МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Физтех-школа Прикладной Математики и Информатики (ФПМИ)
Кафедра интеллектуальных систем

Алексеев Василий Антонович

НЕПОЛНОТА И НЕУСТОЙЧИВОСТЬ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

(научный доклад об основных результатах подготовленной научно-квалификационной работы (диссертации))

Направление 09.06.01 — «Информатика и вычислительная техника»

Направленность (профиль) — «Информатика и вычислительная техника»

Специальность 1.2.3 — «Теоретическая информатика, кибернетика»

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

Москва — 2024

Общая характеристика работы

Актуальность темы. 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 по-разному.) 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.

Авторами[5] предложен подход к обучению тематических моделей, названный аддитивной регуляризацией тематических моделей (Additive Regularization of Topic Models, ARTM). Регуляризаторы поз-

[1] Statistical topic models for multi-label document classification / T. N. Rubin [и др.] // Machine learning. 2012. Т. 88. С. 157—208.

[2] *Ianina A.* [и др.]. 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 [и др.] // SpringerPlus. 2016. Т. 5. С. 1—22. DOI: 10.1186/s40064-016-3252-8. URL: https://doi.org/10.1186/s40064-016-3252-8.

[4] *Steyvers M.* [и др.]. Probabilistic topic models // Handbook of latent semantic analysis. 2007. Т. 427, № 7. С. 424—440.

[5] *Vorontsov K.* [и др.]. Additive regularization of topic models // Machine Learning. 2015. Т. 101, № 1—3. С. 303—323.

воляют 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 свойствами, 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 для решения проблем неполноты и неустойчивости тематических моделей с помощью множественного обучения моделей.

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 когерентность и алгоритмы обучения интерпретируемых тематических моделей в рамках библиотеки `TopicNet`.
3. Разработана библиотека `OptimalNumberOfTopics` для оценки качества тематических моделей по внутренним критериям.
4. Представлен метод `TopicBank` оценки качества тематических моделей с учётом их неустойчивости и неполноты.
5. Предложен многопроходной алгоритм улучшения тематической модели с помощью обратной связи от пользователя `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.

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).

Достоверность полученных результатов обеспечивается проведёнными экспериментами и публикациями. Результаты находятся в соответствии с результатами, полученными другими авторами.

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 использование 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.

Публикации. Основные результаты по теме диссертации изложены в 7 печатных работах, 4 из которых изданы в журналах, рекомендованных ВАК [1—4], 3 работы — в тезисах докладов конференций [5—7]. Помимо этого, ещё 3 работы в данный момент находятся в печати [8—10]. Получены два свидетельства о государственной регистрации программы для ЭВМ [11; 12].

Личный вклад. В работе [6] автором предложены функции внутритекстовой когерентности, создание же полусинтетических датасетов проводилось совместно с Булатовым В. Г. Вклад автора во все основные положения, выносимые на защиту, является решающим.

Содержание работы

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. Далее, приводится пример применения тематического моделирования для анализа транзакций клиентов банка как пример использования тематического моделирования для решения прикладных задач. Глава завершается описанием библиотеки тематического моделирования `TopicNet`, в частности, её преимуществ по сравнению с библиотекой `BigARTM`.

Тематическое моделирование — это направление в статистическом анализе текстов[6]. Задача тематической модели состоит в обнаружении *скрытой тематической структуры* больших коллекций текстовых документов.

Тематические модели используются, например, в информационном поиске[7], категоризации документов[8], анализе данных социальных сетей[9], рекомендательных системах[10], разведочном поиске[11]. После обработки коллекции документов тематическая модель отдаёт набор тем,

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

[7] *Wang C.* [и др.]. 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.

[8] Statistical topic models for multi-label document classification / T. N. Rubin [и др.] // Machine learning. 2012. Т. 88. С. 157—208.

[9] *Varshney D.* [и др.]. 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; *Pinto J. C. L.* [и др.]. 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.

[10] *Wang C.* [и др.]. 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; *Lee S. S.* [и др.]. Dynamic item recommendation by topic modeling for social networks // 2011 Eighth International Conference on Information Technology: New Generations. IEEE. 2011. С. 884—889.

[11] *Ianina A.* [и др.]. 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.

затрагиваемых в документах, информацию о распределении этих тем в документах и слова, характеризующие каждую тему[12].

Пусть D обозначает набор текстовых документов, W — набор всех слов, которые встречаются в документах коллекции (*словарь*). Слово (термин, токен) $w \in W$ может быть буквально отдельным словом или сочетанием слов. После определения того, какие сущности в коллекции следует рассматривать как слова, каждый документ $d \in D$ может быть представлен как упорядоченная последовательность n_d терминов $W_d \subseteq \subseteq W$. Пусть n_{dw} обозначает число, сколько раз термин $w \in W$ встречается в документе $d \in D$.

Получаем, что текстовая коллекция может рассматриваться как выборка троек $\{(d_i, t_i, w_i)\}_{i=1}^{n}$, полученных независимо из дискретного вероятностного распределения $p(d, t, w)$ над конечным пространством $D \times T \times W$.

В тематическом моделировании принимается гипотеза *условной независимости*, которая гласит, что слово относится к теме независимо от того, в каком документе мы наблюдаем это слово в данный момент:

$$p(w \mid d, t) \equiv p(w \mid t)$$

Наконец, считаем, что наблюдаемая коллекция формируется с помощью распределений $p(w \mid t)$ и $p(t \mid d)$. Согласно закону полной вероятности и принятой гипотезе об условной независимости:

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d) = \sum_{t} \phi_{wt} \theta_{td} \tag{1}$$

где $\phi_{wt} \equiv p(w \mid t)$ и $\theta_{td} \equiv p(t \mid d)$. Нахождение этих распределений по сути является целью тематического моделирования.

Итак, вероятностная тематическая модель (1) описывает, что документы коллекции D генерируются как смесь распределений θ_{td} и ϕ_{wt}. Обучение же тематической модели является обратной задачей, то есть необходимо, имея коллекцию D, найти распределения θ_{td} и ϕ_{wt}. Первое распределение ещё называется распределением “тем-в-документе”

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

(столбцы стохастической матрицы Θ *вероятностей тем в документах* размера $T \times D$). Второе распределение — это распределение "слов-в-теме" (столбцы стохастической матрицы Φ *вероятностей слов в темах* размера $W \times T$).

В работе[13] представлена одна из самых первых, и в то же время одна из самых простых и понятных тематических моделей — модель PLSA (Probabilistic Latent Semantic Analysis), где распределения (1) вычисляются путём максимизации логарифма правдоподобия коллекции с линейными ограничениями. Авторами[14] предлагается подход к обучению тематических моделей под названием ARTM (Additive Regularization of Topic Models) который позволяет реализовать в тематических моделях различные свойства: регуляризаторы позволяют сократить возможное множество решений задачи матричного разложения до тех решений, которые удовлетворяют определенным условиям. Таким образом, с одной стороны, регуляризация используется для получения решения с желаемыми свойствами, но регуляризация также может способствовать повышению устойчивости тематических моделей. По сути же подход аддитивной регуляризации заключается в наложении дополнительных ограничений на получаемые модели путем введения в оптимизируемый функционал дополнительного регуляризационного члена $R(\Phi, \Theta)$:

$$\underbrace{\sum_{d \in D} \sum_{w \in d} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td}}_{\mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta} \qquad (2)$$

В **разделе 1.2** говорится о том, что с ростом популярности безналичных способов оплаты расходов банки стали накапливать огромное количество данных о клиентских операциях. Далее же в разделе

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

[14] *Vorontsov K.* [и др.]. Additive regularization of topic models // Machine Learning. 2015. Т. 101, № 1—3. С. 303—323; *Vorontsov K.* [и др.]. 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; Bigartm: Open source library for regularized multimodal topic modeling of large collections / K. Vorontsov [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; TopicNet: Making Additive Regularisation for Topic Modelling Accessible / V. Bulatov [и др.] // Proceedings of The 12th Language Resources and Evaluation Conference. 2020. С. 6745—6752.

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 представления) 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.

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 их датой, MCC-кодом и потраченной суммой. Таким образом, можно было бы читать клиентов как книгу (см. рис. 1), применяя тематическую модель к их истории транзакций. Результатом стало бы латентное *векторное пространство*, представляющее типы потребления, полученные из статистики данных о транзакциях. *Темы* при этом обеспечивают представление любого клиента в виде подмножества типов потребления, описывающих его *интерпретируемым* образом.

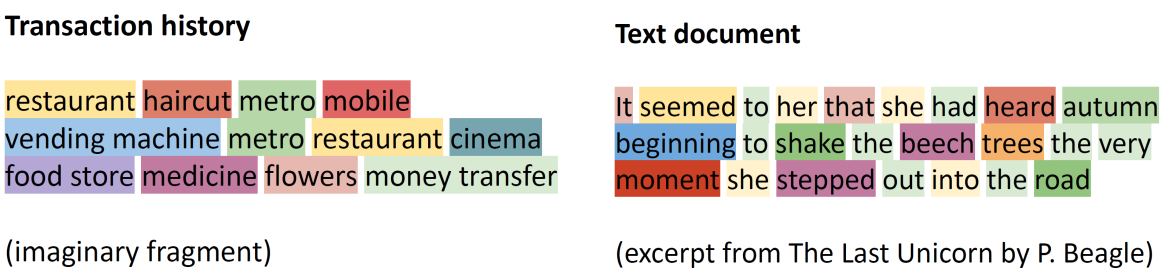

Рис. 1 — transactions. История транзакций — как текстовый документ! Поэтому можно использовать тематическое моделирование для анализа транзакций.

Если рассматривать MCC-код транзакции как токен, то частотой токена можно считать сумму, потраченную на этот MCC-код в документе (истории транзакций клиента).

Проверяется способность предлагаемого подхода создавать интерпретируемые векторные представления клиентов банка на том же уровне. Параллельно проверяется влияние на результат тематического моделирования предварительной обработки данных. Для обучения тематической модели необходимо было определить такие гиперпараметры, как: количество тем, количество шагов EM-алгоритма, коэффициенты регуляризации. Поиск гиперпараметров определялся по значению когерентность модели, так как известно, что он коррелирует с интерпретируемостью [6]. Для сравнения различных тематических моделей значения основных гиперпараметров были зафиксированы и использовались одинаковые для всех моделей: количество шагов EM-алгоритма и количество тем. Коэффициенты же регуляризации настраивались в соответствии с выбранной метрикой. В результате исследования было обнаружено, что наилучшие модели получаются при количестве тем около 30. В таблице 1 приведено описание темы, связанной с отпусками. При этом MCC-коды заменены интерпретируемыми группами таковых (часть предобработки даных).

В **разделе 1.3** описана библиотека для тематического моделирования `TopicNet`. 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 согласован-

Таблица 1 — Тема про отпуска.

Expenses group	**Probability**
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
Gender	**Probability**
Male	0.393
Female	0.607
Age group	**Probability**
17-23	0.138
24-35	0.442
36-54	0.376
55+	0.043

ности темы. Вычислительные эксперименты проводятся на коллекции научно-популярного контента “ПостНаука”.

Традиционные метрики когерентности состоят из двух “частей”: во-первых, они используют информацию из распределения $p(w \mid t)$; во-вторых, они извлекают из текста статистику совстречаемостей слов. Идея автоматических мер когерентности заключается в том, чтобы выяснить, как часто определенные слова появляются вместе в одном контекстном окне, и сравнить это число с частотой, характерной для случайной совстречаемости. Тема считается когерентной, если её слова расположены в тексте группами, а не случайно.

Это напоминает лингвистический феномен текстовой связности[15]: предложения текстов на естественном языке связаны друг с другом с помощью синтаксических и лексических средств, таких как повторы слов, синонимы, гипонимы и другие.

В разделе озвучивается предположение, что тексты на естественном языке разделены на связные сегменты, которые содержат лишь небольшое количество латентных тем. Согласно этому предположению,

[15] *Halliday M. A. K.* [и др.]. Cohesion in english. Routledge, 2014.

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

Отмечается, что частая совстречаемость топовых (самых частых) слов темы является лишь косвенным признаком того, что тема представлена в документах коллекции в виде целостных текстовых фрагментов. Поэтому интерпретируемость темы должна оцениваться не только по согласованности распределения в тексте топовых слов, но и по согласованности распределения всех слов темы в тексте (образуют ли они текстовые сегменты). И можно получить автоматическую меру интерпретируемости модели, исследуя степень согласия распределения её слов с озвученной гипотезой о сегментной структуре.

Представлено несколько автоматических мер когерентности, отличных от традиционных подходов, основанных на совстречаемостях самых частых 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 можно, например, вычислить l_2-расстояние между векторами:

$$\text{SemantiC}_{L_2} = -\langle [\rho(w_i, w_j) \le \text{window}] \|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2 \rangle \tag{3}$$

где $\rho(w_i, w_j)$ — расстояние между словами по тексту (количество других слов между ними), window — окно слов, в котором w_i и w_j считаются близкими по текстовому расстоянию. Знак минус в формуле означает, что когерентность будет тем выше, чем ближе векторы слов. В дополнение к евклидову расстоянию может использоваться мера косинусного сходства:

$$\text{SemantiC}_{Cos} = +\langle [\rho(w_i, w_j) \le \text{window}] \cos(\boldsymbol{w}_i, \boldsymbol{w}_j) \rangle \tag{4}$$

Третий предлагаемый способ оценки семантической близости слов по теме t заключается в вычислении дисперсии между компонен-

тами векторов, соответствующими этой теме:

$$\text{SemantiC}_{Var}\big|_t = -\,\text{Var}\big(\boldsymbol{w}_i(t), \boldsymbol{w}_{i+1}(t), \ldots, \boldsymbol{w}_{i+\text{window}}(t)\big) \tag{5}$$

Перед вычислениями все векторы были умножены на 1000, чтобы увеличить значение результата для когерентности.

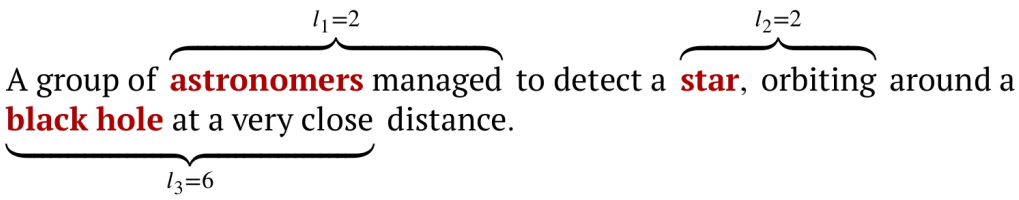

Рис. 2 — Пример, иллюстрирующий идею когерентности TopLen. Пока встречаются слова интересующей темы t, они считаются. Если встречается какое-то не связанное с темой t слово, то оно дает отрицательный штраф. Когда абсолютное значение общего накопленного штрафа оказывается достаточно большим (больше определённого порога), процесс останавливается, и количество подсчитанных к данному моменту слов даёт одно значение длины темы. Далее по всем таким значениям длин можно будет сказать, чему равна "средняя длина темы в тексте": чем она больше, тем лучше распределение темы согласуется с гипотезой о сегментной тематической структуре текста.

Другой метод — *TopLen* (Topic Length) — рассчитывает среднюю продолжительность темы t в тексте, для каждого слова вычисляя скор — разницу между компонентой вектора $\boldsymbol{w}(t)$, соответствующей теме t, и максимальной компонентой среди остальных тем (6). Неотрицательный параметр threshold сглаживает эффект, когда при подсчёте длины темы t когерентность TopLen встречает слова не из темы t: процесс подсчета продолжается до тех пор, пока порог (который в работе был выбран равным 0.01) плюс сумма оценок неотрицательны (для примера см. рис. 2).

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 $\leftarrow \max\left\{ n \ge 0 \,:\, \text{threshold} + \sum_{j=i}^{i+n} \text{score}(w_j, t) \ge 0 \right\}$

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

(6)

Последняя из предлагаемых когерентностей — *FoCon* (Focus Consistency) — оценивает, насколько сильно отличаются соседние слова в тексте, суммируя пары разностей между соответствующими компонентами векторов $p(t \mid w)$ (пара компонент — максимальные компоненты векторов рассматриваемых соседних слов). Знак минус выполняет ту же роль, что с SemantiC: когерентность выше, если слова различаются меньше.

$$\begin{cases} \text{FoCon} = -\sum\limits_{d\in D} \sum\limits_{\substack{w_i, w_j \in W_d \\ j-i=1}} |\boldsymbol{w}_i(t) - \boldsymbol{w}_j(t)| + |\boldsymbol{w}_i(\tau) - \boldsymbol{w}_j(\tau)| \\ t = \arg\max\limits_s \boldsymbol{w}_i(s),\ \tau = \arg\max\limits_s \boldsymbol{w}_j(s) \end{cases} \tag{7}$$

Основной датасет для исследований — это корпус научно-популярных статей, опубликованных в “ПостНауке“ (популярном российском научном интернет-журнале). Исследуется тематическую модель, состоящая из 19 предметных тем и одной фоновой темы (см. таблицу 2).

Оценка интерпретируемости темы — трудоёмкий процесс. Преимуществом когерентностей, основанных на топ-словах, является их

Таблица 2 — Темы “ПостНауки“, 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)

способность сводить темы модели к небольшому обозримому человеком списку слов. Но даже в этом случае получение человеческих оценок о большом количестве тем является сложной задачей.

Предлагается способ обойти эту сложновыполнимую процедуру: вместо того чтобы просить экспертов разметить сырые данные — будет сгенерирован полусинтетический датасет с известной разметкой. Тут помощь оказывает структура датасета “ПостНауки”. Темы статей достаточно общие и разнообразные, при этом большинство документов *монотематические*: т. е. такие, в которых каждое слово может быть отнесено либо к одной из предметных тем (одна и та же предметная темы для большинства слов монотематического документа), либо к фоновой теме. Эти монотематические документы можно использовать для создания полусинтетического датасета. Идея заключается в том, чтобы “разрезать” реальные документы на более мелкие монотематические сегменты, а затем “сшить” их вместе в случайном порядке. Предназначение полусинтетического набора данных — служить в качестве “золотого стандарта”, “ground truth”, по которому можно оценивать тематические модели.

Процедура создания полусинтетического датасета гарантирует, что известны истинные темы для каждого слова. Учитывая эту информацию, можно оценить качество сегментации текста 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 итоговая оценка качества для модели равна сумме этих сумм по всем темам (см. рисунок 3);
- *strict*: для каждой темы t для всех её сегментов подсчитывается количество совпадений темы, предсказанной моделью для слова w в документе d (равной $\arg\max_\tau p(\tau \mid d, w)$), с самой темой t сегмента, к которому принадлежит слово w.

Имея “ground truth”, можно оценить различные меры когерентности. Качество каждой когерентности полагается равным коэффициенту корреляции Спирмена между значениями когерентности и качеством сегментации для ряда тематических моделей.

Для того чтобы получить спектр тематических моделей, дающих различные качества сегментации, было создано несколько различных матриц Φ как взвешенных комбинацию Φ_{good} (известной матрица вероятностей слов в темах для тематической модели датасета “ПостНаука”) и Φ_{bad} (которая была просто набором случайных столбцов, взятых из распределения $\text{Dirichlet}(0.01^{|W|})$):

$$\Phi(\alpha) = \alpha \cdot \Phi_{bad} + (1 - \alpha) \cdot \Phi_{good} \tag{8}$$

Далее, для каждого α 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 посмотреть в таблице 3 и на рисунке 4.

В **разделе 2.3** отмечается, что число тем — один из самых важных параметров тематической модели. Сообществом разработано множество различных процедур для оценки количества тем в текстовом датасете, однако до сих пор не было проведено достаточно полного сравне-

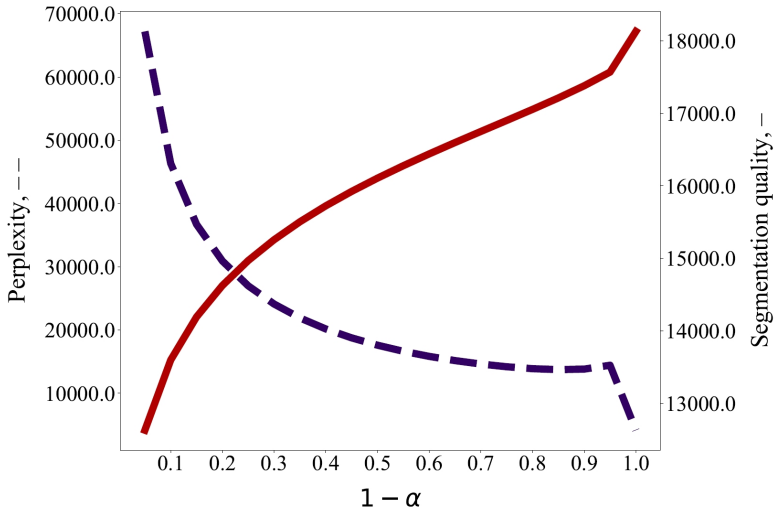

Рис. 3 — Связь между качеством сегментации (soft) и перплексией тематической модели. По оси X отложена доля хороших Ф матриц: единица минус α (степень деградации Ф). Тот факт, что качество сегментации монотонно возрастает при уменьшении перплексии, говорит о том, что предложенный метод оценки качества сегментации действительно может быть использован в качестве меры качества тематических моделей.

ния существующих практик. Представленное в разделе исследование — это попытка восполнить этот пробел, путём изучения 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.

Результаты проведённых экспериментов отражены в таблице 4. Для того чтобы получить осмысленное представление о качестве рассматриваемых внутренних критериев (в плане использования их для определения числа тем), было разработано три показателя. Так, важно

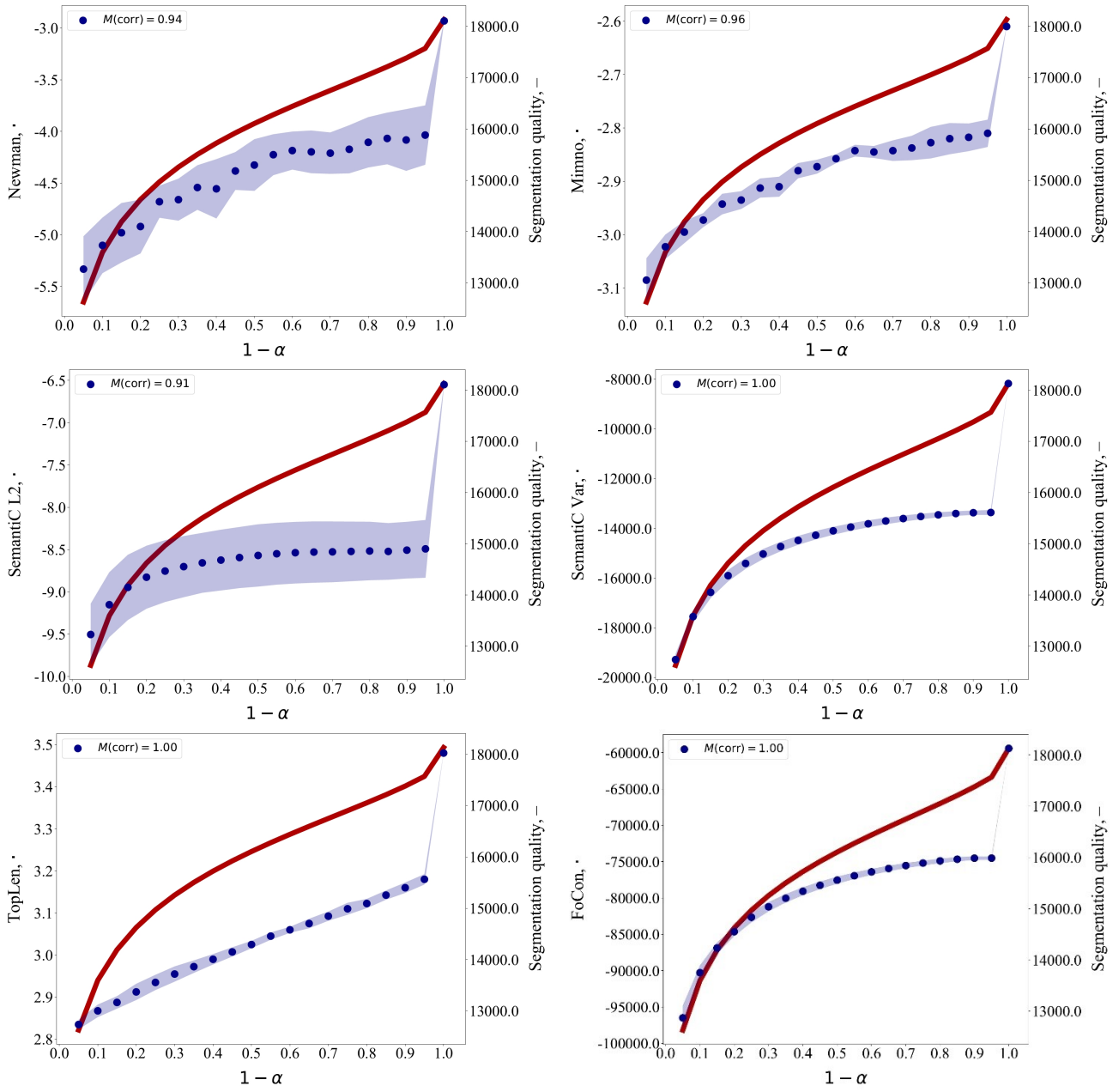

Рис. 4 — Сравнение различных мер когерентности с качеством сегментации в зависимости от α (параметр деградации тематической модели). Значения когерентности на графиках — это значения, усреднённые по серии из четырёх экспериментов с различными матрицами Φ_{bad}.

Таблица 3 — Корреляции Спирмена между когерентностью и качеством сегментации (soft) для полусинтетических датасетов с различными размерами сегментов: 50, 100, 200 и 400 слов — и с 5 темами в каждом 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**

было оценить способность метрики давать оценку количества тем независимую от случайной инициализации модели; “читаемость” графиков изменения метрики от используемого в модели числа тем; и точность метрики в плане предсказания правильного числа тем.

Первый столбец в таблице 4 — это метрика Жаккара, которая рассчитывается следующим образом. Для каждой случайной инициализации было определено оптимальное значение числа тем или диапазон значений (в соответствии с особенностями метрики). Затем было вычислено расстояние Жаккара между пересечением и объединением этих множеств (с исключением тех случаев, когда метрика указывала на границы интервала, в котором проводился поиск числа тем).

Во втором столбце приведена доля того, сколько раз значения метрики были “читаемыми”, то есть попадали в одну из категорий:

- Наличие выраженного минимума или максимума.
- Наличие плато из нескольких близких значений около минимума или максимума.
- Наличие области с чередующимися пиками.

Все остальные типы поведения метрики в зависимости от числа тем можно охарактеризовать либо как не зависящие вообще от количества тем, либо как указывающие (возможно) на некоторое оптимальное значение числа тем вне рассматриваемого диапазона.

Последний показатель для оценки метрик представляет собой среднее значение булевской величины: попадало ли ожидаемое количество тем в диапазон оптимальных значений, показанных метрикой для

Таблица 4 — Сравнение метрик по применимости для определения числа тем (показатели усреднены по нескольким использованным в экспериментах наборам данных). Некоторые из метрик, оценивающих разнообразие тем (diversity), удалены ради компактности таблицы (да и 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

данной модели. Результаты, приведенные в таблице 4, ставят под сомнение представление о том, что количество тем является “чётко определенным свойством” конкретного корпуса (или, по крайней мере, что существующие методы подходят для его определения).

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 "обойти" нестабильность и неполноту тематических моделей, предлагается постепенно (в процессе многократного 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.

Нестабильность и неполнота тематических моделей обусловлены тем, что по природе своей задача тематического моделирования есть по сути задача кластеризации. Для решения этих проблем исследователи часто вынуждены вводить внешние критерии в процесс моделирования, например, оценивать разнообразие найденных тем или требовать, чтобы при разных случайных инициализациях модели получающиеся темы были согласованы. Это может помочь повысить качество финальной тематической модели. Однако это не является полноценным решением проблем неустойчивости и неполноты.

В этой связи кажется возможным организовать процесс поиска тем в коллекции полуавтоматически, используя пользовательские критерии для отбора тем. При таком подходе естественным образом формируется хранилище, или "банк", хороших и плохих примеров тем коллекции, отсюда и название: Topic Bank, или TopicBank. Таким образом, TopicBank — это своего рода *обёртка* над тематическим моделировани-

ем, когда информация о датасете накапливается постепенно в процессе экспериментирования, поиска лучшей модели (5). Подчёркивается, что TopicBank также *не решает* проблем нестабильности и неполноты тематических моделей. Однако TopicBank учитывает эти проблемы и помогает справиться с ними путём многократного обучения тематических моделей. Основные задачи представляемой в процесс тематического моделирования новой сущности TopicBank заключаются в следующем:

- Сбор и хранение интерпретируемых тем (или тем, считающихся хорошими по данному пользователем критерию).
- Использование собранных тем для автоматической оценки качества вновь обученных моделей.

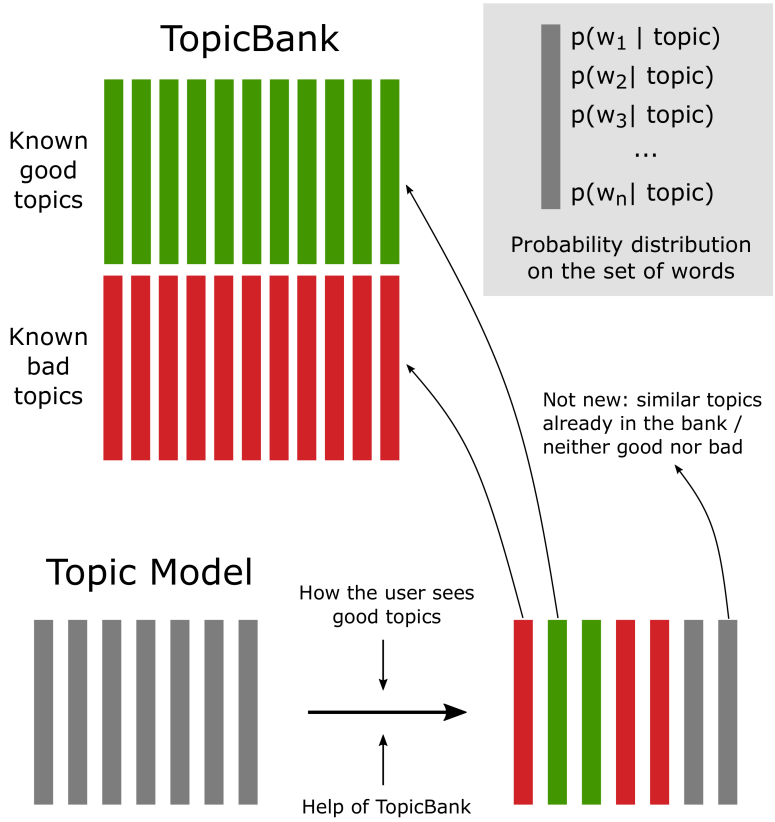

Рис. 5 — Идея Банка тем (TopicBank): в нём накапливаются хорошие и (опционально) плохие темы. Банк нуждается в информации от человека о том, какие темы являются хорошими. Это можно сделать либо автоматически с помощью функции качества, которая сопоставляет тему с булевским значением, означающим, хорошая эта тема или нет; либо с помощью человека, просматривающего и оценивающего темы.

Банк тем создаётся так, чтобы темы, которые он в себе накапливает, в совокупности обладали бы следующими свойствами:

- интерпретируемые (удовлетворяющие используемой оценке качества тем)
- различные (ни одна тема не может быть получена как линейная комбинация других тем)
- составляют решение задачи матричного разложения (дают максимум правдоподобия коллекции)

Набор тем с такими свойствами называется *полным набором тем*.

В **разделе 3.3** представлен метод тематического моделирования с использованием обратной связи от пользователя. Обратная связь заключается в (полу-)автоматическом определении принадлежности 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, которое заключается в выделении новых интерпретируемых тем при сохранении всех уже найденных интерпретируемых тем и уменьшении числа “никаких” тем. Предлагаемое в работе решение (6) основано на применении регуляризаторов сглаживания и декоррелирования в рамках подхода АРТМ. Вычислительный эксперимент проводится на ряде текстовых коллекций естественного языка (7, 8).

отложенные темы

$$L(\Phi,\Theta) + R_{\text{sparse}}(\Phi) + R_{\text{decorr}}(\Phi) + R_{\text{fix}}(\Phi,\widetilde{\Phi}) + R_{\text{decorr}}_{\text{bad}}(\Phi,\widetilde{\Phi}) + R_{\text{decorr}}_{\text{good}}(\Phi,\widetilde{\Phi}) \to \max_{\Phi,\Theta}$$

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

$$R_{\substack{\text{decorr}\\ \text{bad/good}}}(\Phi,\widetilde{\Phi})|_{\tau>0} = -\tau \sum_{t\in T\setminus T_+}\sum_{s\in T_-/T_+}\sum_{w\in W} \phi_{wt}\widetilde{\phi}_{ws} \to \max_{\Phi}$$

Рис. 6 — Максимизация регуляризованного логарифма правдоподобия при обучении `ITAR` модели.

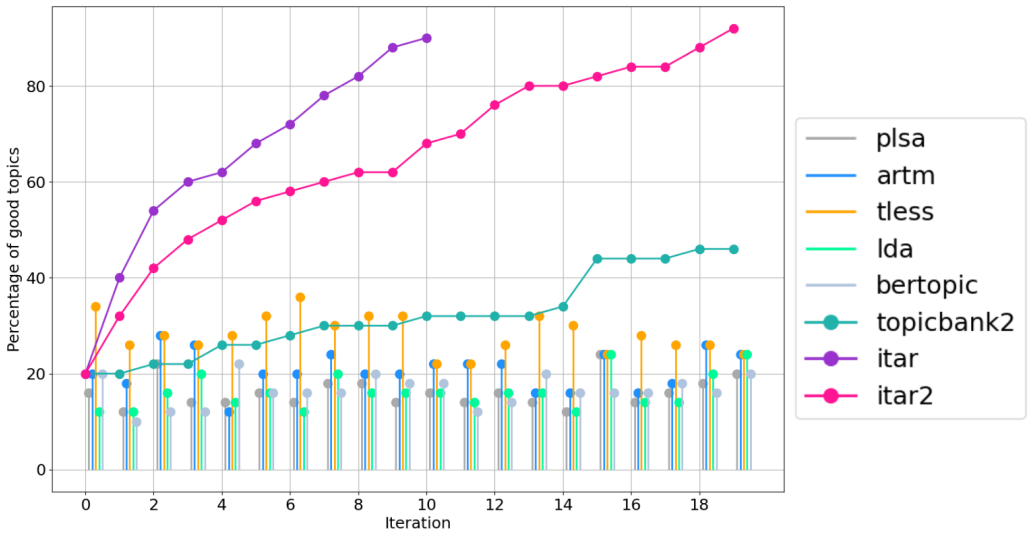

Рис. 7 — Процент хороших тем модели в зависимости от итерации (↑). 20Newsgroups, модели на 50 тем.

Model	PostNauka (20 topics)				RuWiki-Good (50 topics)			
	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Div (↑)	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Div (↑)
plsa	**2,99**	0,74	20	0,60	**3,46**	0,81	26	0,66
artm	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**

Рис. 8 — Некоторые свойства итоговых моделей: перплексия (PPL), средняя когерентность тем (Coh), процент интерпретируемых тем (Good T), различность тем (Div). “ПостНаука”, модели на 20 тем (слева); RuWiki-Good, модели на 50 тем (справа).

В **заключении** приведены основные результаты работы, которые 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 когерентность и алгоритмы 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 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 датасетов естественного языка доказана состоятельность подхода.

Публикации автора по теме диссертации

Список публикаций в изданиях, рекомендованных ВАК

1. *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.
2. 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.
3. 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.
4. 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.

Список публикаций в других изданиях

5. *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.
6. *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.

7. *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.
8. *Bulatov V.*, *Alekseev V.*, *Vorontsov K.* Determination of the Number of Topics Intrinsically: Is It Possible? // International Conference on Analysis of Images, Social Networks and Texts (In Press). — Springer. 2024.
9. *Булатов В. Г.*, *Алексеев В. А.* Determination of the Number of Topics Intrinsically: Is It Possible? // Труды 66-й Всероссийской научной конференции МФТИ (в печати). — 2024.
10. *Горбулев А. И.*, *Алексеев В. А.* Итеративное улучшение аддитивно регуляризованной тематической модели // Труды 66-й Всероссийской научной конференции МФТИ (в печати). — 2024.

Зарегистрированные программы для ЭВМ

11. *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).
12. *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).