический, нежели “полуслучайное блуждание” по пространству гиперпараметров модели. TopicBank помогает исследователю в отборе и хранении хороших (например, по значению когерентности) тем в процессе множественного обучения тематических моделей, что в дальнейшем позволяет обеспечить универсальную оценку ка-

чества вновь обученных моделей — путём сравнения их тем с темами, уже накопленными в банке (отложенными хорошими). Такой подход помогает в итоге выбирать более хорошие тематические модели по сравнению с методами, не использующими знания о ранее встреченных темах.

Экспериментальные результаты показывают, что банк тем позволяет получать модели с более высоким качеством тем, определяемым когерентностью по топ-словам или внутритекстовой. Такими моделями являются, например, все рассмотренные модели с неслучайной инициализацией. Что кажется разумным, поскольку успешная инициализация тематической модели содержит уже хотя бы часть информации об искомой тематической структуре коллекции. Это также хорошо согласуется с идеей о том, что задача тематического моделирования должна упрощаться при наличии дополнительной информации о текстовой коллекции.

В экспериментах для создания банков тем много раз обучались модели одного типа — разной была только случайная инициализация. Однако вообще нет необходимости ограничивать процесс создания банка тем лишь одним типом модели. Кроме этого, разница между моделями, использующимися для извлечения хороших тем, может быть получена не только за счёт изменения типа тематической модели, но и за счёт варьирования некоторых гиперпараметров модели. Например, изменение числа тем в модели может помочь получить темы разной гранулярности (большие, при малом числе тем в модели, или маленькие, при большом числе тем), позволяя восстановить в банке иерархическую структуру тем, порождаемую распределением слов в тексте.

Несмотря на прямой подход к решению проблемы извлечения всех хороших тем коллекции, TopicBank тем не менее не решает её полностью. Основной вклад TopicBank заключается в более осторожном системном подходе к обработке результатов множественного обучения тематических моделей. Но пока нет однозначного ответа, приводит ли вообще процедура по созданию банка тем хотя бы асимптотически к извлечению всех тем из коллекции. Дальнейшее исследование подхода к валидированию моделей с помощью банка тем может быть описано, но не ограничено следующими вопросами:

– Как увеличить качество тем, добавляемых в банк тем?

– Какие ещё характеристики тем, кроме когерентности, стоит принимать во внимание при автоматическом отборе тем? По сути это вопрос о многокритериальной классификации тем на хорошие и плохие.
– Как ускорить процесс сбора тем с высокой когерентностью в процессе множественного обучения моделей? (так чтобы для создания банка требовалось бы как можно меньше моделей)
– Можно ли найти *число* тем в коллекции с помощью банка тем?
– Что будет, если к критериям при создании банка тем, кроме качества отдельных тем и их независимости, добавить такой, чтобы сам банк при этом был хорошей тематической моделью? (чтобы его темы давали модель с низкой перплексией)
– Воспроизведутся ли результаты, если для оценки качества новых тем привлекать людей или LLM? будет ли итоговый банк тем более хорошим (в каком-либо смысле) при таком изменении способа оценки качества тем?

И главным вопросом (который выходит за рамки Банка тем, но к решению которого Банк тем является первым шагом) по-прежнему остается вопрос о том, как построить *сразу одну* тематическую модель, в которой все темы были бы интерпретируемыми и различными, и которая бы хорошо описывала текстовую коллекцию (обладала бы низкой перплексией).

3.2.7 Заключение

Тематическое моделирование — это важный инструмент для исследования данных. Однако существующие алгоритмы обучения тематических моделей по-прежнему приводят к неполным и неустойчивым моделям. Многие работы посвящены преодолению этих недостатков путём изменения способа обучения моделей (создания новых моделей). В данной же работе представлен более простой, понятный и прямой способ исследования данных с помощью тематического моделирования, который фокусируется не на обучении модели, а на её проверке, на её валидации.

Предлагаемый подход состоит из двух шагов. Во-первых, отбор хороших тем с помощью множественного обучения тематических моделей (автоматически без учителя или полуавтоматически с частичным привлечением учителя). Во-вторых, автоматическая валидация новых тематических моделей с использованием отобранных ранее хороших тем. Эксперименты на реальных данных демонстрируют применимость этого подхода.

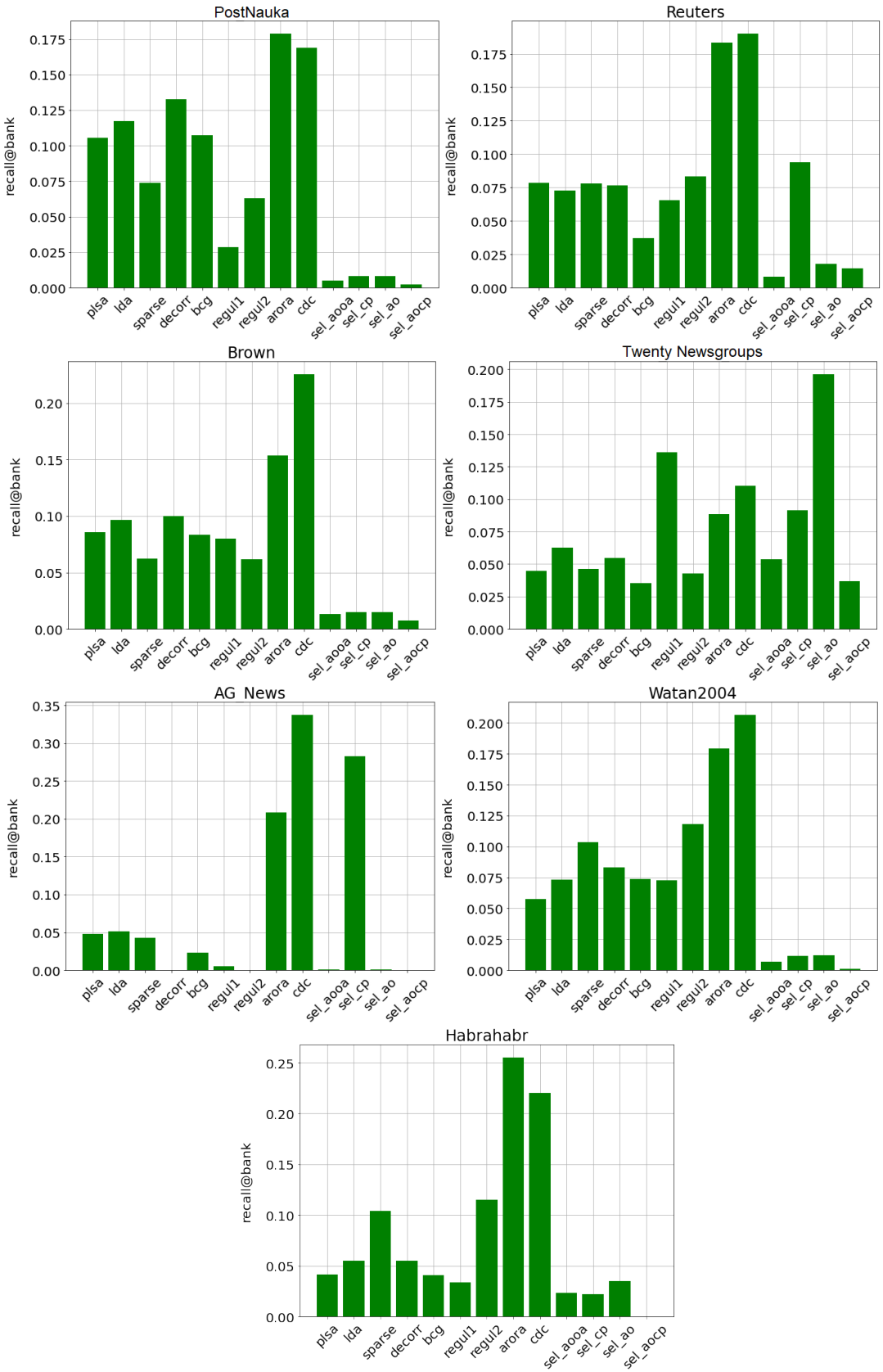


Рисунок 29 — Оценки качества моделей по банку тем для разных текстовых коллекций.

3.3 Регуляризаторы сохранения и отсеивания тем для улучшения тематической модели за несколько проходов

Тематическое моделирование по своей сути является проблемой мягкой кластеризации (известных объектов — документов, над неизвестными кластерами — темами). То есть исходная задача поставлена некорректно. А тематические модели неустойчивы и неполны. Всё это приводит к тому, что процесс поиска хорошей тематической модели (подбор гиперпараметров, обучение модели и оценка качества тем) может быть долгим и трудоёмким. Хочется упростить этот процесс, сделать его более детерминированным и обоснованным. Для этого представляется метод итеративного обучения тематической модели. Суть которого заключается в том, что серия связанных тематических моделей обучается таким образом, чтобы каждая последующая модель была как минимум не хуже предыдущей, то есть сохраняла все хорошие темы, найденные ранее (хорошие в том смысле, что они просто удовлетворяют некоторому критерию, интересующему исследователя, например, являются интерпретируемыми, или когерентными, или релевантными какой-то другой задаче, для которой применяется тематическое моделирование). Связь между моделями достигается с помощью аддитивной регуляризации. Результатом такого итеративного обучения является последняя тематическая модель в серии, которую будем называть итеративно обновляемой аддитивно регуляризованной тематической моделью (iteratively updated additively regularized topic model, ITAR). Эксперименты, проведённые на нескольких коллекциях текстов на естественном языке, показали, что предложенная модель ITAR работает лучше других популярных тематических моделей (LDA, ARTM, BERTopic), её темы разнообразны, а перплексия (мера “удивления” модели при виде данных) умеренная.

3.3.1 Введение

Тематическое моделирование — это направление в области анализа текстов, целью которого является поиск скрытых (латентных) тем в больших объёмах неструктурированных текстовых данных, что даёт возможность автоматически охватывать, понимать большие текстовые коллекции [57]. Предполагается, что каждый текст на естественном языке содержит темы — как интерпретируемые человеком понятия, относящиеся к каким-либо областям реальности. Помимо поиска тем, тематическая модель оценивает вероятности принадлежности документов к каждой из найденных тем. Таким образом, тематические модели также предлагают интерпретируемое представление документов в соответствии с темами [39], которое может быть использовано в других задачах, связанных с обработкой естественного языка, например, при резюмировании текстов и анализе тональности текста. Более того, применимость результатов тематического моделирования распространяется и на другие области, например, социологические исследования [154], мобильные сети [155], здравоохранение [156], дискурс в средствах массовой информации [157], биоинформатику [3], историю [158]. Помимо анализа текстов на естественном языке, тематическое моделирование может применяться и в совершенно других, казалось бы, приложениях, где участвуют совершенно другие сущности, но с теми же отношениями, что есть между словами и документами. Такими сущностями могут быть, например, покупки в историях операций клиентов банка [7] или гены в биологических образцах [159]. Тематическое моделирование — это как уже сложившаяся, так и все ещё активно развивающаяся область анализа текстов [160]. В этой работе мы сосредотачиваемся на вероятностном тематическом моделировании, где темы представляются как вероятностные распределения на множестве слов.

Задача тематического моделирования формулируется как задача матричного разложения — допускающая бесконечно много решений — поэтому задача поставлена некорректно [39]. С целью ограничения количества решений вводятся регуляризаторы. Кроме этого регуляризаторы могут служить для того, чтобы получать тематические модели с желае-

мыми свойствами. Например, регуляризатор разреживания способствует тому, чтобы вероятностное распределение темы было сосредоточено в небольшом числе слов, а не размазано по всему словарю. Регуляризатор декоррелирования способствует получению более отличных друг от друга тем [36]. Однако не все получаемые с помощью тематической модели темы являются хорошими (удовлетворяющими определённом критерию, в общем случае — интерпретируемыми). Кроме хороших, темы бывают плохими (например, когда самые вероятные слова темы на самом деле с точки зрения человека никак не связаны друг с другом, или когда среди самых частых слов темы встречаются “стоп-слова” (служебные) или “фоновые” слова (не несущие никакого смысла, кроме “связки” для языка)). Также можно выделить группу “никаких” тем: таких, которые не являются ни хорошими, ни плохими; темы, которые исследователю было бы не жалко потерять (например, дубликаты уже найденных хороших тем).

Природная некорректность задачи тематического моделирования, вытекающие из неё неполнота и неустойчивость тематических моделей [9; 161] и наличие плохих тем среди находимых моделью приводят к тому, что эксперименты с тематическими моделями могут проходить долго и бессистемно. Выбор гиперпараметров модели, обучение, оценка качества модели, анализ тем модели (автоматический или полуавтоматический). Если модель не устраивает, то всё повторяется заново. При этом найденные хорошие темы *теряются*. И так может быть много раз, пока не будут найдены подходящие гиперпараметры.

Основной идеей данной работы является выстраивание ясного и системного пути от исходной тематической модели до хорошей. Не уходя от необходимости обучения нескольких моделей до получения лучшей, мы предлагаем обучать модели *связанным* образом, одну за другой. Так, чтобы в каждой последующей модели *фиксировались* все хорошие темы, найденные ранее, а плохие — *отсеивались*, то есть чтобы среди оставшихся тем чтобы не было уже найденных ранее плохих тем. Эту последовательность обучаемых друг за другом тематических моделей (а также в зависимости от контекста и самую последнюю, итоговую, модель в последовательности) мы будем называть *итеративной тематической моделью*.

Таким образом, в этом контексте “итерация” — это обучение одной модели.

Для решения задачи по сохранению хороших и отсеиванию плохих тем используются аддитивая регуляризация тематических моделей ARTM. Подход ARTM способствует оптимизации моделей по сумме нескольких критериев [162], что помогает учитывать особенности коллекции текстов и ограничить количество решений задачи тематического моделирования.

Главный вклад работы можно резюмировать следующим образом:

- Представлена итерационно обновляемая аддитивно регуляризованная тематическая модель.
- Введены регуляризаторы по сохранению (фиксации) и удалению (отсеиванию, фильтрации) тем.
- Экспериментами по сравнению предлагаемой модели с рядом других тематических моделей на нескольких коллекциях естественного языка показано, что итеративная тематическая модель способна накапливать наибольший процент хороших тем, которые при этом различны.

3.3.2 Связанные работы

Вероятностное тематическое моделирование

В работе [37] представлена самая простая тематическая модель PLSA, решающая по сути задачу матричного разложения известных частот слов в документах в виде произведения матриц вероятности слов в темах Φ и тем в документах Θ без каких-либо дополнительных ограничений, кроме того, чтоб столбцы матриц вероятностей в самом деле были стохастическими. Авторы [57] предложили модель LDA, ставшую впоследствии очень популярной, где решается всё та же задача матричного разложения, но с дополнительным ограничением на столбцы матриц вероятностей слов в

темах и тем в документах: предполагается, что они порождаются распределением Дирихле.

Аддитивная регуляризация тематических моделей

Работа [39] является отправной точкой в развитии теории аддитивной регуляризации тематических моделей, в русле которой находится и данная работа. В рамках ARTM подхода реализуются и PLSA, и LDA модели. Кроме этого, регуляризаторы предоставляют удобный инструмент для того, чтобы получать тематические модели с желаемыми свойствами, такими как: разреженность тем, различность тем, разделение тем на предметные и фоновые. В работе [135] предложена тематическая модель, обучающаяся без матрицы вероятностей тем в документах, и в которой темы сразу (без дополнительной регуляризации) получаются разреженными.

Нейросетевые тематические модели

В настоящее время большое внимание уделяется возможности использования нейросетей для тематического моделирования [163]: предлагаются нейросетевые тематические модели [113; 164; 165], большие языковые модели также используются для оценки качества получаемых в процессе моделирования тем [114].

В качестве нейросетевой модели, с которой мы собираемся сравнивать вероятностные тематические модели, реализованные в рамках APTM, выбрана BERTopic [113], использующая эмбеддинги слов, полученные из модели BERT [166]. Модель BERTopic состоит по сути из нескольких “блоков”. Первый блок с помощью нейросетевой эмбеддинг модели создаёт эмбеддинги документов. Следующий блок с помощью UMAP [167] понижает размерность этих эмбеддингов перед кластеризацией. Затем

проходит собственно кластеризация документов с помощью алгоритма HDBSCAN [168]. И наконец с помощью TF-IDF [169], когда в качестве “документов” выступают полученные ранее кластеры – группы документов одной темы, BERTopic определяет топ-слова каждой темы.

Внутренние критерии качества тематических моделей

Популярным показателем качества тематической модели является перплексия (14). Часто перплексию даже используют для определения “оптимального” числа тем в текстовой коллекции [119]. Чем меньше перплексия, тем меньше модель испытывает “изумления” при виде данных.

Следующий критерий качества, который мы считаем особенно важным и будем использовать в данной работе, является *различность* тем. Считается, что в хорошей тематической модели темы должны быть различными [5; 116]. В качестве метрики для оценки различности тем в данной работе будем использовать дивергенцию Йенсена – Шеннона [117]:

$$\operatorname{div}(\Phi) = \frac{1}{\binom{T}{2}} \sum_{\substack{\phi_i \neq \phi_j \\ \phi_i, \phi_j \in \Phi}} \sqrt{\frac{1}{2}\Big(\operatorname{KL}(\phi_i \parallel \phi_j) + \operatorname{KL}(\phi_j \parallel \phi_i)\Big)} \tag{28}$$

В работах [100—102] представлена *когерентность* темы (16) — как функция $\operatorname{coh}_{\text{topk}}(t)$ для оценки интерпретируемости темы по совстречаемостям топ-слов.

В работе [6] для оценки качества тем предложен другой подход, названный *внутритекстовой когерентностью*. Авторы выдвигают гипотезу о сегментной структуре текстов [170], которая гласит, что слова тем встречаются в тексте не вперемешку, а группами, сегментами. Таким образом, если тема хорошая, то она согласуется с гипотезой о сегментной структуре и *средняя длина сегмента* слов этой темы будет больше, чем длина сегментов слов плохих тем. Внутритекстовая когерентность $\operatorname{coh}_{\text{intra}}$ как раз оценивает среднюю длину сегмента текста, состоящего из слов темы. Способ такого подсчёта выражается не формулой, а скорее алгоритмом, в резуль-

тате которого вся коллекция просматривается от начала до конца. Это, в частности, делает внутритекстовую когерентность более высислительно затратной по сравнению с когерентностью по топ-словам (для быстрого вычисления которой достаточно один раз заранее пройти по коллекции, чтобы составить матрицу размера $W \times W$ совстречаемостей слов).

Итеративный подход к тематическому моделированию

Итеративное обновление тематической модели — это не что-то новое. Так, само решение задачи тематического моделирования представляет собой итерационный алгоритм. (Хотя в данном контексте “итерация” означает лишь одно обновление матриц Φ и Θ.) Кроме этого, в приложениях тематические модели могут использоваться для анализе коллекций, меняющихся со временем: новостные потоки, научные статьи [171] — где требуется обновление уже обученной модели на вновь поступивших данных.

Данная же работа сконцентрирована на обновлении тематической модели при статичной коллекции. Таким образом, это в некотором смысле “обёртка” над процессом обучения тематической модели. В этой связи данная работа более всего близка [9; 161]. Так, в [9] авторами вводится понятие TopicBank — коллекция хороших различных тем, которые накапливаются в процессе множественного обучения тематических моделей. Авторы предлагают использовать TopicBank как способ оценки качества вновь обученных моделей. Хорошей же тематической моделью он не обязан являться, так как низкая перплексия его как модели не входила в критерий при накоплении тем в банке.

3.3.3 Метод

Итеративная аддитивно регуляризованная тематическая модель Основная идея предлагаемой итеративно обновляемой модели состоит в том, чтобы, имея одну модель, получить следующую, которая бы была гарантированно как минимум не хуже по количеству хороших тем, чем предыдущая (см. рисунок 30). Это достигается использованием двух регуляризаторов: фиксации и отсеивания тем (декорреляции с отложенными хорошими T_+ и плохими T_- темами, столбцы которых собираются в матрицу $\widetilde{\Phi}$):

$$\overbrace{\underbrace{\mathscr{L}(\Phi,\Theta) + R_{\text{sparse}}(\Phi) + R_{\text{decorr}}(\Phi)}_{\text{ARTM}} + R_{\text{fix}}(\Phi,\widetilde{\Phi}) + \underset{\text{bad}}{R_{\text{filter}}}(\Phi,\widetilde{\Phi}) + \underset{\text{good}}{R_{\text{filter}}}(\Phi,\widetilde{\Phi}) \to \max_{\Phi,\Theta}}^{\text{ITAR}} \tag{29}$$

Регуляризатор фиксации тем — действует как регуляризатор сглаживания (10), только теперь вместо равномерного распределения (KL-дивергенцию с которым регуляризатор стремится уменьшить для выбранных тем) выступает то, которое хотим сохранить:

$$R_{\text{fix}}(\Phi,\widetilde{\Phi})|_{\tau\gg 1} = \tau \sum_{t\in T_+}\sum_{w\in W} \widetilde{\phi}_{wt} \ln \phi_{wt} \to \max_{\Phi} \tag{30}$$

Регуляризатор отсеивания тем — действует как регуляризатор декоррелирования (11), только теперь темы обучаемой модели декоррелируются не друг с другом, а с отложенными темами:

$$\underset{\text{bad/good}}{R_{\text{filter}}}(\Phi,\widetilde{\Phi})|_{\tau>0} = -\tau \sum_{t\in T'}\sum_{s\in T_-/T_+}\sum_{w\in W} \phi_{wt}\widetilde{\phi}_{ws} \to \max_{\Phi} \tag{31}$$

где $T' = T \setminus T_+$ — свободные темы в новой модели (которые не фиксируем). Декоррелировать предлагается с отложенными и плохими, и хорошими темами. Идея декорреляции с плохими кажется понятной: не хо-

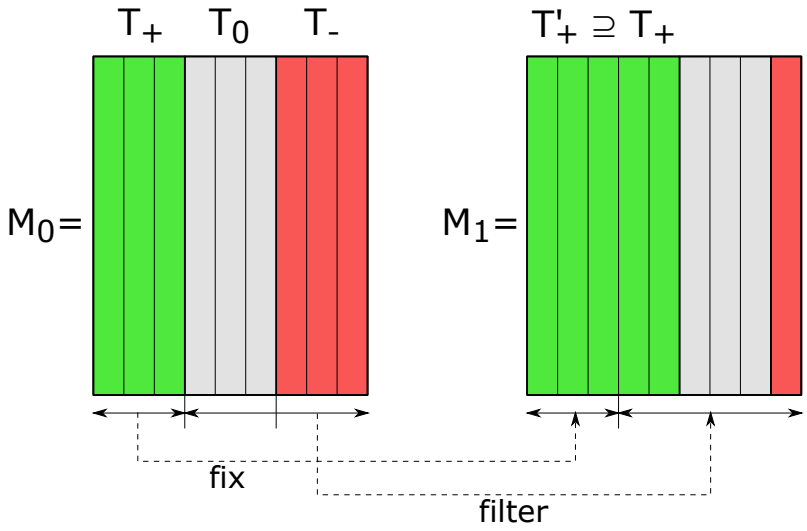


Рисунок 30 — Идея итерационного подхода к улучшению тематической модели. Темы исходной модели M_0 автоматически или полуавтоматически размечаются на хорошие T_+, плохие T_- и “никакие” T_0 (такие, которые не жалко потерять, не плохие, но и не релевантные для исследования, например, это могут быть дубликаты тем из T_+). Далее обучается новая тематическая модель M_1, так чтобы в ней сохранились все темы T_+, а тем T_-, наоборот, чтобы не было. Таким образом модель M_1 по количеству хороших тем T'_+ получается как минимум не хуже модели M_0, а, возможно, и лучше: $T'_+ \supseteq T_+$.

тим, чтобы в новой тематической модели снова появились уже виденные ранее плохие темы. Декорреляция же с хорошими имеет тот смысл, чтоб модель не пыталась повторно найти те же самые хорошие темы (хотим различные хорошие).

Поправка от регуляризатора декоррелирования с отложенными плохими темами на М-шаге:

$$\phi_{wt}\frac{\partial}{\partial\phi_{wt}}R^{\text{filter}}_{\text{bad}} = -\tau[t \in T']\phi_{wt}\sum_{s\in T^-}\widetilde{\phi}_{ws}$$

где выражение $[t \in T']$ означает булевский индикатор, то есть $[t \in T'] = (1 \text{ if } t \in T' \text{ else } 0)$.

Видно, что эффект такого регуляризатора на самом деле — декорреляция со “средней плохой” темой $\sum_{s\in T^-}\widetilde{\phi}_{ws}$! Хотя по первоначальной задумке хотелось, чтобы новая тема обучалась непохожей *ни на одну* из отложенных плохих. Поэтому мы представляем ещё одну версию отсеиваю-

щего темы регуляризатора, призванного исправить этот недостаток:

$$R_{\substack{\text{filter2}\\ \text{bad/good}}}(\Phi, \widetilde{\Phi})|_{\tau>0} = -\frac{\tau}{2} \sum_{t \in T'} \sum_{s \in T_-/T_+} \left(\sum_{w \in W} \phi_{wt} \widetilde{\phi}_{ws} \right)^2 \to \max_{\Phi} \tag{32}$$

$$\phi_{wt} \frac{\partial}{\partial \phi_{wt}} R_{\substack{\text{filter2}\\ \text{bad}}} = -\tau [t \in T'] \phi_{wt} \sum_{s \in T_-} \widetilde{\phi}_{ws} \sum_{u \in W} \phi_{ut} \widetilde{\phi}_{us}$$

Можно заметить, что значения, посчитанные по второй версии регуляризатора (32), получаются на порядки меньше, чем значения регуляризатора первой версии (31) (это значит, что для одинакового по величине эффекта на модель коэффициент τ второго регуляризатора должен быть на порядки больше, чем τ первого).

Итеративную модель с использованием первой версии отсеивающего регуляризатора (31) мы будем обозначать далее как *ITAR*, а модель со второй версией регялизатора (32) — как *ITAR2*.

3.3.4 Эксперимент

Методология В экспериментальной части мы хотим проверить следующие моменты: действительно ли количество хороших тем в предложенной модели увеличивается итеративно? превосходит ли итеративная модель другие тематические модели по итоговому числу хороших тем?

Методология следующая. Имеется несколько тематических моделей. Их предполагается сравнить по количеству хороших тем на нескольких коллекциях текстов. Также в рамках отдельных экспериментов будем обучать модели на $T = 20$ и $T = 50$ тем (считаем, что хорошие тематические модели должны получаться при произвольном числе тем [10]). Для каждой модели проводится несколько (а именно 20) обучений с разными начальными инициализациями. При этом итеративная модель обновляется от итерации к итерации, остальные же модели на каждой итерации обучаются целиком заново. Итоговая итеративная модель — модель на последней

итерации, итоговая не итеративная модель — лучшая по числу хороших тем из всей серии.

Хорошими считаются темы с высокой когерентностью [6; 102]. Порог высокой когерентности находится также экспериментально путём перцентильного анализа всех тем, полученных от неитеративных вероятностных тематических моделей (80% — принят за порог по когерентности, чтобы считать тему хорошей; 20% — порог для того, чтобы считать тему плохой; для каждого датасета и для каждого числа тем в модели абсолютные пороговые значения когерентностей получались разными). В качестве окна совстречаемости при расчёте когерентности (16) в данной работе использовался один документ [100]; количество топ-слов было выбрано равным $k = 20$; также, вместо "обычной PMI" [172; 173], использовалась её положительная версия (positive PMI [174]):

$$\text{PMI}_+(w_i, w_j) = \max\left(\ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, 0\right)$$

Эксперимент был произведён на языке Python с использованием библиотек с открытым исходным кодом TopicNet[6] [8] и BigARTM[7] [39].

Исходный код представленной итеративной тематической модели, а также исходный код и результаты проведённых экспериментов также находятся в открытом доступе.[8]

Данные Используется несколько текстовых коллекций: часть на русском, часть на английском языке (см. таблицу 9). Все датасеты уже были заранее предобработаны (выделение модальностей, лемматизация, удаление стоп-слов). 20Newsgroups[9] — очень известный датасет по тематическому моделированию. Коллекция PostNauka была впервые использована в [6; 175], RuWiki-Good и ICD-10[10] датасеты собраны и представлены авторами [8]. RTL-Wiki-Person был впервые использован в [71; 103].

[6] https://github.com/machine-intelligence-laboratory/TopicNet.
[7] https://github.com/bigartm/bigartm.
[8] https://github.com/machine-intelligence-laboratory/OptimalNumberOfTopics.
[9] http://qwone.com/~jason/20Newsgroups.
[10] https://en.wikipedia.org/wiki/ICD-10.

Единственная предобработка, которая проводилась с датасетами — фильтрация словаря. Так, данные, как правило, были мультимодальные (например, обычный текст, биграммы, заголовок), но в экспериментах использовалась лишь одна основная модальность (обычный текст). Кроме фильтрации по модальности, применялась фильтрация токенов по частоте: $\text{df}_{\min} = 5$, $\text{df}_{\max} = 0.5$ для RuWiki-Good; $\text{df}_{\min} = 2$, $\text{df}_{\max} = 0.5$ для RTL-Wiki-Person и для ICD-10 (где df — частота появления токена в документах коллекции, абсолютная (сколько документов) или относительная (доля документов)).

Все использованные датасеты находятся в открытом доступе.[11]

Таблица 9 — Датасеты, которые использовались в экспериментах (D — количество документов, Len — средняя длина документа, W — размер словаря (после отсеивания очень редких и очень частых слов), Lang — язык, BOW — индикатор, означающий наличие текста в Bag-of-Words формате, NWO — индикатор, означающий наличие текста с естественным порядком слов). Из 20Newsgroups датасета использовалась только обучающая подвыборка документов.

Dataset	**D**	**Len**	**W**	**Lang**	**BOW**	**NWO**
PostNauka	3404	421	19186	Ru	✓	✓
20Newsgroups	11301	93	52744	En	✓	✓
RuWiki-Good	8603	1934	61688	Ru	✓	
RTL-Wiki-Person	1201	1600	37739	En	✓	
ICD-10	2036	550	22608	Ru		✓

Модели Для сравнения с итеративной тематической моделью (29) используются следующие тематические модели. *PLSA*, модель с единственным гиперпараметром T [37]; *LDA*, у которой столбцы Φ и Θ порождаются распределениями Дирихле [57]; *Sparse* — модель с аддитивной регуляризацией [39], состоящей в разреживании и сглаживании тем (10) (сглаживание применяется лишь к одной дополнительной теме — фоновой); *Decorr* — модель с аддитивной регуляризацией, состоящей в декоррелировании (11) и сглаживании (10) тем (где сглаживание снова применяется только к одной дополнительной, фоновой, теме); *TLESS* — модель без

[11] https://huggingface.co/TopicNet.

матрицы Θ, с темами, которые по умолчанию получаются разреженными [135]; *BERTopic* — нейросетевая тематическая модель [113]; *TopicBank* — итеративно обновляемая модель, но без регуляризаторов [9].

Все модели, кроме BERTopic, реализуются в рамках TopicNet/ARTM фреймворка.

Базовая модель, которая обучается с разными начальными приближениями на итерациях при обучении ITAR модели — это ARTM модель с разреживающим, сглаживающим и декоррелирующим регуляризаторами, и ещё с дополнительными регуляризаторами фиксирования (30) и отсеивания тем (29) (для модели ITAR отсеивающий регуляризатор определяется соотношением (31), для модели ITAR2 — соотношением (32)).

Все модели, кроме PLSA, обладают одним или несколькими настраиваемыми гиперпараметрами. Гиперпараметры моделей с регуляризацией (Sparse, Decorr, ITAR, ITAR2) — это по сути лишь коэффициенты регуляризации τ. Коэффициенты регуляризации для моделей Sparse и Decorr использовались относительные и подбирались по сетке: $\{-0.05, -0.1\}$ для разреживания, $\{0.05, 0.1\}$ для сглаживания, и $\{0.01, 0.02, 0.05, 0.1\}$ для декорреляции (выбирался коэффициент, приводящий к модели с минимальной перплексией). Коэффициент регуляризации фиксирующего регуляризатора моделей ITAR и ITAR2 был положен равным абсолютному значению 10^9 для всех датасетов, кроме RuWiki-Good (для которого коэффициент был выставлен равным 10^{12}). Для отсеивающего регуляризатора коэффициенты регуляризации побирались также абсолютные, по сетке: $\{10, 100, \ldots, 10^{10}\}$ для регуляризатора модели ITAR (31), и $\{10, 100, \ldots, 10^{12}\}$ для регуляризатора модели ITAR2 (32) (выбирался минимальный коэффициент, который приводил к заметному, но не очень большому (порядка 10%), ухудшению перплексии). В результате были зафисированы следующие коэффициенты регуляризации: $\tau = -0.05$ для разреживания; $\tau = 0.05$ для сглаживания; $\tau = 0.01$ для декоррелиции; $\tau = 10^5$ для отсеивания тем в модели ITAR для датасетов PostNauka, 20Newsgroups, ICD-10, и $\tau = 10^6$ для датасетов RuWiki-Good и RTL-Wiki-Person; $\tau = 10^8$ для отсеивания тем в модели ITAR2 для датасетов PostNauka, 20Newsgroups, ICD-10, $\tau = 10^{10}$ для датасета RuWiki-Good, и $\tau = 10^9$ для RTL-Wiki-Person. Для модели LDA использовались симметричные априорные распределения (сравнение их

с асимметричными [65] не показало заметной разницы; “эвристические” же априорные распределения [176] привели даже к несколько большему значению перплексии).

Модель BERTopic отличается тем, что в ней нет возможности явно задать нужное число тем. Но число тем подбиралось близким к желаемому (20 или 50) с помощью подмодуля по одноуровневой “плоской” кластеризации библиотеки HDBSCAN.[12] Таким образом, при инициализации модели BERTopic подбирался и задавался только параметр ϵ (который отвечал за итоговое число тем в модели).[13] Остальные значения гиперпараметров принимали значение по умолчанию.

Модель TopicBank используется в двух версиях: *TopicBank* — как в оригинальной статье [9], и *TopicBank2* — когда при создании банка тем используется множественное обучение моделей ARTM (разреживание, сглаживание и декорреляция), а не PLSA. Помимо изменения базовой модели, в *TopicBank2* изменён порог когерентности по отбору тем в банк: вместо 90% перцентили по значению когерентностей тем вновь обученной модели [9] использовался такой же порог, как при отборе хороших тем для ITAR модели.

На каждой итерации обучения при инициализации моделей выставлялся новый `seed`, равный номеру итерации (этот параметр определяет инициализацию Φ матрицы для APTM моделей, и поведение UMAP для модели BERTopic).

3.3.5 Результаты и обсуждение

Когерентность по топ-словам Эксперименты показали, что итеративная модель содержит больше всего хороших тем. При этом хорошими могут быть больше 80% тем итеративной модели (см. рисунок 31). При этом её темы различны, а перплексия всей модели, хоть и не самая маленькая среди рассмотренных моделей, но умеренная (см. подробные результаты по нескольким датасетам в таблице 10, и только процент хороших тем в

[12] https://github.com/scikit-learn-contrib/hdbscan/pull/398.

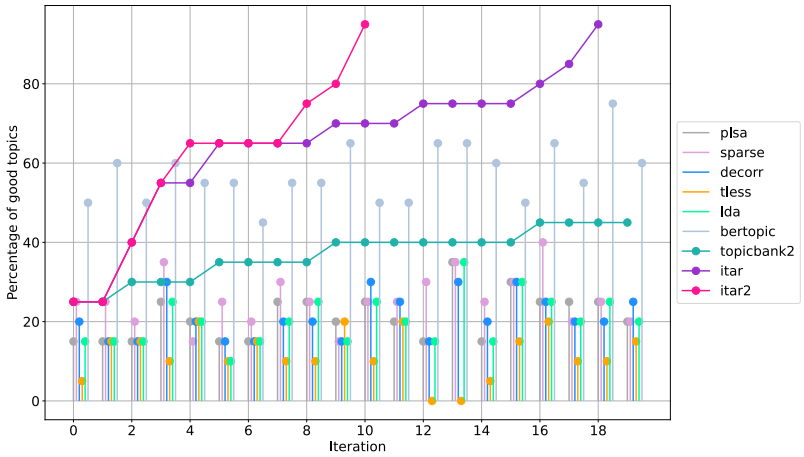
[13] `cluster_selection_epsilon`.

модели для всех датасетов — в таблице 11). То, что перплексия итеративной модели выше минимальной, объяснимо, так как это модель, обученная с дополнительными ограничениями (с регуляризацией). Минимальной же перплексией обладают самые простые модели PLSA и LDA.

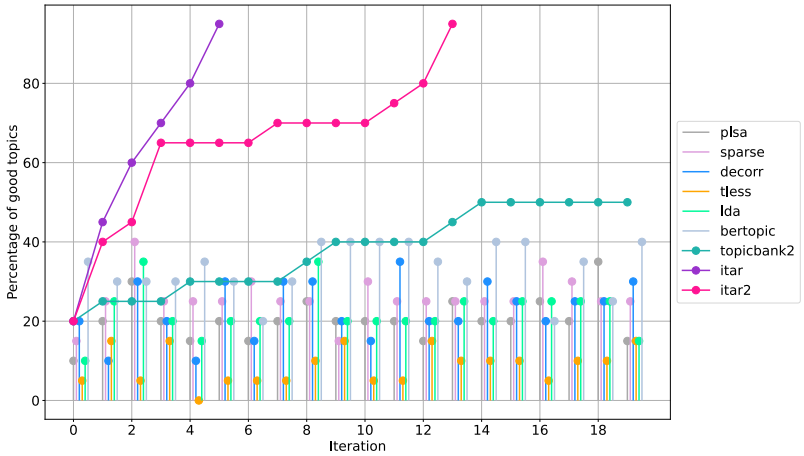
Относительно регуляризации было также замечено следующее. Так как фиксация тем в итеративной модели всё больше ограничивает её свободу (по мере того, как хороших тем становится всё больше), может наступить момент, когда хороших тем становится так много, что оставшиеся свободные темы "вырождаются", становясь нулевыми. Чтобы этого избежать, в обучение итеративной модели был заложен критерий останова по числу накопленных хороших тем: при обучении модели на $T = 20$ считалось, что накопление уже хотя бы $20 - 2 = 18$ хороших тем достаточно; при обучении модели на $T = 50$ тем достаточным для останова числом хороших считалось $50 - 5 = 45$ тем (таким образом, по 90% тем в обоих случаях). Но в итеративной модели могло получиться и больше 90% хороших тем, ведь на итерации их может быть добавлено несколько.

Абляционное исследование (Ablation Study) Итеративное обновление модели обеспечивается применением нескольких регуляризаторов: фиксации хороших тем (30), декорреляции с отложенными плохими (31) и отложенными хорошими темами. Но каков вклад каждого из регуляризаторов в итоговое качество модели? Все ли перечисленные регуляризаторы одинаково важны?

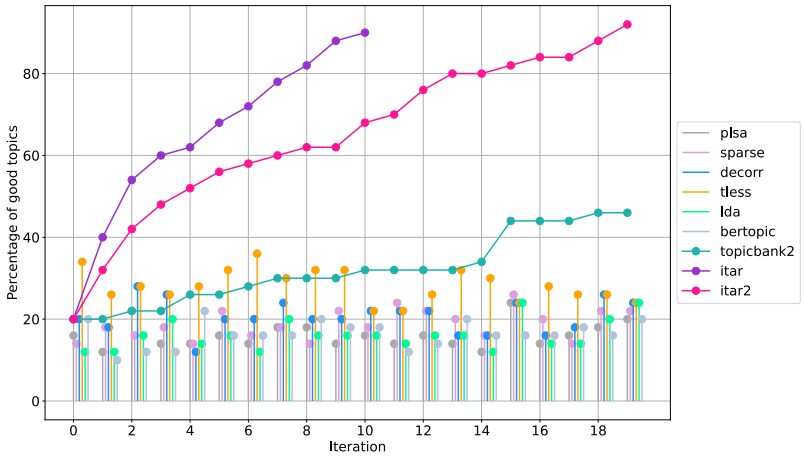
Чтобы это выяснить, был проведён отдельный эксперимент, заключавшийся в том, что кроме "полной" ITAR модели со всеми регуляризаторами было обучено ещё несколько итеративных моделей, в каждой из которых был "выключен" один или два регуляризатора. Полученные результаты для датасета PostNauka при $T = 20$ приведены в таблице 12. Откуда виден вклад каждого из регуляризаторов: фиксация хороших тем увеличивает итоговый процент хороших тем, а также перплексию; декорреляция с плохими снижает частоту появления плохих тем в моделях, обучаемых на отдельных итерациях; декорреляция с хорошими приводит в более различным темам.



а) RuWiki-Good, $T = 20$.



б) RTL-Wiki-Person, $T = 20$.



в) 20Newsgroups, $T = 50$.

Рисунок 31 — Процент хороших тем в модели в зависимости от итерации (↑). В итерационных моделях (TopicBank2, ITAR, ITAR2) каждая последующая модель обучается на основе предыдущей, отсюда и монотонная зависимость (в отличие от неитеративных моделей).

Внутритекстовая когерентность Подсчёт внутритекстовой когерентности предполагается проводить на тексте с естественным порядком

Таблица 10 — Некоторые свойства итоговых моделей: перплексия (ppl, ↓), средняя когерентность тем (coh, ↑), процент интерпретируемых тем (T_+, ↑), различность тем (div, ↑). PostNauka, модели на 20 тем (слева); RuWiki-Good, модели на 50 тем (справа). Видно, что итеративные модели ITAR и ITAR2 имеют наибольший процент хороших тем (T_+). При этом темы различны (div), а перплексия всей модели умеренная (ppl). Для моделей BERTopic, TopicBank и TopicBank2 в столбце для перплексии приведено два значения в формате ppl_1 / ppl_2: второе ppl_2 — это "честная" перплексия, посчитанная как для тематической модели с ровно такими темами, как в обученных BERTopic, TopicBank и TopicBank2 соответственно. Первое же значение перплексии ppl_1 посчитано при добавлении к темам упомянутых моделей ещё одной фоновой темы (заведомо плохой, но которая не берётся в рассчёт при вычислении других показателей качества модели). Это даёт моделям BERTopic, TopicBank и TopicBank2 больше свободы, позволяя достигать более низкой перплексии. Таким образом, сравнение по перплексии с другими моделями получается более "интересным" (конкурентным).

Model	**PostNauka (20 topics)**				**RuWiki-Good (50 topics)**			
	ppl/1000	coh	T_+, %	div	ppl/1000	coh	T_+, %	div
plsa	**2.99**	0.74	20	0.60	**3.46**	0.81	26	0.66
sparse	3.33	0.84	40	0.66	3.85	0.85	28	0.68
decorr	3.15	0.79	40	0.61	3.62	0.86	30	0.67
tless	3.65	0.75	30	**0.75**	4.98	0.71	24	0.72
lda	**2.99**	0.73	25	0.58	**3.48**	0.83	24	0.65
bertopic	4.26/5.93	**1.16**	75	0.67	3.17/5.06	**1.34**	70	0.67
topicbank	4.22/6.11	0.98	30	0.60	7.39/12.94	**1.33**	20	0.68
topicbank2	4.12/8.11	**1.10**	70	0.67	6.09/11.30	1.16	44	0.69
itar	3.79	1.02	**90**	**0.76**	4.62	1.12	**86**	**0.77**
itar2	3.75	1.00	**90**	0.74	4.53	1.23	**96**	**0.77**

слов [6], поэтому среди всех датасетов (см. таблицу 9) использовались только те, где был доступен лемматизованный текст в естественном порядке.

При оценке хорошести тем по значениям их внутритекстовой когерентности итерационная модель также монотонно улучшается, однако по итогу оказывается сравнима с лучшими не итерационными моделями

Таблица 11 — Процент хороших тем (T_+, ↑), соответствующих итоговым моделям на каждом из исследуемых датасетов. Для краткости введены следующие новые обозначения: PN (PostNauka), 20NG (20Newsgroups), RWG (RuWiki-Good), RTL (RTL-Wiki-Person) и ICD (ICD-10).

Model	**20 topics**					**50 topics**				
	PN	20NG	RWG	RTL	ICD	PN	20NG	RWG	RTL	ICD
plsa	20	25	35	35	25	20	24	26	24	20
sparse	40	30	40	40	35	34	26	28	28	26
decorr	40	30	30	35	30	32	28	30	30	26
tless	30	35	20	15	35	34	36	24	24	50
lda	25	25	35	35	25	26	24	24	26	20
bertopic	75	55	75	40	**90**	68	22	70	52	88
topicbank	30	20	15	20	30	16	22	20	16	22
topicbank2	70	45	45	50	75	48	46	44	42	28
itar	**90**	**95**	**95**	**95**	**90**	**92**	**90**	**86**	**80**	**92**
itar2	**90**	**95**	**95**	**95**	**100**	**88**	**92**	**96**	**88**	**90**

(см. рисунок 32). Дело в том, что значение внутритекстовой когерентности темы, в отличие от её когерентности по топ-словам, *зависит от других тем*. Таким образом, фиксация хороших тем по внутритекстовой когерентности (таких, которые встречаются в тексте в виде длинных однородных сегментов) ещё сильнее ограничивает свободу модели, чем фиксация тем, отобранных по совстречаемостям топ-слов (так как размер коллекции фиксирован, то просто не может быть слишком много тем с большим количеством длинных сегментов, где каждый сегмент относится лишь к одной теме).

Было замечено, что усиливающаяся регуляризация (когда фиксируется всё больше и больше тем) может приводить к тому, что внутритекстовая когерентность по некоторым темам оказывается нулевой (это значит, что ни для одного слова в тексте максимум вероятностей $p(t \mid w)$ не приходится на эти темы [6]). Поэтому, помимо критерия останова по числу хороших тем, как в эксперименте с когерентностью Ньюмана (см. раздел 3.3.5), применялся следующий: обучение итеративной модели останавливалось, если у хотя бы одной темы внутритекстовая когерентность становилась равной нулю.

Таблица 12 — Влияние разных частей ITAR модели на итоговый результат на примере датасета PostNauka при обучении моделей на 20 тем Формат имени: “itar_[есть ли фиксация хороших]-[есть ли декорреляция с плохими]-[есть ли декорреляция с хорошими]”. В столбце “# iters” показано, сколько итераций заняло обучение (в процентах от максимального числа в 20 итераций). В столбце T_- в процентах от числа тем в одной модели (20) представлено суммарное число плохих тем, найденных итерационной моделью за все итерации обучения, от первой до последней (в отличие от столбца T_+, где показан процент хороших тем в модели на последней финальной итерации). Столбцы ppl, coh, T_+, div имеют такой же смысл, как в таблице 10. По T_+ видно, что фиксация хороших тем ожидаемо увеличивает долю хороших тем в модели; T_- показывает, что декорреляция с плохими снижает частоту появления плохих тем; а div демонстрирует, что декорреляция ещё и с хорошими приводит в более различным темам.

Model	**PostNauka (20 topics)**					
	# iters, %	ppl/1000	coh	T_+, %	T_-, %	div
itar	**50**	3.79	**1.02**	**90**	**100**	**0.76**
itar_0-0-1	85	**3.30**	0.81	35	275	0.66
itar_0-1-0	**60**	**3.31**	0.86	50	350	0.71
itar_0-1-1	85	**3.31**	0.93	50	325	0.71
itar_1-0-0	70	3.56	0.90	60	230	0.69
itar_1-0-1	90	3.65	0.95	75	200	0.72
itar_1-1-0	90	3.75	**1.05**	**95**	**95**	**0.75**

В экспериментах с разным числом тем в моделях (20 и 50) относительные пороги когерентности, по которым определялись хорошие и плохие темы, оставались неизменными (см. секцию 3.3.4). Однако абсолютные пороги были разные. При этом для всех в случае когерентности Ньюмана для всех датасетов (3.3.4) наблюдалось увеличение абсолютных порогов при изменении числа тем в моделях с 20 до 50, в случае же внутритекстовой когеретности — уменьшение. Последнее объясняется тем, что темы в моделях предполагаются равносильными [40], и увеличение числа

тем в модели приводит к уменьшению размера тем,[14] отсюда и уменьшение средней длины сегмента темы. Повышение же Ньюмановской когерентности можно объяснить так. Уменьшение тем — отчасти объясняется расщеплением больших тем на более маленькие [9]. И когерентность Ньюмана повысится, если одна большая неоднородная тема (у которой совстречаемости топ-слов не очень высокие) расщепится не несколько более маленьких, но уже однородных (таких, у которых топ-слова распределены в тексте более компактно, близко друг к другу).

Интересно было проверить наличие связи между когерентностями: по совстречаемостям и внутритекстовой. В работе [9] было показано, что эти когерентности связаны: в результате отбора тем по значениям одной когерентности, в итоге среди отложенных тем росло среднее значение другой, то есть значения двух когерентностей коррелируют. При обучении же итеративной модели тоже происходит явный отбор тем по значениям выбранной когерентности (например, внутритекстовой). Но тогда можно посмотреть, сколько среди таким тем в среднем содержится тем, когерентных по другой метрике когерентности (по совстречаемостям). Если связь между когерентностями существует, то её можно будет оценить по такой "концентрации" одних тем среди других. Представляются две величины для оценки наличия связи между когерентностями. *Относительная плотность тем, хороших по когерентности по совстречаемостям топ-слов* coh_{topk}*, среди тем модели*. Например, если в модели на 20 тем есть 2 хороших темы, то плотность хороших тем будет $2/20 = 10\%$. Если известно, что в среднем в обучаемых моделях плотность хороших тем составляет 20%, то относительная плотность хороших тем для модели из примера составит $10/20 = 0.5$. Таким образом, если относительная плотность хороших тем выше единицы, то это означает, что в данной модели хороших тем содержится в среднем больше, чем во всех исследуемых тематических моделях. Следующая величина — *относительная плотность хороших относительно когерентности по совстречаемостям топ-слов тем среди тем, хороших по внутритекстовой когерентности*. Например, если в модели на 20 тем есть 2 хороших по внутритекстовой когерентности темы,

[14] Размер темы, или объём темы (topic capacity) — это количество текста, занятого темой [40]: $n_t \equiv \equiv \sum_{d \in D} n_d \theta_{td}$. Существует также несколько иной способ определения размера темы, который непосредственно следует из обозначений, введенных в разделе 1.1: $n_t = \sum_{d \in D} n_{td} = \sum_{d \in D} \sum_{w \in d} n_{dw} p_{tdw}$.

из которых лишь 1 является также хорошей по когерентности Ньюмана, то значение $\tau_+^{\text{topk@intra}}$ для этой модели составит (1/2)/0.2 = 2.5. Аналогично, чем данная плотность выше единицы, тем больше среди внутритекстово когерентных тем содержится тем, когерентных также по совстречаемостям топ-слов. Исходя из постановки эксперимента (см. раздел 3.3.4), средний процент тем в модели, хороших по оценке coh_{topk}, составляет 20%, а среднее значение τ_+^{topk} равно 1.0. В итоге показано, что итерационные модели, обученные с тем, чтобы накапливать интерпретируемые по внутритекстовой когерентности темы, содержат в итоге большой процент и когерентных по Ньюману тем (см. таблицу 13). Более того, таких тем не просто много в целом во всей модели, но высокий процент когерентных по Ньюману тем содержится и среди самих внутритекстово когеретных тем (см. таблицу 14). Это указывает на положительную связь между двумя мерами согласованности тем: по совстречаемостям топ-слов и внутритекстовой — различными подходами к оценке интерпретируемости тем.

3.3.6 Заключение

В работе представлена итеративно обновляемая тематическая модель как серия обучаемых последовательно друг за другом связанных тематических моделей. Процесс устроен так, чтобы в итерационная модель накапливала уже найденные и “искала” новые хорошие темы. Итеративное обновление модели (переход от только что обученной к вновь обучаемой) реализовано в рамках АРТМ подхода: за накопление тем отвечает регуляризатор фиксирования тем (регуляризатор типа сглаживания), а за поиск новых хороших — регуляризатор декоррелирования с отложенными темами (регуляризатор типа декоррелирования). Таким образом, связь между моделями осуществляется с помощью регуляризации, отсюда и название: итеративная аддитивно регуляризованная тематическая модель (iterative additively regularized topic model, ITAR).

На нескольких коллекциях текстов естественного языка проведены эксперименты по сравнению модели ITAR с другими тематическими мо-

Таблица 13 — Связь между когерентностью по совстречаемостям топ-слов и внутритекстовой когерентностью. Для итерационных моделей (TopicBank2, ITAR, ITAR2) для первой и последней итераций обучения представлены следующие показатели: τ_+^{topk} (↑) означает относительную плотность тем, хороших по когерентности по совстречаемостям топ-слов coh_{topk}, среди тем модели. Далее, $\tau_+^{\text{topk@intra}}$ (↑) означает относительную плотность хороших относительно когерентности по совстречаемостям топ-слов тем среди тем, хороших по внутритекстовой когерентности. По результатам эксперимента, среднее значение τ_+^{topk} для тем моделей равно 1.0 ± 0.3 (точная погрешность зависит от датасета и количества тем в моделях, то есть она уникальна для каждого эксперимента; однако по результатам наблюдений вышло, что 0.3 в среднем может служить хорошей оценкой погрешности) — то есть совпадает с теоретическим значением в пределах погрешности. TopicBank2 имеет низкую начальную плотность τ_+^{topk} в основном потому, что после первой итерации в банке тем обычно находится ещё очень небольшое число тем (TopicBank как тематическая модель накапливает темы, от итерации к итерации, постепенно увеличивая их количество).

Model	First iteration		Last iteration	
	τ_+^{topk}	$\tau_+^{\text{topk@intra}}$	τ_+^{topk}	$\tau_+^{\text{topk@intra}}$
PostNauka (20 topics)				
topicbank2	0.5		0.5	**1.7**
itar	1.5	**3.3**	**2.8**	**2.1**
itar2			**2.5**	**2.1**
20Newsgroups (50 topics)				
topicbank2	0.4		0.6	**1.7**
itar	1	**2.2**	**1.7**	**2.1**
itar2			**1.4**	**1.7**

делями. Показано, что представляемая итеративная модель превосходит все остальные по количеству хороших тем — где хорошесть темы определяется по значению её когерентности на основе совстречаемости топ-слов (PMI) — при этом её темы различны, а перплексия умеренная.

Таблица 14 — Связь между когерентностью по совстречаемостям топ-слов и внутритекстовой когерентностью, оценённая по общей совокупности тем всех моделей. Для датасетов с естественным порядком слов и количеством тем T в одной модели, $\tau_+^{\text{topk@intra}}$ означает относительную плотность когерентных по топ-словам тем среди внутритекстово когерентных тем. Однако в данной таблице значения этой плотности приведены не *локальные*, по темам отдельных моделей (как в таблице 13), а *глобальные*: оценённые по всем накопленным темам всех моделей, которые использовались при вычислении абсолютных порогов когерентностей для определения принадлежности темы к числу хороших или плохих (см. раздел 3.3.4). Ожидаемое значение глобальной плотности τ_+^{topk} когерентных по топ-словам тем среди всех тем также составляет 1.0. Посчитанные же экспериментально значения τ_+^{topk} составляют: 1.00 ± 0.07 для $T = 20$, и 1.00 ± 0.05 для $T = 50$. (Погрешности оценены путём многократного сэмплирования из общей совокупности тем, полученных от всех моделей, нескольких в количестве T_+^{intra} и оценке плотности на этой совокупности, где T_+^{intra} есть общее число тем, когерентных по $\mathrm{coh}_{\text{intra}}$.) Таким образом, значения $\tau_+^{\text{topk@intra}}$ выше единицы свидетельствуют о том, что среди внутритекстово когерентных тем в среднем содержится большое число тем, когерентных по Ньюману.

Dataset	20 topics $\tau_+^{\text{topk@intra}}$	50 topics $\tau_+^{\text{topk@intra}}$
PostNauka	2.11	1.73
20Newsgroups	1.18	1.71
ICD-10	1.53	1.85

3.3.7 Ограничения

Стоит отметь несколько ограничений, допущенных в процессе исследования и/или присущих полученному результату.

Применение итерационного подхода к улучшению тематической модели предполагает использование APTM фреймворка. Так, вопрос о возможности итеративно улучшить модель BERTopic не исследовался.

Тем не менее, можем отметить, что BERTopic всё-таки имеет некоторые возможности для (полу)контролируемого тематического моделирования. Наиболее похожей на предложенный в работе регуляризатор фиксации тем для модели ITAR является “направленное тематическое моделирование” (“guided topic modeling”). Это техника, которая позволяет исследователю “направлять” модель BERTopic в нужном направлении (отсюда и название) с помощью подачи на вход модели набора так называемых “начальных тем” (“seed topics”), причём тем именно как последовательностей слов, а не как вероятностных распределений на словах. Однако эти начальные темы в итоговой модели не (строго) сохраняются.[15]

Не по всем критериям возможно накопление тем. Итеративная модель эффективна, если критерий хорошести темы не зависит от других тем.

Получение итерационной модели требует множественного обучения моделей. Если датасет большой, то это может быть не эффективно. (Даже если датасет не большой, это всё равно не очень удобно.)

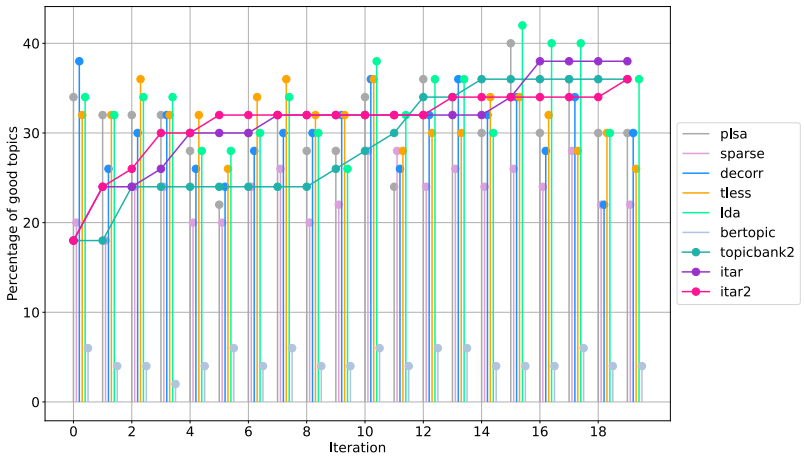
В качестве оценок хорошести тем использовались только автоматически вычисляемые меры когерентности. Однако человеческая оценка интерпретируемости тем является более надёжной. (Но в то же время она дороже и её труднее получить).

Не выявлено видимой разницы между ITAR и ITAR2 моделями. Во всех рассмотренных случаях обе показывали похожие результаты (таким образом, к использованию можно рекомендовать более простую модель ITAR). Однако кажется, что при большом числе отложенных плохих тем регуляризатор (31), участвующий в обучении ITAR модели, не сможет эффективно отсеивать плохие темы.

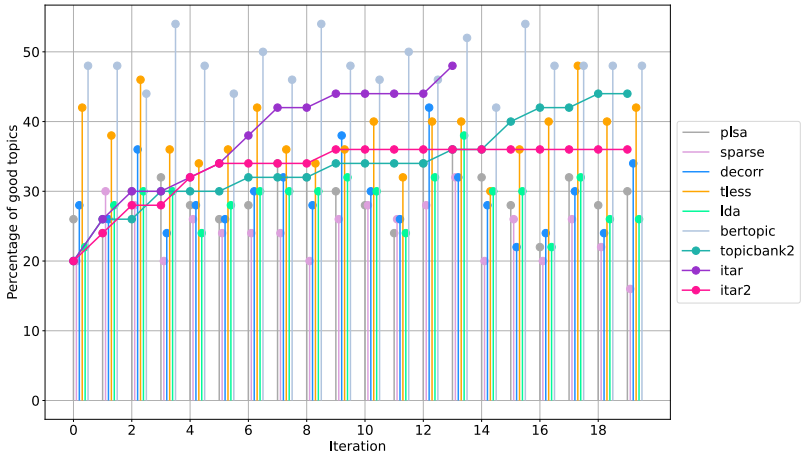
Из сказанного, можно отметить несколько возможных направлений дальнейших исследований по развитию подхода. Более тщательное и системное сравнение эффективности отсеивания плохих тем регуляризаторами (31) и (32). Ускорение обучения итеративной модели (в идеале — вообще за одну итерацию). Исследование возможности использования других критериев для отбора тем (например, человеческую оценку интерпретируемости или оценку, полученную с помощью LLM [114]). Исполь-

[15] https://maartengr.github.io/BERTopic/getting_started/guided/guided.html.

зование Twitter, Reddit или других больших датасетов, с тем чтобы понять, насколько применимы предложенные модели для анализа больших данных, накопленных пользователями в Интернете. Увеличение средней когерентности накапливаемых в итеративной модели тем. Исследование возможности снижения перплексии обученной итеративной модели. Изучение вопроса о возможности получить все 100% хороших тем в модели. Создание внутритекстовой когерентности, считающейся для тем независимо (или, по крайней мере, не так зависимо, как предложено в [6]).



а) 20Newsgroups, $T = 50$.



б) ICD-10, $T = 50$.

Рисунок 32 — Процент хороших тем модели в зависимости от итерации (↑). Хорошесть темы определяется по значению её *внутритекстовой* когерентности. Так как оценки качества разных тем по внутритекстовой когерентности не являются независимыми, в данном случае итеративной модели сложнее накапливать хорошие темы. (В чём можно убедиться на графике для ITAR2 модели, когда больше половины итераций прошло вообще без добавления новых тем; Более того, видно, что модель TopicBank2 может получиться лучше, чем ITAR2, потому что в случае TopicBank модели на разных итерациях обучаются независимо друг от друга, а потому и накопленные хорошие темы не оказывают влияния на оценку качества новых тем; в ITAR2 же применяется ещё и попарная декорреляция с отложенными хорошими, что ещё больше сужает область поиска новых тем. График же для ITAR прерывается (до достижения максимальной итерации), потому что хороших тем накопилось такое количество, что их фиксация с помощью регуляризации привела к вырождению оставшихся тем).

Заключение

Основные результаты работы заключаются в следующем.

1. Предложена внутритекстовая когерентность как метод оценки интерпретируемости темы по распределению её слов в тексте. В отличие от часто используемой на практике когерентности Ньюмана по самым частым словам темы, внутритекстовая когерентность при оценке интерпретируемости темы *полностью* учитывает её распределение по тексту коллекции, что делает её более надёжным внутренним критерием качества тематических моделей.
2. Реализованы алгоритмы вычисления когерентности и обучения интерпретируемых тематических моделей в рамках библиотеки `TopicNet`. Когерентность — в качестве "скора", алгоритмы обучения — в качестве "рецептов" в разделе "cooking machine" библиотеки. Помимо когерентности, также в рамках работы над библиотекой `TopicNet` опубликованы несколько датасетов естественного языка на русском и английском — с целью предоставления всем желающим возможности проводить собственные эксперименты по тематическому моделированию (как в рамках библиотеки `TopicNet`, так с помощью других инструментов).
3. Разработана библиотека `OptimalNumberOfTopics` для оценки качества тематических моделей по внутренним критериям. На момент публикации библиотека содержала самый большой набор доступных критериев среди всех аналогичных библиотек с открытым кодом. Отдельно стоит отметить возможность оценивать качество тематических моделей по внутритекстовой когерентности и по "Банку тем" — доступные исключительно в `OptimalNumberOfTopics`.
4. Представлен метод `TopicBank` оценки качества тематических моделей с учётом их неустойчивости и неполноты. Реализован как часть библиотеки `OptimalNumberOfTopics`. Основное его назначение, помимо определения "оптимального числа тем" — полуавтоматическая оценка качества тематических моделей.

5. Предложен и реализован многопроходной алгоритм улучшения тематической модели с помощью обратной связи от пользователя `ITAR`, повышающий устойчивость и полноту итоговой модели по сравнению с одиночными моделями. Данный алгоритм является развитием идеи “Банка тем”. Сравнением модели `ITAR` с другими тематическими моделями на ряде текстовых коллекций естественного языка показана состоятельность подхода.

Список литературы

1. Statistical topic models for multi-label document classification / T. N. Rubin, A. Chambers, P. Smyth [и др.] // Machine learning. — 2012. — Т. 88. — С. 157—208.
2. *Ianina A.*, *Golitsyn L.*, *Vorontsov K.* Multi-objective topic modeling for exploratory search in tech news // Conference on Artificial Intelligence and Natural Language. — Springer. 2017. — С. 181—193.
3. An overview of topic modeling and its current applications in bioinformatics / L. Liu, L. Tang, W. Dong [и др.] // SpringerPlus. — 2016. — Т. 5. — С. 1—22.
4. *Steyvers M.*, *Griffiths T.* Probabilistic topic models // Handbook of latent semantic analysis. — 2007. — Т. 427, № 7. — С. 424—440.
5. *Vorontsov K.*, *Potapenko A.* Additive regularization of topic models // Machine Learning. — 2015. — Т. 101, № 1—3. — С. 303—323.
6. *Alekseev V.*, *Bulatov V.*, *Vorontsov K.* Intra-text coherence as a measure of topic models' interpretability // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue. — 2018. — С. 1—13.
7. Topic Modelling for Extracting Behavioral Patterns from Transactions Data / E. Egorov, F. Nikitin, V. Alekseev [и др.] // 2019 International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI). — IEEE. 2019. — С. 44—444.
8. TopicNet: Making Additive Regularisation for Topic Modelling Accessible / V. Bulatov, V. Alekseev, K. Vorontsov [и др.] // Proceedings of The 12th Language Resources and Evaluation Conference. — 2020. — С. 6745—6752.
9. TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation / V. Alekseev, E. Egorov, K. Vorontsov [и др.] // Data & Knowledge Engineering. — 2021.

10. *Bulatov V.*, *Alekseev V.*, *Vorontsov K.* Determination of the Number of Topics Intrinsically: Is It Possible? // Recent Trends in Analysis of Images, Social Networks and Texts. — Cham : Springer Nature Switzerland, 2024. — С. 3—17. — ISBN 978-3-031-67008-4. — DOI: 10.1007/978-3-031-67008-4_1.

11. *Алексеев В. А.*, *Булатов В. Г.* Внутритекстовая когерентность как мера интерпретируемости тематических моделей текстовых коллекций // Труды 60-й Всероссийской научной конференции МФТИ. — 2017. — С. 84—86.

12. *Алексеев В. А.* Банк тем: сбор интерпретируемых тем с помощью множественного обучения тематических моделей и их дальнейшее использование для оценки качества тематических моделей // Труды 64-й Всероссийской научной конференции МФТИ. — 2021. — С. 149—151.

13. *Алексеев В. А.*, *Воронцов К. В.* Банк тем: сбор интерпретируемых тем с помощью множественного обучения тематических моделей и их дальнейшее использование для оценки качества тематических моделей // Математические методы распознавания образов: Тезисы докладов 20-й Всероссийской конференции с международным участием. — 2021. — С. 313—315.

14. *Bulatov V. G.*, *Alekseev V. A.* Determination of the Number of Topics Intrinsically: Is It Possible? // Труды 66-й Всероссийской научной конференции МФТИ (в печати). — 2024.

15. *Горбулев А. И.*, *Алексеев В. А.* Итеративное улучшение аддитивно регуляризованной тематической модели // Труды 66-й Всероссийской научной конференции МФТИ (в печати). — 2024.

16. *Gorbulev A.*, *Alekseev V.*, *Vorontsov K.* Iterative Improvement of an Additively Regularized Topic Model // International Conference on Analysis of Images, Social Networks and Texts (In Press). — Springer. 2024.

17. *Свидетельство о государственной регистрации программы для ЭВМ*. Система разведочного поиска / К. В. Воронцов, А. В. Гончаров, Е. О.

Егоров [и др.] ; федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)». — № 2020616114 ; заявл. 18.06.2020 ; опубл. 30.06.2020, RU 2020617007 (Рос. Федерация).

18. *Свидетельство о государственной регистрации программы для ЭВМ.* Программа сегментации и профилирования поведения пользователей транзакционных систем / К. В. Воронцов, А. В. Гончаров, Е. О. Егоров [и др.] ; федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)». — № 2021614410 ; заявл. 30.03.2021 ; опубл. 18.05.2021, RU 2021617632 (Рос. Федерация).

19. *Blei D. M.* Probabilistic topic models // Communications of the ACM. — 2012. — Т. 55, № 4. — С. 77—84.

20. *Wang C.*, *Blei D. M.* Collaborative topic modeling for recommending scientific articles // Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2011. — С. 448—456.

21. *Varshney D.*, *Kumar S.*, *Gupta V.* Modeling information diffusion in social networks using latent topic information // Intelligent Computing Theory: 10th International Conference, ICIC 2014, Taiyuan, China, August 3-6, 2014. Proceedings 10. — Springer. 2014. — С. 137—148.

22. *Pinto J. C. L.*, *Chahed T.* Modeling multi-topic information diffusion in social networks using latent Dirichlet allocation and Hawkes processes // 2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems. — IEEE. 2014. — С. 339—346.

23. *Lee S. S.*, *Chung T.*, *McLeod D.* Dynamic item recommendation by topic modeling for social networks // 2011 Eighth International Conference on Information Technology: New Generations. — IEEE. 2011. — С. 884—889.

24. *Ianina A.*, *Golitsyn L.*, *Vorontsov K.* Multi-objective topic modeling for exploratory search in tech news // Artificial Intelligence and Natural Language: 6th Conference, AINL 2017, St. Petersburg, Russia, September 20–23, 2017, Revised Selected Papers 6. — Springer. 2018. — С. 181—193.

25. Applications of topic models / J. Boyd-Graber, Y. Hu, D. Mimno [и др.] // Foundations and Trends® in Information Retrieval. — 2017. — Т. 11, № 2/3. — С. 143—296.

26. *Narayan S.*, *Cohen S. B.*, *Lapata M.* Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. — 2018. — С. 1797—1807.

27. A Reinforced Topic-Aware Convolutional Sequence-to-Sequence Model for Abstractive Text Summarization / L. Wang, J. Yao, Y. Tao [и др.] // International Joint Conference on Artificial Intelligence. — 2018.

28. Scoring Sentence Singletons and Pairs for Abstractive Summarization / L. Lebanoff, K. Song, F. Dernoncourt [и др.] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 07.2019. — С. 2175—2189. — DOI: 10.18653/v1/P19-1209. — URL: https://www.aclweb.org/anthology/P19-1209.

29. Topic Modeling Users' Interpretations of Songs to Inform Subject Access in Music Digital Libraries / K. Choi, J. H. Lee, C. Willis [и др.] // Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries. — ACM. 2015. — С. 183—186.

30. Song Lyrics Summarization Inspired by Audio Thumbnailing / M. Fell, E. Cabrio, F. Gandon [и др.] // RANLP. — 2019. — С. 328—337.

31. Modeling Storylines in Lyrics / K. Watanabe, Y. Matsubayashi, K. Inui [и др.] // IEICE TRANSACTIONS on Information and Systems. — 2018. — Т. 101, № 4. — С. 1167—1179.

32. Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models / T. Funnell, A. W. Zhang, D.

Grewal [и др.] // PLoS computational biology. — 2019. — Т. 15, № 2. — e1006799.

33. *Antons D.*, *Joshi A. M.*, *Salge T. O.* Content, contribution, and knowledge consumption: Uncovering hidden topic structure and rhetorical signals in scientific texts // Journal of Management. — 2019. — Т. 45, № 7. — С. 3035—3076.

34. *Lautamatti L.* Observations on the development of the topic in simplified discourse // AFinLAn vuosikirja. — 1978. — С. 71—104.

35. Knowledge discovery through directed probabilistic topic models: a survey / A. Daud, J. Li, L. Zhou [и др.] // Frontiers of computer science in China. — 2010. — Т. 4, № 2. — С. 280—301.

36. *Vorontsov K.*, *Potapenko A.* Additive regularization of topic models // Machine Learning. — 2014. — Дек. — Т. 101. — С. 1—21. — DOI: 10.1007/s10994-014-5476-6.

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

38. *Vorontsov K.*, *Potapenko A.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // International Conference on Analysis of Images, Social Networks and Texts. — Springer. 2014. — С. 29—46.

39. BigARTM: Open source library for regularized multimodal topic modeling of large collections / K. Vorontsov, O. Frei, M. Apishev [и др.] // Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers 4. — Springer. 2015. — С. 370—381.

40. *Veselova E.*, *Vorontsov K.* Topic balancing with additive regularization of topic models // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. — 2020. — С. 59—65.

41. *Chirkova N.*, *Vorontsov K.* Additive regularization for hierarchical multimodal topic modeling // Journal Machine Learning and Data Analysis. — 2016. — Т. 2, № 2. — С. 187—200.

42. *Ianina A.*, *Vorontsov K.* Regularized multimodal hierarchical topic model for document-by-document exploratory search // 2019 25th Conference of Open Innovations Association (FRUCT). — IEEE. 2019. — С. 131—138.

43. *Vorontsov K.*, *Potapenko A.*, *Plavin A.* Additive regularization of topic models for topic selection and sparse factorization // International Symposium on Statistical Learning and Data Sciences. — 2015.

44. *Jordan M. I.*, *Mitchell T. M.* Machine learning: Trends, perspectives, and prospects // Science. — 2015. — Т. 349, № 6245. — С. 255—260.

45. *Khandani A. E.*, *Kim A. J.*, *Lo A. W.* Consumer credit-risk models via machine-learning algorithms // Journal of Banking & Finance. — 2010. — Т. 34, № 11. — С. 2767—2787.

46. *Chakraborty C.*, *Joseph A.* Machine learning at central banks. — 2017.

47. *Hamid A. J.*, *Ahmed T. M.* Developing prediction model of loan risk in banks using data mining // Machine Learning and Applications: An International Journal (MLAIJ) Vol. — 2016. — Т. 3.

48. Time-aware user identification with topic models / C. Lesaege, F. Schnitzler, A. Lambert [и др.] // 2016 IEEE 16th International Conference on Data Mining (ICDM). — IEEE. 2016. — С. 997—1002.

49. A Combination Approach to Web User Profiling / J. Tang, L. Yao, D. Zhang [и др.] // ACM Trans. Knowl. Discov. Data. — New York, NY, USA, 2010. — Дек. — Т. 5, № 1. — 2:1—2:44. — ISSN 1556-4681. — DOI: 10.1145/1870096.1870098.

50. *Baldassini L.*, *Serrano J. A. R.* client2vec: towards systematic baselines for banking applications // arXiv preprint arXiv:1802.04198. — 2018.

51. A balanced view for customer segmentation in CRM / J. Yoon, S. Hwang, D. Kim [и др.] // AMCIS 2003 Proceedings. — 2003. — С. 67.

52. Modelling web-based banking systems: Story boarding and user profiling / K.-D. Schewe, R. Kaschek, C. Matthews [и др.] // International Conference on Conceptual Modeling. — Springer. 2002. — С. 427—439.

53. Machine learning techniques applied to profile mobile banking users in India / M. Carr, V. Ravi, G. S. Reddy [и др.] // International Journal of Information Systems in the Service Sector (IJISSS). — 2013. — Т. 5, № 1. — С. 82—92.

54. *Gladstone J. J.*, *Matz S. C.*, *Lemaire A.* Can Psychological Traits Be Inferred From Spending? Evidence From Transaction Data // Psychological science. — 2019. — С. 0956797619849435.

55. *Ładyżyński P.*, *Żbikowski K.*, *Gawrysiak P.* Direct marketing campaigns in retail banking with the use of deep learning and random forests // Expert Systems with Applications. — 2019. — Т. 134. — С. 28—35.

56. *Reisenbichler M.*, *Reutterer T.* Topic modeling in marketing: recent advances and research opportunities // Journal of Business Economics. — 2019. — Т. 89, № 3. — С. 327—356.

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

58. Topic sentiment mixture: modeling facets and opinions in weblogs / Q. Mei, X. Ling, M. Wondra [и др.] // Proceedings of the 16th international conference on World Wide Web. — ACM. 2007. — С. 171—180.

59. *Hospedales T.*, *Gong S.*, *Xiang T.* Video behaviour mining using a dynamic topic model // International journal of computer vision. — 2012. — Т. 98, № 3. — С. 303—323.

60. Unsupervised dialogue intent detection via hierarchical topic model / A. Popov, V. Bulatov, D. Polyudova [и др.] // RANLP. — 2019. — С. 932—938.

61. CatBoost: unbiased boosting with categorical features / L. Prokhorenkova, G. Gusev, A. Vorobev [и др.] // Advances in Neural Information Processing Systems. — 2018. — С. 6638—6648.

62. The human touch: How non-expert users perceive, interpret, and fix topic models / T. Y. Lee, A. Smith, K. Seppi [и др.] // International Journal of Human-Computer Studies. — 2017. — Т. 105. — С. 28—42.

63. *Agrawal A.*, *Fu W.*, *Menzies T.* What is wrong with topic modeling? and how to fix it using search-based software engineering // Information and Software Technology. — 2018. — Т. 98. — С. 74—88.

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

65. *Wallach H.*, *Mimno D.*, *McCallum A.* Rethinking LDA: Why priors matter // Advances in neural information processing systems. — 2009. — Т. 22.

66. *Vorontsov K. V.* Additive Regularization for Topic Models of Text Collections // Doklady Mathematics. — 2014. — Т. 89, № 3. — С. 301—304.

67. Fast and modular regularized topic modelling / D. Kochedykov, M. Apishev, L. Golitsyn [и др.] // 2017 21st Conference of Open Innovations Association (FRUCT). — IEEE. 2017. — С. 182—193.

68. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections / K. Vorontsov, O. Frei, M. Apishev [и др.] // AIST'2015, Analysis of Images, Social networks and Texts. — Springer International Publishing Switzerland, Communications in Computer, Information Science (CCIS), 2015. — С. 370—384.

69. On Smoothing and Inference for Topic Models / A. Asuncion, M. Welling, P. Smyth [и др.] // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. — Montreal, Quebec, Canada, 2009. — С. 27—34.

70. *Řehůřek R.*, *Sojka P.* Software Framework for Topic Modelling with Large Corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. — Valletta, Malta : ELRA, 05.2010. — С. 45—50. — http://is.muni.cz/publication/884893/en.

71. *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. — 2015. — С. 399—408.

72. Topic Modeling for the Social Sciences / D. Ramage, E. Rosen, J. Chuang [и др.] // Neural Information Processing Systems (NIPS) Workshop on Applications for Topic Models: Text and Beyond. Т. 5. — 2009. — С. 1—4. — URL: https://nlp.stanford.edu/dramage/papers/tmt-nips09.pdf.

73. *McCallum A. K.* Mallet: A machine learning for language toolkit // http://mallet.cs.umass.edu. — 2002.

74. *Li W.*, *McCallum A.* Pachinko allocation: DAG-structured mixture models of topic correlations // Proceedings of the 23rd international conference on Machine learning. — ACM. 2006. — С. 577—584.

75. *Pol M.*, *Walkowiak T.*, *Piasecki M.* Towards CLARIN-PL LTC Digital Research Platform for: Depositing, Processing, Analyzing and Visualizing Language Data // International Conference on Reliability and Statistics in Transportation and Communication. — Springer. 2017. — С. 485—494.

76. STTM: A Tool for Short Text Topic Modeling / J. Qiang, Y. Li, Y. Yuan [и др.] // arXiv preprint arXiv:1808.02215. — 2018.

77. *Yin J.*, *Wang J.* A dirichlet multinomial mixture model-based approach for short text clustering // Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. — ACM. 2014. — С. 233—242.

78. *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.

79. Topic modeling of short texts: A pseudo-document view / Y. Zuo, J. Wu, H. Zhang [и др.] // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. — 2016. — С. 2105—2114.

80. Short and sparse text topic modeling via self-aggregation / X. Quan, C. Kit, Y. Ge [и др.] // Twenty-Fourth International Joint Conference on Artificial Intelligence. — 2015.

81. Improving topic models with latent feature word representations / D. Q. Nguyen, R. Billingsley, L. Du [и др.] // Transactions of the Association for Computational Linguistics. — 2015. — Т. 3. — С. 299—313.

82. Familia: A Configurable Topic Modeling Framework for Industrial Text Engineering / D. Jiang, Y. Song, R. Lian [и др.] // arXiv preprint arXiv:1808.03733. — 2018.

83. *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.

84. *Gao J.*, *Toutanova K.*, *Yih W.-t.* Clickthrough-based latent semantic models for web search // Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. — ACM. 2011. — С. 675—684.

85. Mining geographic knowledge using location aware topic model / C. Wang, J. Wang, X. Xie [и др.] // Proceedings of the 4th ACM workshop on Geographical information retrieval. — ACM. 2007. — С. 65—70.

86. *Lian R.* Project Title. — 2019. — https://github.com/baidu/Familia/issues/81.

87. *Frei O.*, *Apishev M.* Parallel non-blocking deterministic algorithm for online topic modeling // International Conference on Analysis of Images, Social Networks and Texts. — Springer. 2016. — С. 132—144.

88. *Yanina A.*, *Golitsyn L.*, *Vorontsov K.* Multi-objective Topic Modeling for Exploratory Search in Tech News // Communications in Computer and Information Science, vol 789. AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017 / под ред. A. Filchenkov, L. Pivovarova, J. Žižka. — Springer International Publishing, Cham, 2018. — С. 181—193.

89. *Potapenko A.*, *Popov A.*, *Vorontsov K.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks // Communications in Computer and Information Science, vol 789. AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017. — Springer, Cham, 2017. — С. 167—180.

90. *Chirkova N. A.*, *Vorontsov K. V.* Additive Regularization for Hierarchical Multimodal Topic Modeling // Journal Machine Learning and Data Analysis. — 2016. — Т. 2, № 2. — С. 187—200.

91. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections / K. Vorontsov, O. Frei, M. Apishev [и др.] // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. — Melbourne, Australia : ACM, 2015. — С. 29—37.

92. *Sokolov E.*, *Bogolubsky L.* Topic Models Regularization and Initialization for Regression Problems // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. — Melbourne, Australia : ACM, 2015. — С. 21—27.

93. Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Text Content / M. Apishev, S. Koltcov, O. Koltsova [и др.] // MICAI 2016, 15th Mexican International Conference on Artificial Intelligence. Т. 10061. — Springer, Lecture Notes in Artificial Intelligence, 2016. — С. 166—181.

94. Mining Ethnic Content Online with Additively Regularized Topic Models / M. Apishev, S. Koltcov, O. Koltsova [и др.] // Computacion y Sistemas. — 2016. — Т. 20, № 3. — С. 387—403.

95. *Vorontsov K. V.*, *Potapenko A. A.*, *Plavin A. V.* Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // The Third International Symposium On Learning And Data Sciences (SLDS 2015). April 20-22, 2015. Royal Holloway, University of London, UK. / под ред. A. G. et al. — Springer International Publishing Switzerland 2015, 2015. — С. 193—202.

96. *Skachkov N.*, *Vorontsov K.* Improving topic models with segmental structure of texts // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue. — 2018. — С. 652—661.

97. *Mavrin A.*, *Filchenkov A.*, *Koltcov S.* Four Keys to Topic Interpretability in Topic Modeling // Conference on Artificial Intelligence and Natural Language. — Springer. 2018. — С. 117—129.

98. *Vorontsov K. V.*, *Potapenko A. A.* Additive Regularization of Topic Models // Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications. — 2015. — Т. 101, № 1. — С. 303—323.

99. *Potapenko A.*, *Popov A.*, *Vorontsov K.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks // Artificial Intelligence and Natural Language: 6th Conference, AINL 2017, St. Petersburg, Russia, September 20–23, 2017, Revised Selected Papers 6. — Springer. 2018. — С. 167—180.

100. Optimizing semantic coherence in topic models / D. Mimno, H. Wallach, E. Talley [и др.] // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics. 2011. — С. 262—272.

101. *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.

102. Automatic evaluation of topic coherence / D. Newman, J. H. Lau, K. Grieser [и др.] // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Association for Computational Linguistics. 2010. — С. 100—108.

103. Reading tea leaves: How humans interpret topic models / J. Chang, S. Gerrish, C. Wang [и др.] // NIPS 2009. — 2009. — Т. 22.

104. Perplexity—a measure of the difficulty of speech recognition tasks / F. Jelinek, R. L. Mercer, L. R. Bahl [и др.] // The Journal of the Acoustical Society of America. — 1977. — Т. 62, S1. — S63—S63.

105. *Koltcov S.* Application of Rényi and Tsallis entropies to topic modeling optimization // Physica A: Statistical Mechanics and its Applications. — 2018. — Т. 512. — С. 1192—1204.

106. *Mehta V.*, *Caceres R. S.*, *Carter K. M.* Evaluating topic quality using model clustering // 2014 IEEE Symposium on Computational Intelligence and Data Mining.

107. *Gerlach M.*, *Peixoto T. P.*, *Altmann E. G.* A network approach to topic models // Science advances. — 2018.

108. *Zamzami N.*, *Bouguila N.* MML-Based Approach for Determining the Number of Topics in EDCM Mixture Models // Advances in Artificial Intelligence. — 2018.

109. *Than K.*, *Ho T. B.* Fully sparse topic models // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. — 2012.

110. *Blei D. M.*, *Lafferty J. D.* Topic models // Text mining. — Chapman, Hall/CRC, 2009. — С. 101—124.

111. *Halliday M. A. K.*, *Hasan R.* Cohesion in english. — Routledge, 2014.

112. *Schmidt B. M.* Words alone: Dismantling topic models in the humanities // Journal of Digital Humanities. — 2012. — Т. 2, № 1. — С. 49—65.

113. *Grootendorst M.* BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv preprint arXiv:2203.05794. — 2022.

114. LLM Reading Tea Leaves: Automatically Evaluating Topic Models with Large Language Models / X. Yang, H. Zhao, D. Phung [и др.] // arXiv preprint arXiv:2406.09008. — 2024.

115. On finding the natural number of topics with latent dirichlet allocation: Some observations / R. Arun, V. Suresh, C. V. Madhavan [и др.] // Pacific-Asia conference on knowledge discovery and data mining. — 2010.

116. A density-based method for adaptive LDA model selection / J. Cao, T. Xia, J. Li [и др.] // Neurocomputing. — 2009.

117. *Deveaud R.*, *SanJuan E.*, *Bellot P.* Accurate and effective latent concept modeling for ad hoc information retrieval // Document numérique. — 2014.

118. *Greene D.*, *O'Callaghan D.*, *Cunningham P.* How many topics? stability analysis for topic models // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. — 2014.

119. *Griffiths T. L.*, *Steyvers M.* Finding scientific topics // Proceedings of the National academy of Sciences. — 2004.

120. A heuristic approach to determine an appropriate number of topics in topic modeling / W. Zhao, J. J. Chen, R. Perkins [и др.] // BMC bioinformatics. — 2015.

121. *Krasnov F.*, *Sen A.* The number of topics optimization: clustering approach // Machine Learning and Knowledge Extraction. — 2019.

122. *Bryant M.*, *Sudderth E. B.* Truly nonparametric online variational inference for hierarchical Dirichlet processes // Advances in Neural Information Processing Systems. — 2012.

123. *Gialampoukidis I.*, *Vrochidis S.*, *Kompatsiaris I.* A hybrid framework for news clustering based on the DBSCAN-Martingale and LDA // International Conference on Machine Learning and Data Mining in Pattern Recognition. — 2016.

124. *Murzintcev Nikita N. C.* ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. — 2020. — URL: https://CRAN.R-project.org/package=ldatuning.

125. *Hou-Liu J.* Benchmarking and improving recovery of number of topics in latent Dirichlet allocation models. — 2018.

126. *Guille A.*, *Soriano-Morales E.-P.* TOM: A library for topic modeling and browsing. //.

127. *Tang J.*, *Zhang M.*, *Mei Q.* "Look Ma, No Hands!"A Parameter-Free Topic Model // arXiv preprint arXiv:1409.2993. — 2014.

128. *Fan A.*, *Doshi-Velez F.*, *Miratrix L.* Assessing topic model relevance: Evaluation and informative priors // Statistical Analysis and Data Mining: The ASA Data Science Journal. — 2019.

129. *Soleimani H.*, *Miller D. J.* Parsimonious topic models with salient word discovery // IEEE Transactions on Knowledge and Data Engineering. — 2014.

130. Metagenes and molecular pattern discovery using matrix factorization / J.-P. Brunet, P. Tamayo, T. R. Golub [и др.] // Proceedings of the National Academy of Sciences. — 2004.

131. Emerging Topics in Brexit Debate on Twitter Around the Deadlines / E. del Gobbo, S. Fontanella, A. Sarra [и др.] // Social Indicators Research. — 2020.

132. Configuring latent dirichlet allocation based feature location / L. R. Biggers, C. Bocovich, R. Capshaw [и др.] // Empirical Software Engineering. — 2014.

133. *Tan Y.*, *Ou Z.* Topic-weak-correlated latent dirichlet allocation // 2010 7th International Symposium on Chinese Spoken Language Processing. — IEEE. 2010.

134. *Potapenko A.*, *Vorontsov K.* Robust PLSA performs better than LDA // European Conference on Information Retrieval. — 2013.

135. *Irkhin I.*, *Bulatov V.*, *Vorontsov K.* Additive regularizarion of topic models with fast text vectorizartion (in Russian) // Computer research and modeling. — 2020. — Т. 12, № 6. — С. 1515—1528.

136. *Korobov M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. Т. 542 / под ред. M. Y. Khachay, N. Konstantinova, A. Panchenko [и др.]. — Springer International Publishing, 2015. — С. 320—332. — (Communications in Computer and Information Science). — ISBN 978-3-319-26122-5. — DOI: 10.1007/978-3-319-26123-2_31.

137. *Barua A.*, *Thomas S. W.*, *Hassan A. E.* What are developers talking about? an analysis of topics and trends in stack overflow // Empirical Software Engineering. — 2014.

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

139. *De Waal A.*, *Barnard E.* Evaluating topic models with stability. — 2008.

140. *Koltcov S.*, *Koltsova O.*, *Nikolenko S.* Latent dirichlet allocation: stability and applications to studies of user-generated content // Proceedings of the 2014 ACM conference on Web science. — ACM. 2014. — С. 161—165.

141. *Greene D.*, *O'Callaghan D.*, *Cunningham P.* How many topics? stability analysis for topic models // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. — Springer. 2014. — С. 498—513.

142. *Balagopalan A.* Improving topic reproducibility in topic models. — University of California, Irvine, 2012.

143. *Mehta V.*, *Caceres R. S.*, *Carter K. M.* Evaluating topic quality using model clustering // 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). — IEEE. 2014. — С. 178—185.

144. Computing a nonnegative matrix factorization–provably / S. Arora, R. Ge, R. Kannan [и др.] // Proceedings of the forty-fourth annual ACM symposium on Theory of computing. — ACM. 2012. — С. 145—162.

145. *Arora S.*, *Ge R.*, *Moitra A.* Learning topic models–going beyond SVD // 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science. — IEEE. 2012. — С. 1—10.

146. A practical algorithm for topic modeling with provable guarantees / S. Arora, R. Ge, Y. Halpern [и др.] // International Conference on Machine Learning. — 2013. — С. 280—288.

147. *Dobrynin V.*, *Patterson D.*, *Rooney N.* Contextual document clustering // European Conference on Information Retrieval. — Springer. 2004. — С. 167—180.

148. *Lewis D. D.* Reuters-21578 (Dataset). — 1997. — URL: http://www.daviddlewis.com/resources/testcollections/reuters21578/.

149. *Kucera H.*, *Francis W. N.* Brown (Dataset). — 1961. — URL: http://korpus.uib.no/icame/brown/bcm.html.

150. *Lang K.* 20 Newsgroups (Dataset). — 1995. — URL: http://qwone.com/~jason/20Newsgroups/.

151. *Gulli A.* AG News (Dataset). — 2015. — URL: http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html.

152. *Abbas M.* Watan-2004 (Dataset). — 2004. — URL: https://sites.google.com/site/mouradabbas9/corpora/text-corpora.

153. *Vorontsov K.*, *Potapenko A.*, *Plavin A.* Additive regularization of topic models for topic selection and sparse factorization // International Symposium on Statistical Learning and Data Sciences. — Springer. 2015. — С. 193—202.

154. *DiMaggio P.*, *Nag M.*, *Blei D.* Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding // Poetics. — 2013. — Т. 41, № 6. — С. 570—606. — ISSN 0304-422X. — DOI: 10.1016/j.poetic.2013.08.004.

155. *Aqui J.*, *Hosein M.* Mobile Ad-hoc Networks Topic Modelling and Dataset Querying // 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC). — IEEE. 2022. — С. 1—6. — DOI: 10.1109/ICMNWC56175.2022.10031921.

156. Topic Modelling of Croatian News During COVID-19 Pandemic / P. K. Bogović, A. Meštrović, S. Beliga [и др.] // 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO). — 2021. — С. 1044—1051. — DOI: 10.23919/MIPRO52101.2021.9597125.

157. *Rabitz F.*, *Telešienė A.*, *Zolubienė E.* Topic modelling the news media representation of climate change // Environmental Sociology. — 2021. — Т. 7, № 3. — С. 214—224. — DOI: 10.1080/23251042.2020.1866281.

158. *Grajzl P.*, *Murrell P.* Toward understanding 17th century English culture: A structural topic model of Francis Bacon's ideas // Journal of Comparative Economics. — 2019. — Т. 47, № 1. — С. 111—135.

159. *Zhao W.*, *Zou W.*, *Chen J. J.* Topic modeling for cluster analysis of large biological and medical datasets // BMC Bioinformatics. Т. 15. — Springer. 2014. — С. 1—11.

160. Topic modeling algorithms and applications: A survey / A. Abdelrazek, Y. Eid, E. Gawish [и др.] // Information Systems. — 2023. — Т. 112. — С. 102131.

161. *Suhareva A. V.*, *Voroncov K. V.* Postroenie polnogo nabora tem verojatnostnyh tematičeskih modelej (in Russian) // Intellektual'nye sistemy. Teorija i priloženija. — 2019. — T. 23, № 4. — C. 7—23.

162. *Vorontsov K.* Additive regularization for topic models of text collections // Doklady Mathematics. T. 89. — Pleiades Publishing. 2014. — C. 301—304.

163. *Dai S.-C.*, *Xiong A.*, *Ku L.-W.* LLM-in-the-loop: Leveraging large language model for thematic analysis // arXiv preprint arXiv:2310.15100. — 2023.

164. *Wang X.*, *Yang Y.* Neural topic model with attention for supervised learning // International conference on artificial intelligence and statistics. — PMLR. 2020. — C. 1147—1156.

165. ANTM: Aligned Neural Topic Models for Exploring Evolving Topics / H. Rahimi, H. Naacke, C. Constantin [и др.] // Transactions on Large-Scale Data-and Knowledge-Centered Systems LVI: Special Issue on Data Management-Principles, Technologies, and Applications. — Springer, 2024. — C. 76—97.

166. BERT: Pre-training of deep bidirectional transformers for language understanding / J. Devlin, M.-W. Chang, K. Lee [и др.] // arXiv preprint arXiv:1810.04805. — 2018.

167. *McInnes L.*, *Healy J.*, *Melville J.* Umap: Uniform manifold approximation and projection for dimension reduction // arXiv preprint arXiv:1802.03426. — 2018.

168. hdbscan: Hierarchical density based clustering / L. McInnes, J. Healy, S. Astels [и др.] // J. Open Source Softw. — 2017. — T. 2, № 11. — C. 205.

169. *Sparck Jones K.* A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. — 1972. — T. 28, № 1. — C. 11—21.

170. *Skachkov N.*, *Vorontsov K.* Improving topic models with segmental structure of texts // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2018. — C. 652—661.

171. Incremental topic modeling for scientific trend topics extraction / N. Gerasimenko, A. Chernyavskiy, M. Nikiforova [и др.] // Proceedings of the International Conference Dialogue. — 2023.

172. *Fano R. M.* Transmission of information: a statistical theory of communications. — MIT press, 1968.

173. *Church K.*, *Hanks P.* Word association norms, mutual information, and lexicography // Computational Linguistics. — 1990. — Т. 16, № 1. — С. 22—29.

174. *Dagan I.*, *Marcus S.*, *Markovitch S.* Contextual word similarity and estimation from sparse data // 31st Annual Meeting of the Association for Computational Linguistics. — 1993. — С. 164—171.

175. Quality evaluation and improvement for hierarchical topic modeling / A. Belyy, M. Seleznova, A. Sholokhov [и др.] // Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2018. — 2018. — С. 110—123.

176. *Rosen C.*, *Shihab E.* What are mobile developers asking about? A large scale study using stack overflow // Empirical Software Engineering. — 2016. — Т. 21. — С. 1192—1223.