

Тематические модели не моделируют темы: эпистемологические замечания и шаги на пути к лучшим практикам

Анна Шадрова 1

1 Университет имени Гумбольдта в Берлине, Германия

Автор, ответственный за переписку: Анна Шадрова, anna.shadrova@hu-berlin.de

Абстрактный

Социальные науки и цифровые гуманитарные науки недавно приняли метод машинного обучения для тематического моделирования для решения исследовательских вопросов в своих областях. Это проблематично по ряду причин, некоторым из которых еще не уделялось должного внимания в ходе дебатов. В этой статье добавляются эпистемологические проблемы, связанные с интерфейсом между моделированием темы и лингвистическими концепциями, а также аргументированное встраивание доказательств, полученных посредством моделирования темы. Делается вывод о том, что тематическое моделирование в его нынешнем состоянии методологической интеграции не отвечает требованиям самостоятельного метода исследования. Он исходит из относительно нереалистичных предположений, недетерминирован, не может быть эффективно проверен на основе разумного числа конкурирующих моделей, не привязан к четко определенному лингвистическому интерфейсу и не моделирует научные темы в смысле тем или содержания. Эти особенности являются внутренними и делают интерпретацию его результатов склонной к апофении (человеческая склонность воспринимать случайные наборы элементов как значимые паттерны) и предвзятости подтверждения (человеческая склонность перцептивно предпочитать паттерны, соответствующие ранее существовавшим предубеждениям). Хотя частичная проверка статистической модели возможна, концептуальная проверка потребует расширенной триангуляции с другими методами и человеческими оценками, а также выяснения того, коррелирует ли статистическая различимость лексического совпадения каким-либо надежным образом с концептуальными темами. [1](#page1)

Ключевые слова

Тематическое моделирование; цифровые гуманитарные науки; извлечение информации для научных исследований

ВВЕДЕНИЕ \_

Тематическое моделирование — это метод машинного обучения, который классифицирует слова по так называемым темам и вычисляет предполагаемую долю таковых для документов и корпусов (наборов текстов). Более поздние реализации также позволяют анализировать тематические корреляции с текстовыми метаданными, такие как изменения во времени или распределения между авторами ( [Блей и Лафферти](#page23) , [2006](#page23) ; [Робертс](#page26) [и другие.](#page26) , [2019](#page26) , [2014](#page27) ). Первоначально он был разработан и продолжает использоваться для поиска информации. и целей классификации текста в прикладных контекстах ( [Bao et al.](#page23) , [2009](#page23) ; [Асунсьон и др.](#page23) , [2010](#page23) ; [Рэймидж и др.](#page26) , [2011](#page26) ; [Ван и Блей](#page27) , [2011](#page27) ; [Чуанг и др.](#page24) , [2013](#page24) ; [Си и др.](#page27) , [2014](#page27) ; [Чжун и др.](#page28) , [2015](#page28) ; [Ван Дер Хофт и др.](#page27) , [2016](#page27) ; [Лю и др.](#page25) , [2016](#page25) ; [Бойд-Грабер и др.](#page23) , [2017](#page23) ; [Кун](#page25) , [2018](#page25) ; [Лю](#page25) [и другие.](#page25) , [2019](#page25) ; [Ребер](#page26) , [2019](#page26) ; [Корфиатис и др.](#page25) , [2019](#page25) и многие другие). Совсем недавно он также набрала обороты в контексте так называемого дистанционного чтения [2](#page1) в цифровых гуманитарных науках и

* Я благодарен Фредерику Клаверу и анонимному рецензенту за их очень конструктивные и полезные комментарии. Все остальные огрехи и ошибки конечно же мои.

2 Дистанционное чтение, в отличие от внимательного чтения, описывает анализ поверхностных паттернов, извлеченных из больших корпусов с использованием вычислительных методов анализа текста. За приложения и критический вклад в

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 1 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

социальные науки, где в настоящее время он все чаще используется для ответа на вопросы предметного исследования, касающиеся распределения содержания в художественном тексте ( [Asgari et al.](#page23) , [2013](#page23) ; [Тангерлини](#page27) [и Леон ард](#page27) , [2013](#page27) ; [Джокеры и Мимно](#page25) , [2013](#page25) ; [Андервуд](#page27) , [2014](#page27) ; [Голдстоун и Андервуд](#page24) , [2014](#page24) ; [Вейтин и Хергет](#page27) , [2017](#page27) ; [Митрофанова и Седова](#page26) , [2017](#page26) ; [Шох¨](#page27) , [2017](#page27) ; [Эрлин](#page24) , [2018](#page24) ; [Наварро-Колорадо](#page26) , [2018](#page26) ; [Джейкобс](#page25) , [2018](#page25) ; [Зиг](#page27) , [2019](#page27) ; [Даллоф¨ и Берглунд](#page24) , [2019](#page24) ; [Лю и Джин](#page25) , [2020](#page25) ), судебные решения ( [Livermore et al.](#page26) , [2016](#page26) ; [Панагис и др.](#page26) , [2016](#page26) ; [Картер и др.](#page23) , [2016](#page23) ; [Закон](#page25) , [2016](#page25) ; [Ван и др.](#page27) , [2017](#page27) ; [рис](#page26) , [2017](#page26) ; [молодой](#page28) , [2019](#page28) ; [Лампах и Дьевре](#page25) , [2018](#page25) ), политические и юридические дебаты ( [молодые](#page27) , [2012](#page27) ; [Грин и Кросс](#page24) , [2016](#page24) ; [Стерлинг и др.](#page27) , [2019](#page27) ; [Гриммер](#page24) , [2010](#page24) ), освещение в СМИ ( [DiMaggio et al.](#page24) , [2013](#page24) ; [Берталан и Руис](#page23) , [2019](#page23) ; [Люстра и др.](#page23) , [2018](#page23) ; [Ян](#page27) [и другие.](#page27) , [2011](#page27) ), или академических журналов ( [Chen et al.](#page23) , [2020](#page23) ; [Линдштедт](#page25) , [2019](#page25) ; [Ван и др.](#page27) , [2017](#page27) ). [3](#page2)

Это расширение является эпистемологически проблематичным по ряду причин. Хотя вопросы, связанные с субъективностью маркировки тем и отсутствием математической проверки тематических моделей, обсуждались с самого начала их применения ( [Chang et al.](#page23) , [2009](#page23) ; [Рамирес и др.](#page26) , [2012](#page26) ; [Арора и др.](#page23) , [2013](#page23) ; [Чуанг и др.](#page24) , [2013](#page24) , [2015](#page24) ), некоторые более глубокие проблемы до сих пор остаются в значительной степени нерешенными, хотя была проведена некоторая работа, показывающая практические проблемы (чрезмерно смешанные, неинтерпретируемые, неполные темы и т. д.), возникающие при недостаточной подготовке данных ( [Schmidt](#page27) , [2012](#page27) ; [Бойд-Грабер и др .](#page23) , [2014](#page23) ). Аргумент в этой статье касается структурных и концептуальных свойств тематического моделирования и его лингвистических основ, в частности

1. базовая модель латентного, неизменного и однозначно определяемого семантического пространства, лежащего в основе моделирования тем, что сильно расходится с теорией лингвистических исследований и исследований текста. Темы, темы или категории содержания не являются четко определенными лингвистическими понятиями, их границы, как правило, размыты и динамичны, и они могут быть построены только в ходе глубокого аналитического процесса, а не извлечены в качестве базовой информации, поскольку они в значительной степени получены из неявное значение, а не совпадение слов как таковое;
2. тот факт, что статистические распределения слов, могут ли они быть более или менее отчетливыми, не гарантируют различения тем с лингвистической точки зрения, а это означает, что даже статистически четко определенные и надежные модели не могут гарантировать достоверность концепции;
3. эвристический, недетерминистический характер тематического моделирования, который противоречит воспринимаемой объективности его результатов, в то время как большое количество моделей для любого данного корпуса эффективно избегает проверки между репрезентативной пропорцией возможных моделей;
4. оценка распространенности тем в документах с помощью статистических средств, что является чрезмерным упрощением сложностей количественного определения значения и дополнительно усложняется тексто-лингвистическими особенностями, такими как длина текста, сложность или жанр;
5. противоречие между фактически узким доказательным охватом результатов тематического моделирования, т. е. тем фактом, что они потребуют больших методологических усилий и контекстуализации, чтобы интегрироваться в научный аргумент в какой-то убедительной форме, с обычной практикой включения результатов в качестве «исследовательского исследования». » без дополнительной интеграции.

Пункты а)-г) в равной степени относятся к тематическому моделированию в прикладных контекстах, для которых оно изначально разрабатывалось. Однако в этих областях результаты подлежат немедленной внешней проверке, например, посредством изменения поведения клиентов в рекомендательной системе или фактического синтеза гипотетически возможных молекул, ср. [Ван Дер Хофт и др.](#page27) ( [2016](#page27) ). В отличие от этого, сами темы и категории контента являются частью конечного результата текстового исследования в академическом контексте. Они обычно не могут быть проверены или подтверждены извне, поэтому весь вес их

дебаты, см. [Моретти](#page26) ( [2013 г.](#page26) ); [аскари](#page23) ( [2014 г.](#page23) ); [Янике¨ и др.](#page25) ( [2015](#page25) ); [Андервуд](#page27) ( [2017](#page27) ); [Друкер](#page24) ( [2017](#page24) ).

3 Дополнительные ссылки и сопоставление технологий с исследовательскими вопросами можно найти в [Макфарланд и др.](#page26) ( [2013](#page26) ). [Бойд-Грабер и др.](#page23) ( [2017](#page23) ) составили 150-страничный подробный обзор приложений и их технического контекста.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 2 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

приемлемость их аргументированного включения в теоретические и/или эмпирические модели, на которых могут основываться будущие исследования. Тематическое моделирование не предоставляет моделей такого рода.

Цель этой статьи состоит в том, чтобы обеспечить систематический обзор упомянутых вопросов с точки зрения исследований в области количественной корпусной лингвистики, качественных текстовых исследований и эпистемологии, чтобы разграничить сферы применения, а также эпистемологические ограничения.

Он начинается с краткого обсуждения того, что составляет тему, тему или содержание в различных субдисциплинах текстовых областей исследований, таких как лингвистика, литературоведение и социальные науки. Затем в нем представлены аргументы в пользу того, почему тематическое моделирование не является автоматизацией процесса построения категорий, сначала представлено концептуальное и в значительной степени нетехническое введение в тематическое моделирование (раздел [III](#page5) ); а затем более тщательное обсуждение типа данных, полученных в результате тематического моделирования, и того, как они а) противоречат лингвистическим концепциям и б) трудно интегрируются в научные аргументы (раздел [IV](#page7) ).

Это актуально в контексте того, что в последние годы области цифровых гуманитарных наук и текстовых исследований переходят от более качественной к более количественной работе, где методологическая сложность не поспевает за развитием технологий. Поскольку для дальнейшего развития необходимы подходы к более структурированной оценке, систематизации и валидации методов, некоторые предложения на этот счет сделаны в разделе [В](#page18) .

II ЧТО ТАКОЕ ТЕМЫ И КАК ИХ МОДЕЛИРОВАТЬ?

Будучи компетентными взрослыми, говорящими на каком-либо языке, мы часто хорошо понимаем, о чем текст, даже если это очень короткий текст, например, одно предложение, или синтаксически нерекомендуемый, как газетный заголовок («Шеф Трампа сотрудников Meadows подверглись критике за преодоление пандемии и других кризисов», The Washington Post онлайн, 27 октября 2020 г. [4](#page3) ). Мы выводим эту идею, т. е. предмет, тему, тему, предметность или содержание текста, частично из содержащихся в нем слов, а частично из нашего таксономического и онтологического понимания мира. Например, в заголовке выше мы можем вывести определенные аспекты значения из таких слов, как начальник штаба, пандемия или кризисы. Но для того, чтобы правильно экстраполировать тему, нам также нужен контекст текущей мировой политики (например, кто такой Трамп) и понимание контекстуализированного значения под огнем (в профессиональной обстановке в отличие от военной осады) или обращение (пандемия в отличие от животного зоопарка). Однако, даже с этими знаниями, определение соответствующей метки темы по-прежнему не является тривиальной задачей. В зависимости от лежащей в основе онтологии и цели классификации она может варьироваться от политики США до определенного периода жизни Марка Медоуза и мировых событий, вызывающих профессиональные кризисы.

Есть по крайней мере четыре в значительной степени отдельных направления научных исследований, которые работают с идеей темы, темы или категории содержания: лингвистика, литературоведение, информатика и социальные науки. Их соответствующие представления о том, что представляет собой тема, прагматически и теоретически разнообразны - в лингвистике понятие темы в значительной степени уместно в контексте информационной структуры и обозначает единицу информации в предложении или дискурсе, к которому относится предложение, или референт, о котором говорится что-то важное; в литературоведении тема представляет собой неявную историю или более широкую категорию значения, полученную из слов произведения и вплетенную в исторический или социальный контекст; в социальных науках тема или категория содержания — это неявное или явное значение, полученное от слов, принадлежащих к аналогичным семантическим категориям. Информатика также использует термины «субъект» или «предметность», описывая всеобъемлющие ярлыки классификации, которые облегчают эффективный поиск в библиотеках или поисковых системах.

4 [https://www.washingtonpost.com/politics/meadows-trump-coronavirus/2020/1](https://www.washingtonpost.com/politics/meadows-trump-coronavirus/2020/10/26/475f03d2-122f-11eb-82af-864652063d61_story.html) [0/26/475f03d2-12 2f-11eb-82af-864652063d61 story.html](https://www.washingtonpost.com/politics/meadows-trump-coronavirus/2020/10/26/475f03d2-122f-11eb-82af-864652063d61_story.html)

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 3 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

Например, в предложении «Король умер, а затем умерла королева» лингвист может сказать, что тема предложения — «Король» в первом предложении и «Королева» во втором; литературовед мог бы сказать, что тема — утрата или горе; и социолог может сказать, что речь идет о смерти, королевской власти или истории определенной страны в определенное время. Набор потенциальных категорий зависит от контекстуализации исследователем в рамках (потенциальной) таксономии каждого предмета.

Однако объем такой контекстуализации или категоризации не определен однозначно, и сама концепция темы, темы и т. Д. В дальнейшем обсуждается в разных дисциплинах. По лингвистике [Гутсос](#page24) ( [1997](#page24) , 1) отмечает, что

«[В] обзоре работы по теме и теме бросается в глаза почти полное отсутствие консенсуса среди лингвистов относительно природы, определяющих характеристик и сферы применения используемых понятий. Единственное мнение, которое, по-видимому, широко распространено, состоит в том, что в этом районе (...) полно проблем. Одной из причин этого должно быть постоянное распространение литературы. Трудно просмотреть даже существующие обзоры (...). Этот постоянный интерес, безусловно, свидетельствует о важности этого района. Тем не менее весьма сомнительно, чтобы используемые термины вышли из области интуиции и приобрели точное значение».

Автор продолжает называть не менее 15 определений, охватывающих диапазон от «узнаваемой единицы (идей, слов и т. д.)» до «(...) объединяющей нити, проходящей через текст в целом». [Шлобински и Шютце¨-Коберн](#page27) ( [1992](#page27) , 114) и [Келер](#page25) ( [2004](#page25) , 238) даже предполагают, что тема эпифеномен, возникающий из текстовой когерентности и не имеющий собственной лингвистической реальности. Подобные дебаты с поразительно схожими результатами произошли в тематике, поддисциплине литературоведения — то, что делает тему, мотив, предмет и т. Д., Не определяется однозначно в этой области, а скорее привлекает расходящиеся и часто противоречащие концепции. ( [Ван Пир](#page27) ( [2002](#page27) ), [Петтерс- сон](#page26) ( [2002](#page26) , 238), [Риммон- Кенан](#page26) ( [1995](#page26) )) или даже подвергается сомнению в отношении ценности его вклад в первую очередь ( [Sollors](#page27) , [2002](#page27) ).

Такое обилие подходов связано с тем, что темы (темы, темы и т. д.) не являются лингвистической категорией как таковой — многое из того, о чем идет речь в тематическом или контент-анализе, не включено в фактические слова или их совместное употребление. В качестве [Криппендорф](#page25) ( [2004a](#page25) , 10) размышляет о контент-анализе в социальных науках:

«Контент-аналитики должны предсказывать или делать выводы о явлениях, которые они не могут наблюдать напрямую. Неспособность наблюдать интересующие явления, как правило, является основной мотивацией для использования контент-анализа. Имеются ли у анализируемого источника причины скрывать то, что аналитик желает знать (...), или интересующие явления недоступны в принципе (например, отношение или состояние ума человека, или исторические события) или просто труднодоступны для понимания? оценивая иначе (например, что определенные аудитории СМИ могли бы узнать из просмотра телевизора), аналитик ищет ответы на вопросы, выходящие за рамки текста».

Таким образом, в большинстве определений темы являются не прямыми производными от данных, а интерпретацией аспектов данных в контексте, где контекст обеспечивается знанием ученого предмета, таксономий и дискуссий в этой области. Таким образом, тематический или контент-анализ в научном смысле представляет собой не процесс извлечения существующих сущностей, а процесс построения категорий. [5](#page4) Таким образом, на вопрос о том, включены ли темы в какие-либо данные, можно ответить только в

5 Даже если данные сопоставляются с существующими категориями в поле, это все равно требует построения тех же категорий из данных.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 4 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

процессно-ориентированный способ. Если определенный тип функции, т. е. метод, принимает определенный вход и дает определенный результат, то относительно функции можно сказать, что результат оправдан. Поскольку функция имеет решающее значение для достоверности результатов, каждая дисциплина создала обширную методологию качественного исследования вопросов такого рода. [6](#page5) Такие подходы, как герменевтика или обоснованная теория, обычно не могут считаться столь точными, как количественные подходы, но они делают сильный акцент на роли контекстуализации посредством существующих таксономий, влияния субъективности исследователя и когнитивных искажений, таких как предвзятость подтверждения (тенденция человеческого разума воспринимать образы, которые соответствуют существующим ожиданиям) или апофения (считывание образов в случайные данные, такие как облачные животные, ср. [Диксон](#page24) ( [2012](#page24) )). Оставшаяся часть статьи направлена на то, чтобы показать, что тематическое моделирование не является автоматизацией таких видов анализа, и что те же самые оговорки сохраняются и фактически усиливаются за счет массивного и последовательного сокращения информации при тематическом моделировании.

III ЧТО ТАКОЕ ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ?

Тематическое моделирование — это метод машинного обучения без присмотра, который классифицирует слова в документах на совпадающие наборы слов, так называемые темы. Неконтролируемый означает, что классификация основана только на самом содержании корпуса; никакая дополнительная информация, такая как золотой стандарт маркированного контента, не предоставляется. Однако обычно требуется некоторая предварительная обработка, чтобы избежать тривиальных или случайных тем. Это включает в себя фильтрацию часто встречающихся слов, таких как артикли или предлоги, и перевод всех слов в нижний регистр (это может иметь последствия для контента в таких языках, как немецкий, где существительные пишутся с большой буквы). Большинство реализаций предлагают возможность указать другие слова, которые не следует учитывать в модели, например слова, которые могут исказить результаты из-за двусмысленности или сверхспецифичности.

Статистическая модель, извлеченная из корпуса текстов, строится на предположении, что в мире существует ограниченное и фиксированное количество повторяющихся тем, т. е. «вещей, о которых можно говорить» [7 .](#page5) и что это отражено в корпусах через частое совпадение тематических слов в отдельных текстах. Например, некоторые слова, которые могут часто встречаться вместе, — это здоровье, врач и больница, что предполагает тему медицины или здравоохранения. Каждый документ моделируется как стохастически распределенная смесь всего нескольких тем. Например, более вероятно, что тема медицины связана с темой прав человека, чем тема настольных игр, хотя технически возможно и то, и другое; и ни один документ обычно не содержит все темы, существующие в модели (что делает вероятность некоторых тем близкой к нулю в некоторых документах). Более формально, лежащая в основе модель представляет собой стохастическую модель со смешанным членством скрытых семантических пространств.

Предполагается, что темы содержат слова с определенной и неизменной вероятностью (слова, не относящиеся к теме, имеют вероятность, близкую к нулю). Таким образом, темы в тематическом моделировании представляют собой распределения слов и вероятностно распределены по документам. Таким образом, тематическая модель представляет собой распределение распределений (слов).

Эти распределения можно вычислить с помощью ряда алгоритмов: латентного семантического индексирования (LSI) или латентного семантического анализа (LSA) ( [Deerwester et al.](#page24) , [1990](#page24) ), Неотрицательная матрица

* Анализ содержания: [Криппендорф](#page25) ( [2004а](#page25) ); [Макнамара и др.](#page26) ( [2005 г.](#page26) ); [Кимберли](#page25) ( [2011 г.](#page25) ); [Майринг](#page26) ( [2004 г.](#page26) ); [Гранехейм и Лундман](#page24) ( [2004 г.](#page24) ); [Криппендорф](#page25) ( [2004б](#page25) ); Тематика в литературоведении: [Хасан](#page25) ( [1967 г.](#page25) ); [Риммон-Кенан](#page26) ( [1995 г.](#page26) ); [Петтерссон](#page26) ( [2002](#page26) ).

7 «Общим для всех неконтролируемых тематических моделей является идея о том, что язык организован скрытыми измерениями, о которых субъекты могут даже не знать. Применительно к повседневной речи базовые (неконтролируемые) тематические модели обычно определяют области обсуждения, такие как вождение и знаки остановки, и отличают их, скажем, от свиданий». [Макфарланд](#page26) [и другие.](#page26) ( [2013](#page26) , 5). Интересно, что автор обращается к бытовым предметам, относительно четко определенным и различимым, в отличие от нечетких и градиентных категорий научного исследования.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 5 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

Факторизация (NMF) ( [Куанг и др.](#page25) , [2015](#page25) ), и наиболее широко используемое в настоящее время латентное распределение Дирихле (LDA) ( [Blei et al.](#page23) , [2003](#page23) ) и их модификации. [8](#page6) Не вдаваясь в подробности, грубое описание необходимо для того, чтобы подготовить следующую аргументацию. [9](#page6)

В тематическом моделировании используются два различных набора алгоритмов. Один основан на уменьшении размерности векторных пространств (скрытое семантическое индексирование/анализ (LSI/LSA), неотрицательная матричная факторизация (NMF)). Они детерминированы, что означает, что они всегда будут давать одну и ту же модель для одних и тех же данных, независимо от того, как часто запускается алгоритм или какое слово он использует в качестве отправной точки. Эти алгоритмы используют различные типы уменьшения размерности, но оба идентифицируют темы на основе матрицы совпадения слов (также называемой матрицей терминов документа). Это матрица, содержащая частоту совпадения каждого слова со всеми другими словами для каждого документа, которая затем разлагается на вектор слов, относящихся к каждой теме, и другой вектор тем, относящихся к каждому документу. По этим векторам вычисляется сходство, обычно через косинусное расстояние. Думайте о векторах как о стрелках в двумерном пространстве. Если две стрелки указывают в разных направлениях, их корни встретятся под углом, который можно использовать для сравнения: чем более похоже направление, тем меньше угол с максимальной разницей в 180°. Векторы в тематическом моделировании не являются двумерными, но применяется тот же принцип.

С помощью этих алгоритмов количество тем выводится из матрицы, и одна и та же матрица всегда дает одни и те же темы как по составу, так и по распределению. LSI/LSA и NMF были первыми типами тематических моделей, но в значительной степени были вытеснены более продвинутым алгоритмом скрытого распределения Дирихле (LDA). В LDA количество тем не выводится из матрицы терминов документа, а выбирается вручную. Основываясь на желаемом количестве тем, LDA сначала случайным образом назначает слова темам, а затем, стремясь максимизировать функцию различимости, итеративно меняет местами слова между темами. Окончательная модель — это модель, которая, в зависимости от точки инициализации, дает наиболее отчетливое или разделимое распределение тем. [10](#page6)

Это согласуется с лежащим в основе (хотя и проблематичным) предположением о том, что темы дискретны и различны, то есть между темами нет преемственности или градиента; что они отмечены определенными словами; и что темы различаются в зависимости от документа - хотя некоторые очень частые темы могут встречаться в большинстве документов с разной скоростью, в целом предполагается, что темы работают как классификатор и не распределяются чрезмерно равномерно.

Поскольку функция максимизируется до k выбранных кластеров итеративно, начиная с точки инициализации, LDA не является детерминированным. В зависимости от точки инициализации (первоначальное случайное назначение слов темам) результаты могут различаться и действительно различаются. Функция может быть решена для любого количества тем, точно так же, как можно разделить вещи, которыми владеешь, на любые предметы.

* Также существуют вероятностные версии LSA/LSI и NMF. Однако оказывается, что они математически эквивалентны и представляют собой только частные случаи LDA, см. [де Паулу Фалейрос и де Андраде Лопес](#page24) ( [2016](#page24) ); [Джиролами](#page24) [и Кабань](#page24) ( [2003 г.](#page24) ); [Дин и др.](#page24) ( [2006](#page24) ).

9 Техническое, но очень доступное сравнение основных алгоритмов можно найти здесь: [https://medium.com/@souravboss.bose/comprehensive-topic-modelling-with-nm](https://medium.com/@souravboss.bose/comprehensive-topic-modelling-with-nmf-lsa-plsa-lda-lda2vec-part-1-20002a8e03ae) [f-lsa-plsa-lda-lda2vec-part-1-20002a8e03ae](https://medium.com/@souravboss.bose/comprehensive-topic-modelling-with-nmf-lsa-plsa-lda-lda2vec-part-1-20002a8e03ae) . Для сравнения LDA и NMF см. [Чен](#page24) [и другие.](#page24) ( [2019](#page24) ). Обзор LDA и его приложений с некоторыми техническими подробностями см. [Блей и др.](#page23) ( [2010](#page23) ) и [Блей](#page23) ( [2012](#page23) ). О некоторых модификациях LDA см. [Ю и др.](#page28) ( [2017](#page28) ).

10 Имя Дирихле в LDA относится к распределению Дирихле, названному в честь математика XIX века Йохана Петера Густава Лежена Дирихле. Распределение Дирихле — это многомерное распределение плотности вероятности, используемое в байесовской статистике. Максимально кластеризованное распределение, т. е. распределение, при котором высокая вероятность присваивается нескольким членам в небольшом диапазоне, аппроксимируется в LDA с помощью байесовского моделирования смешанного членства. Детали не имеют отношения к аргументу в этой статье и не будут далее обсуждаться.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 6 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

количество (достаточно больших) выдвижных ящиков. Очевидно, что качество или полезность разбиения может пострадать из-за слишком большого или слишком малого количества ящиков, и то же самое касается тем в LDA: нет никакого способа заранее определить идеальное количество тем. Попытки приблизить идеальное число включают комбинацию с другими статистическими подходами, такими как анализ основных компонентов (PCA), который определяет статистически лучшее количество кластеров, которое затем можно использовать как k для количества тем; или путем добавления неотрицательной матричной факторизации в предварительную обработку (последнее реализовано в [Арора и др.](#page23) ( [2014](#page23) )). Однако ни анализ основных компонентов, ни NMF не обеспечивают окончательного разделения тем – это очевидно из того факта, что от него отказались в пользу LDA из-за часто неудовлетворительных результатов. В целом LDA дает гораздо более впечатляющие результаты по сравнению со старыми алгоритмами с точки зрения интерпретируемости и удобства использования тем в контекстах поиска информации, но он не может определить общую «лучшую» модель без дополнительных входных данных.

Хотя о технических и математических аспектах различных алгоритмов тематического моделирования можно сказать больше, для аргументации в этой статье достаточно знать, что

темы - это распределения слов,

тематические модели - это распределения распределений (слов) по документам,

тематические модели — это модели со смешанным членством, что означает, что в документе могут быть представлены несколько тем, и слова могут встречаться в нескольких темах,

наиболее часто используемый алгоритм моделирования тем, LDA, является вероятностным и максимизирует различимость тем, т. е. распределяет слова по темам таким образом, что темы наиболее четко отличаются друг от друга,

количество тем в LDA выбирается вручную,

LDA является эвристическим, а не детерминистическим: одни и те же данные могут давать и действительно дают разные модели для одного и того же количества тем, если выбрана другая точка инициализации.

IV КАКИЕ ДОКАЗАТЕЛЬСТВА ПРЕДОСТАВЛЯЮТ ТЕМАТИЧЕСКИЕ МОДЕЛИ?

Есть несколько неотъемлемых аспектов тематического моделирования в его текущих реализациях, которые ограничивают объем доказательств, которые оно представляет:

1. Тематическое моделирование является неполным, эвристическим и не поддается проверке, поэтому оно не обеспечивает уникальную или объективно лучшую модель содержания корпуса даже статистически;
2. Тематическое моделирование основано на массивной и непредсказуемой редукции лингвистической размерности и соответствующих нереалистичных предположениях, поэтому оно не обеспечивает концептуальной достоверности по отношению к лингвистической реальности;
3. Тематическое моделирование не дает надежной количественной оценки значения, его количественные результаты (оценка распространенности темы в документах и по метаданным, таким как год или автор), следовательно, не являются точным или почти точным измерением распределения контента;
4. Тематическое моделирование не позволяет сознательно устанавливать качественные параметры, поэтому его научное применение довольно ограничено;
5. Темы, полученные в результате тематического моделирования, по-прежнему являются конструкциями, а не объективными наблюдениями.

Они будут рассмотрены с методологической и лингвистической точки зрения более подробно в следующих разделах.

4.1 Тематическое моделирование не завершено

Тематическое моделирование может обоснованно включать в анализ только слова определенного частотного спектра. Включение всех лексем в топическую модель, как правило, неэффективно из-за высоких вычислительных затрат. Он также ограничен с точки зрения статистической релевантности: редкое слово, встречающееся в любом количестве из его ограниченного общего числа вхождений рядом с более частыми словами, не будет опрокидываться.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 7 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

общая статистика по сравнению с более высокой силой кластеризации этих частых слов. [11](#page8) Хотя может показаться, что это влияет только на несколько слов в конце частотного диапазона, на самом деле все наоборот. Только очень небольшая часть лексем встречается достаточно часто, чтобы сделать какое-либо разумное статистическое заключение. На самом деле слова распределены таким образом, что только некоторые из них встречаются с высокой частотой (например, в английском языке the, to или is). До половины всех слов в больших корпусах встречаются только один раз (так называемые hapax legomena, или просто гапаксы), другая большая часть — дважды, трижды и т. д. Это часто моделируется как распределение Ципфа или степенное распределение ( [Baayen](#page23) , [2002](#page23) ; [Zipf](#page28) , [1965](#page28) ) и подразумевает, что количество различных слов (лексем), используемых в корпусах, очень велико. За вычислительная эффективность, пакет Rstm ( [Roberts et al.](#page26) , [2019](#page26) ) работает с 10 000 наиболее часто встречающихся лексем. Хотя это может охватывать широкий лексический диапазон, это далеко не полно: например, в корпусе решений Федерального конституционного суда Германии, содержащем 3312 решений за 1951–2017 годы ( [Mollers¨ et al.](#page26) , [2021](#page26) ), остается неиспользованными 48 000 лексем, или 82,8%.

Это ограничение также подразумевает, что тематическое моделирование значительно (фактически, экспоненциально) более неполно по отношению к включенным лексемам в более крупные корпуса — более частые лексемы достигают насыщения в относительно небольших корпусах, в то время как количество гапаксов продолжает расти. Хотя гапаксы нельзя использовать для статистического анализа, методологический вывод о том, что более 80% лексического материала, составляющего корпус, нельзя использовать для оценки распределения его содержания, безусловно, неудобен с научной точки зрения. Даже исключение одного слова из темы может изменить ее интерпретацию — рассмотрим, например, набор ftable, bar, boxg (возможно, помеченный как мебель) по сравнению с тем же набором с дополнительным графиком, который может изменить его на количественные результаты исследования. . Если исключить большинство слов, то разнообразие результирующих тем неизбежно будет ограничено.

Кроме того, неизвестно, что релевантные для темы слова также распределяются по степенному закону, т. е. что тема состоит из некоторых прототипических и многих других родственных, но редких слов. Вполне возможно, что все наиболее характерные слова расположены в более низком частотном спектре распределения. Фактически отсутствует какая-либо количественная модель внутритематического распределения, что делает оценку эффекта неполноты практически невозможной.

4.2 Тематическое моделирование эвристично и не поддается проверке

LDA — это эвристика, которая инициализирует тематические распределения из одного слова и максимизирует различимость оттуда. Другими словами, он берет лексему и распределяет другие лексемы в соответствии с выбранной. Это, очевидно, зависит от точки инициализации: если мы возьмем какой-нибудь предмет в нашем доме и поместим его в коробку, тем самым определяя коробку как вместилище для подобных предметов, а затем расположим все остальные предметы по совпадению, результаты будут различаться в зависимости от первый объект. Если первым объектом был гамак, мы можем найти гамак, соломенную шляпу, лимонад, если мы начнем с шезлонга, мы можем найти стул, соломенную шляпу, лимонад, но не гамак. Обе являются разными категориями, определяемыми порядком ввода и классами, которые существовали до этого момента, и ни одна из них не является неправильной. Но какой из них лучше? Это зависит от цели классификации, по которой моделирование тем не учитывается.

1. Следует отметить, что все статистические вычисления над словами в корпусе эпистемологически проблематичны. Лексикон, вероятно, не соответствует предположениям об эргодичности, т. е. независимости от пути ( [Debowski](#page24) , [2018](#page24) ), и стационарность ( [Piantadosi](#page26) , [2014](#page26) ), т. е. неизменности во времени и пространстве, которые занимают центральное место в стохастической теории и статистике. Фактически представляется, что концепция фиксированной вероятности слов слишком странна во многих отношениях, чтобы ее можно было считать эпистемологически безопасно применимой ( [Шадрова](#page27) , [в печати](#page27) ). Однако, поскольку в тематическом моделировании относительная частота используется в основном описательно, без обращения к внешней совокупности или ожидаемым значениям в конечном результате, эти проблемы не будут рассматриваться в данной статье.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 8 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

Приведенный выше пример не совсем справедлив, потому что LDA, скорее всего, отсортирует гамак в одну и ту же группу из-за его совместного появления с другими предметами. Слова могут встречаться в одной и той же теме, даже если не все они встречаются в документе, т. е. темы могут содержать взаимоисключающие слова. Однако общее вычисление основано на максимизации отличительности, что означает, что оно попытается найти наиболее отчетливое распределение, наблюдаемое из определенной начальной точки, что обязательно приводит к различным моделям для одних и тех же данных. Кроме того, модели очень восприимчивы к изменениям данных, точно так же, как изменится распределение вещей из нашего дома по коробкам, если есть только один предмет, который не подходит ни к одной из определенных нами коробок. Эмпирически это проявляется как высокая чувствительность к незначительным изменениям в корпусе. Например, [Вендель и др.](#page27) ( [в печати](#page27) ) обнаруживают массовые изменения в распределении тем и преобладании между тематическими моделями корпуса решений Федерального конституционного суда Германии, основанного на более чем 3000 текстов, и того же корпуса с добавлением еще 10 текстов.

Поскольку тематические модели различаются в зависимости от порядка слов, которыми они питаются, существует множество различных моделей, т. е. распределений тем для каждого корпуса, и еще больше различных тем. Это общепризнанная проблема при применении тематического моделирования и обычно решается одним из трех способов: (1) с помощью фиксированных точек инициализации на основе предварительной обработки или путем проверки модели либо (2) вычислительными мерами, либо ( 3) человеческий рейтинг ( [Boyd-Grabe r](#page23) [и другие.](#page23) , [2014](#page23) ). Все три являются эпистемологически проблематичными. Определение точки инициализации на основе предварительной обработки, например путем добавления анализа основных компонентов (PCA) или неотрицательной матричной факторизации (NMF, реализованной в пакете R stm ( [Roberts et al.](#page26) , [2019](#page26) ) с [Арора и др.](#page23) ( [2014](#page23) )), оценивает хорошую отправную точку и статистически оптимальное количество тем. Однако это делается на основе алгоритмов, которые сами по себе не дают убедительных результатов (отсюда и разработка LDA в первую очередь). В то время как в сочетании LDA и NMF или PCA согласованность тем достигает более удовлетворительного уровня в прикладных контекстах, чем любая из отдельных процедур, уменьшение размерности до n оптимальных тем не может обеспечить концептуальную гарантию идеального распределения тем.

Решения (2) и (3), человеческие и машинные оценки (показатели) согласованности темы ( [Chang et al.](#page23) , [2009](#page23) ; [Лау и др.](#page25) , [2014](#page25) ; [Бхатия и др.](#page23) , [2018](#page23) ; [Весслен](#page27) , [2018](#page27) ; [Макфарланд и др.](#page26) , [2013](#page26) ), одинаково ограничены, так как их можно вычислить только для существующей модели. Они могут только предложить оценку того, достигает ли модель минимальной согласованности, а не подтверждение модели по сравнению с другими моделями. Это означает более низкий порог, но этого недостаточно для научных целей - обычно целью исследования является не поиск какой-либо минимально последовательной модели, а модель, которая наиболее описательно данных в свете конкретного исследовательского вопроса.

Чтобы полностью проверить модель для научных целей, необходимо сравнить все или, по крайней мере, репрезентативную пропорцию возможных тематических моделей корпуса. Поскольку существует диапазон вероятных номеров тем для любого корпуса и, в отличие от информационного поиска, точное количество категорий имеет значение для таксономического упорядочения, для каждого из этих номеров тем потребуется собственное вычисление из всех точек инициализации.

Допустим, мы хотим сравнить модели из 100–200 тем для большого корпуса, увеличивая их шагами по 25 (100, 125, 150, 175, 200), что дает пять моделей для сравнения — это все еще довольно грубо, модель которая лучше всего соответствует данным, вполне может быть одной из 127 или 138 тем. Используя 10 000 наиболее часто встречающихся лексем stm, для каждой из них можно рассчитать отдельную тематическую модель. Для пяти тематических моделей и 10 000 лексем имеется набор из 50 000 потенциально различных тематических моделей. Некоторые из них могут перекрываться, но трудно предсказать, какие это будут, а также масштаб или значимость их перекрытия.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 9 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

Даже если бы кто-то вложил вычислительную мощность, а затем посвятил себя ручной проверке и перекрестному сравнению 50 из этих моделей (1225 комбинаций от 100 до 200 тем), это все равно оставило бы неопределенность в 99,9%. В этом случае проверка человеком окончательно становится нижним порогом приемлемости более или менее случайной модели, а не гарантией качества в смысле выбора лучшей или даже одной из лучших эмпирических моделей.

Следовательно, на практике невозможно проверить тематические модели на предмет их комбинаторной мощности, и мы фактически вынуждены выбирать одну из случайного и небольшого набора, который вычисляется первым. Единственная проверка, которая может быть обеспечена для полного набора моделей, скрыто существующих в данных, — это проверка с помощью таких показателей, как нагрузка, энтропия или другие меры согласованности ( [McFarland et al.](#page26) , [2013](#page26) ; [Бойд-Грабер и др.](#page23) , [2014](#page23) ; [Весслен](#page27) , [2018](#page27) ) — и даже это вариант только в том случае, если у вас есть вычислительная мощность, готовая вычислить тысячи и тысячи тематических моделей, что на практике почти невозможно в обычных исследовательских контекстах. Однако даже в этом случае статистическая различимость тем не имеет лингвистического коррелята: в лингвистике нет понятия, которое связывало бы определенные степени статистической различимости с определенными качественными аспектами, такими как качество или связность тем. Таким образом, ни предварительная обработка с помощью PCA или NMF, ни показатели когерентности не могут обеспечить достоверность концепции.

4.3 Тематическое моделирование нелингвистично

Тематическое моделирование работает с частотами слов, но это не лингвистическая модель. Он основан на идее, что значение — это скрытая структура, образованная или выраженная через совместно встречающиеся слова, и что это значение может проявляться в более или менее отдельных единицах, которые можно извлечь статистически. Хотя в идее о том, что значение возникает из комбинации слов, есть определенный здравый смысл, одного этого недостаточно для точного и исчерпывающего лингвистического описания.

В тематическом моделировании концепция слов, работающих вместе для построения темы, представлена пространственной близостью (как правило, в пределах документа, но возможны и меньшие разделы), и предполагается, что часто встречающиеся слова образуют более тесные или более связные семантические группы, чем менее часто встречающиеся слова. В лингвистике эта идея моделируется в подразделе дистрибутивной семантики (см. [Барони и др.](#page23) ( [2014](#page23) ) и [Фабр и Ленчи](#page24) ( [2015](#page24) ) для обзора). Он широко используется в компьютерной лингвистике, особенно в прикладных контекстах, например, при встраивании слов ( [Peters et al.](#page26) , [2018](#page26) ; [Девлин и др.](#page24) , [2018](#page24) ; [Леви и Голдберг](#page25) , [2014](#page25) ; [Лю и др.](#page26) , [2015](#page26) ; [Этаярадж](#page24) , [2019](#page24) ) или анализ настроений ( [Medhat et al.](#page26) , [2014](#page26) ; [Бакши и др.](#page23) , [2016](#page23) ). Однако ограничения получения информации более высокого уровня из совпадений слов сами по себе также часто обсуждаются и приводят к плодотворному развитию комбинаций с другими моделями значения, такими как визуальная информация, графы знаний и реляционная семантика ( [Herbelot](#page25) , [2013](#page25) ; [Бруни и др.](#page23) , [2014](#page23) ; [Фрид и Дах](#page24) , [2014](#page24) ; [Шпеер и Лоури-Дуда](#page27) , [2017](#page27) ; [Ленгерих и др.](#page25) , [2017](#page25) ; [Тома и др.](#page27) , [2017](#page27) ). Это связано с тем, что то, что кажется Прямолинейная модель — слова, которые встречаются вместе, образуют более крупные единицы связного значения — на самом деле представляет собой масштабное, последовательное и в значительной степени непредсказуемое сокращение лингвистической размерности. Чтобы проиллюстрировать лишь некоторые проблемы, возникающие при тематическом моделировании:

1. Статистическая различимость не равна тематической различимости, связности или детализации. Моделирование тем с помощью LDA основано на функции максимизации, предполагая, что более высокая статистическая различимость способствует лучшим темам. В лингвистике нет понятия, которое бы охватывало это. Это подразумевает, что отчетливо встречающиеся слова действительно образуют семантические единицы (т. е. топики), в то время как лингвистика только констатирует, что совместно встречающиеся слова могут образовывать семантические единицы, но никоим образом не определяет это количественно.

В языке похожие слова не обязательно обозначают схожие темы, в то время как разные слова могут. Редкие слова, часто встречающиеся вместе, могут образовывать темы, но частые слова, которые

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 10 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

также могут быть более рассредоточены. С этим очень трудно справиться в количественной модели, которая сводит все виды лингвистических аспектов к одному измерению. На самом деле, лексико-количественно очень похожие строки могут даже иметь совершенно разные значения. [12](#page11) Возьмем, к примеру, королеву Нарнии, которая ушла в отставку, и королеву Англии, которая ушла в отставку. Несмотря на почти идентичные формулировки, эти два предложения обозначают совершенно разные жанры и семантику (онтологии, импликации, следствия, семантические поля и фреймы). Если бы корпус далее содержал Король Гондора достиг своих целей, а Король Испании достиг своих целей, основываясь только на этой информации, любой алгоритм, основанный на сходстве строк, правильно отсортировал бы первые два предложения в одну тему, а вторые два в другую. -другие, хотя на самом деле их своеобразие мало что говорит об их содержании. Точно так же предложения В самолете находился врач. Он не мог помочь и в самолете находился доктор философии. Он не мог не обозначать разные темы и даже жанры.

1. Языковая динамика вездесуща и последовательна, и тематическое моделирование не может этого объяснить. Язык не стационарен. Она меняется не только на протяжении длительных периодов истории, но и постоянно. Двумя процессами и лингвистическими особенностями, влияющими на появление слов в документе и корпусе, являются стандартизация ( [Ferguson ,](#page24) [1997](#page24) ; [Лайтинен](#page25) , [2004](#page25) ; [Шмидлин](#page27) , [2011](#page27) ) и производительность (чеканка новых слов, [Баайен](#page23) ( [1994 г.](#page23) ); [Бертрам и др.](#page23) ( [1999](#page23) )).

Это не только приводит к количественным неточностям, но и меняет качественный результат модели. Когда в мире возникает новая область исследования, движение, процесс и т. д., сначала возникает множество терминов, обозначающих ее. Со временем термины стандартизируются и дифференцируются. К тому времени одни и те же или похожие понятия упоминаются с меньшим количеством терминов. Некоторые из ранее перекрывающихся терминов могут стать свободными для использования в других понятиях или просто исчезнуть. Со временем слова не просто добавляются в лексикон, но и дифференцируются по смыслу. Однако этот процесс инициируется не единожды, а постоянно — всякий раз, когда происходит что-то новое, язык адаптируется, чтобы быть в состоянии описать это (продуктивность, диверсификация), и говорящие сходятся на определенных сопоставлениях этих понятий и слов (стандартизация). Этот метапроцесс пересекается между всеми вновь возникающими процессами в мире – первоначальный набор терминов (предновый-процесс) также был результатом эволюционных процессов диверсификации и стандартизации.

Со временем в корпусе появляется все больше и больше слов, хотя и не обязательно в активном словарном запасе каждого говорящего. В тематической модели они по праву должны быть отсортированы по более разнообразным темам. Однако некоторые из этих слов на самом деле могут относиться к одним и тем же понятиям, в то время как другие обозначают действительно новые понятия, а некоторые могут частично совпадать.

В целом тенденция будет заключаться в переоценке диверсификации тем с течением времени, т. е. в результате будет возникать артефакт более высокого разнообразия тем. Соответственно, диверсификация тем с течением времени фактически является обычным наблюдением в сообщаемых тематических моделях ( [Wang et al.](#page27) , [2017](#page27) ; [Писаревская и др.](#page26) , [2020](#page26) ; [Лаубихлер и др.](#page25) , [2019](#page25) ), хотя и без обсуждения влияния языковой динамики. Конечно, верно (и несколько тривиально), что в мире инноваций постоянно возникают новые вещи, поэтому в любом временном ряду дискурса следует ожидать определенной степени диверсификации. По-прежнему необходимо отделить ожидаемое тривиальное разнообразие от лингвистического артефакта и наблюдения, относящегося к подполю.

1. Это, конечно, также зависит от типа значения, на которое ссылаются, например, интенсионал — концептуальное значение — против экстенсионала — представления в мире — против прагматического значения или функциональности против условного истинности значения и т. д. (см. [Гербелот и Ганесалингам](#page25) ( [2013 г.](#page25) ); [Он рбелот](#page25) ( [2013 г.](#page25) ); [Льюис и Стидман](#page25) ( [2013](#page25) ) для некоторых мыслей, касающихся распределительной и вычислительной семантики). Объем данной статьи не позволяет вдаваться в подробности.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 11 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

Моделирование тем также предполагает, что темы статичны — слова, составляющие тему, не меняются — и что меняется только преобладание тем:

«Как только латентным темам доверяют с помощью различных средств, социологи могут начать изучать, как они меняются с течением времени. При этом можно определить приливы и отливы различных языковых доменов или областей исследований в рамках одной области». [Макфарланд и др.](#page26) ( [2013](#page26) , 8).

Однако в действительности очевидно, что это не так. Например, тема компьютерных сетей в 1990 году по сравнению с 2020 годом охватывала совершенно разные концепции, и некоторые из соответствующих концепций возникли, а затем тем временем исчезли. Некоторые даже утверждают, что темы развиваются и развиваются в рамках одного текста или более мелкой единицы:

«(...) Дело стало ощутимо меняться, а именно, возвращаться интерес к тематике, т. е. к улавливанию имевшейся, но рассеянной по тексту информации, как зыбучие туманы (...). Попытка уловить рассредоточенную информацию — движущуюся цель без фиксированного значения, поскольку темы действительно не остаются на месте на протяжении даже умеренно короткого предложения — вот к чему стремится тематика», [Хогенрад и др.](#page25) ( [2003](#page25) , 222).

В то время как корпуса могут быть разделены по времени, моделирование изменяющихся во времени единиц является концептуально, а именно онтологически сложной задачей. Даже если рассчитать модели для каждого года или десятилетия, их интерфейсы потребуют определения (какие темы в t 1 соответствуют тем или группам тем в t 2 ), и каждый из этих интерфейсов будет страдать от той же неопределенности. относительно валидности и эвристики как общей модели. Внутренний текстовый язык и динамика темы даже структурно не поддаются учету в рамках тематической модели, поскольку тематические модели вычисляются с помощью так называемого подхода с использованием мешка слов, при котором теряется всякая внутренняя структура.

4.4 Распространенность темы не является точной количественной оценкой значения

Одно из основных требований, предъявляемых к текстовым исследованиям, заключается в оценке того, как определенные аспекты значения меняются с течением времени или между факторами. Заманчиво рассматривать распространенность темы как представление того, «сколько корпус посвящен теме x». Однако во многих отношениях это сомнительное упрощение.

1. То, как моделирование темы оценивает распространенность темы, не совпадает с тем, как темы обозначаются и воспринимаются. Оценщики-люди а) учитывают только наиболее часто встречающиеся или наиболее характерные слова темы, обычно первые 10 или меньше, и б) улавливают основную концепцию на основе избирательного восприятия, они фильтруют нерелевантные слова в своей категоризации - тема, которая состоит из 9 слов. явно принадлежащие к общей теме, вероятно, будут классифицированы как хорошая тема, даже если 10-е слово не подходит.

Тематическая распространенность, т. е. расчетная доля темы в корпусе, с другой стороны, вычисляется не только из очень релевантных для темы слов, но и из всех слов, отсортированных по одной и той же теме. Сюда входит ряд концептуально слабо связанных терминов, которые все же количественно способствуют преобладанию темы, и некоторые слова, несколько случайно отсортированные в одну и ту же тему. Например, если тематическая модель имеет тему, составленную из набора fnews, журналист, телевидение, станция, передача, отчет, статья, ведущий, радио, catg, оценщик, скорее всего, скажет, что это тема СМИ, и назначит подсчитана его распространенность. Однако распространенность включает все значения слова «кошка» и только некоторые другие тематические слова. Если бы следующими пятью словами в теме — которые оценщик мог бы проигнорировать из-за их маркировки, потому что они могут рассматривать только наиболее часто встречающиеся или отличительные слова для своей маркировки — были бы еда, постель, собака, прогулка, веселье, то распространенность темы была бы включать части документов, не имеющие ничего общего со средствами массовой информации,

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 12 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

но вместо этого с кормом для кошек или лежаками для собак. Изменение количества слов, рассматриваемых для маркировки, только смещает проблему, потому что тема может состоять из сотен слов. Если не найти категорию, которая соответствует всем словам, отсортированным по теме, оценка распространенности неизбежно будет неточной, поскольку она рассчитывается на основе объединенных тем. Таким образом, будучи неполным в отношении лексического инвентаря корпуса, тематическое моделирование на самом деле чрезмерно полно для оценки распространенности темы.

1. Определение и количественная оценка одновременности — сложная задача с лингвистической точки зрения, требующая учета таких параметров, как жанр или тип и длина текста. Считать слова тяжело. Среди лингвистов распространена шутка, что никто не знает, что такое слово. Слова довольно сложно определить, потому что они взаимодействуют с аспектами всех лингвистических слоев (синтаксического, морфологического, фонетического, семантического, прагматического, графематического), но известно, что подсчет слов и осмысление подсчета слов в машинном контексте очень раздражает. среди корпусных лингвистов — настолько, что корпусная лингвистика обычно считает токены (строки, разделенные пробелом), чтобы вообще избежать обсуждения. Например, если семантическое слово состоит из двух графематических слов (например, «Холодная война»), это имеет последствия для количественной модели, хотя лексикографически оно обычно рассматривается просто как одно слово.

Это также несет в себе концептуальную проблему — если все слова, которые разделены на письме, не учитываются и не преобразуются в непрерывные строки в ходе предварительной обработки, их объединенное значение теряется для модели. Точно так же модель не может различить омографы (слова, которые пишутся одинаково, но обозначают разное значение, например, государственный бар, бар на гистограмме и бар, где подают напитки). Это имеет качественные последствия, но также систематически искажает количественные результаты модели.

Кроме того, аналогичное значение может иметь большее или меньшее количество слов (Buffalo buffalo Buffalo buf-falo buffalo buffalo Buffalo buffalo vs. buffalo from Buffalo, которые произносятся буйволами от Buffalo от buffalo, в свою очередь, buffalo buffalo от Buffalo). [13](#page13) В то время как встречаемость строки buffalo составляет 100% в первом примере, во втором она составляет менее 50%, хотя второе предложение является просто объяснением первого; и, кроме того, технически не все строки buffalo являются одним и тем же словом, поскольку некоторые из них являются глаголами, некоторые прилагательными и некоторыми существительными. Для точной количественной модели необходимо принять решения по этим вопросам, а последствия метода требуют учета при интерпретации выходных данных.

Это также влияет на базовый уровень того, что считается совместным возникновением. В то время как стилометрия и поиск информации обычно рассматривают документы, иногда размером с книгу, для их анализа, лингвистические идеи совместного появления обычно связаны со словами, находящимися в некотором приближении друг к другу. Это может быть определено расстоянием в окне из n слов друг от друга (обычно от 3 до 10), позиционно за счет совместного появления в одном и том же предложении, предложении, абзаце или главе или синтаксически через определенные отношения, такие как глагол и объект. Подсчет слов как совместно встречающихся, когда они разделены примерно на 200 страниц, является натяжкой концепции, потому что маловероятно, что читатель сделает семантическую абстракцию между ними так же, как в окне из 5 слов. Это вызывает нежелательные последствия для тематических абстракций. Например, в этом разделе слово «кошка» встречается несколько раз. Ранее в этой статье слово таксономия встречалось несколько раз. При внимательном чтении читатель не счел бы это статьей о таксономии кошек. Однако статистически по сравнению с другими работами

13 [https://en.wikipedia.org/wiki/Buffalo buffalo Buffalo buffalo buffalo buff](https://en.wikipedia.org/wiki/Buffalo_buffalo_Buffalo_buffalo_buffalo_buffalo_Buffalo_buffalo) [фало буйвол буйвол](https://en.wikipedia.org/wiki/Buffalo_buffalo_Buffalo_buffalo_buffalo_buffalo_Buffalo_buffalo)

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 13 | [http://jdmdh.episc iences.org](http://jdmdh.episciences.org) |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

Автор, это статья по таксономии кошка +, хотя эти термины даже не встречаются в одном и том же разделе.

Чем длиннее текст, тем больше вероятность того, что его слова будут охватывать различные и часто несвязанные темы, но подход «мешок слов» не может сделать этого различия. Разделение текста на части для тематического моделирования может быть (и было) сделано, но это не решает проблему предоставления определения того, что составляет часть текста. Это концептуально неразрешимо и неразрешимо только с помощью лингвистических средств. В научных приложениях тематического моделирования длина текста варьируется от твитов ( [Chen et al.](#page24) , [2019](#page24) ) к романам ( [Лю и Джин](#page25) , [2020](#page25) ) и сборники романов ( [Тангерлини и Леонард](#page27) , [2013](#page27) ).

1. Сверхслучайные совпадения слов практически бессмысленны для большинства слов в корпусе. Это подвергает анализ риску чрезмерной интерпретации статистической различимости. Для любого гапакса в корпусе совпадение с любым другим словом статистически крайне маловероятно, и различимость часто может быть достигнута просто благодаря тому факту, что большинство слов редки и могут встречаться только в ограниченном числе контекстов. Из-за большого количества возможных словосочетаний почти все в корпусе вряд ли будет встречаться случайно. [14](#page14)
2. Количественное определение значения концептуально сложно. Тематическое моделирование сначала извлекает совпадающие наборы слов, а затем оценивает их распространенность в корпусе. Обычно это понимается как

(а) слова, принадлежащие к одному и тому же набору слов, образуют связное значение, и (б) их преобладание или соотношение в корпусе выражает их отношение к корпусу или авторам.

Следствие (а) заключается в том, что слова, которые встречаются одновременно в определенной различимости, т. е. встречаются друг с другом, но не в разных темах, образуют более связные семантические единицы. Смысл (б) заключается в том, что для того, чтобы говорить о чем-то (теме), я использую одни и те же слова, и чем более это актуально для меня, тем больше я использую одни и те же слова. Хотя оба имеют определенную степень здравого смысла, они рискованны в количественном анализе: оба предполагают скалярную или постепенную модель актуальности и релевантности. Поскольку распространенность выражается в процентах, еще одним следствием является линейность: математически тема может составлять от 0% до 100% корпуса, а 2% вдвое релевантнее 1%.

Это метатеоретическое предположение [15](#page14) это подразумевает принятие предпосылки о том, что значение поддается количественной оценке только на основе совпадений строк, что оно масштабируется линейно (больше одинаковых слов означает большее количество одинаковых значений). Однако экстраполяция относительной частоты слова проблематична, потому что ограничения этой частоты, как правило, неизвестны. Инжир. [1](#page15) иллюстрирует это с юмором.

Первая проблема заключается в том, что слова ведут себя иначе. Поскольку во всех корпусах большинство слов являются гапаксами, их относительная частота больше зависит от размера корпуса (один на миллион токенов или один на миллиард токенов?), чем от их релевантности. Кроме того, есть несколько процессов, определяющих лексическую частоту, которые имеют мало общего с их релевантностью к теме, например взрывной рост и прайминг (повышенная локальная вероятность появления слова после его использования, [Бок](#page23) ( [1986 г.](#page23) ); [Хоуи](#page25) ( [2012 г.](#page25) ); [Грис](#page24) ( [2005 г.](#page24) ); [Мэдсен](#page26) [и другие.](#page26) ( [2005 г.](#page26) ); [Пьерумбер](#page26) ( [2012](#page26) )), структура текста (некоторые слова, как правило, достоверно встречаются в

14 Комбинаторный потенциал слов обычно недооценивается. Например, [Шадрова](#page27) ( [2020](#page27) ) вычисляет комбинаторный потенциал лексем глагола и объекта винительного падежа по мере их появления (т. е. количество лексем глагола, которые содержат лексемы винительного падежа, умноженное на количество лексем винительного падежа) в небольшом корпусе, содержащем только 21 относительно короткий (<1000 токенов) а тематически похожие тексты на несколько величин превышают предполагаемое количество атомов во Вселенной.

15 «Примером такого метатеоретического предположения может быть то, что чем больше раз данный термин появляется в тексте, тем выше вероятность того, что статья посвящена понятию, выраженному этим термином. Таким образом, метатеоретические предположения шире и менее конкретны, чем теории. Это более или менее сознательные или бессознательные предположения, лежащие в основе теоретической, эмпирической и практической работы». [Хьёрланд \_](#page25) ( [1998](#page25) , 607).

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 14 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

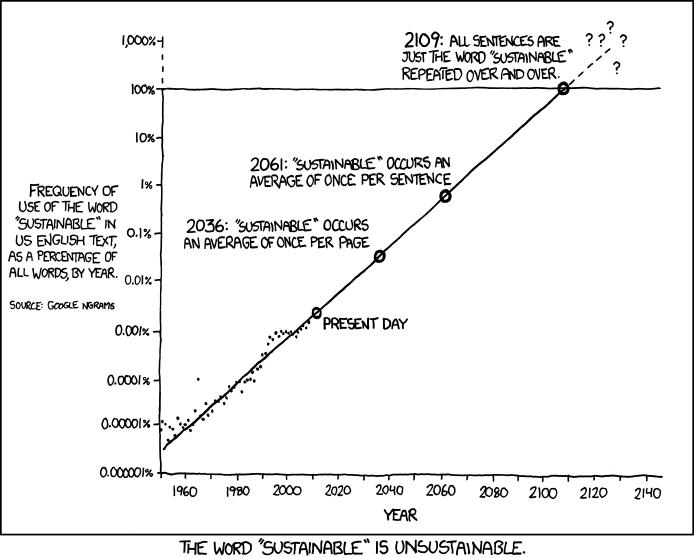


Рисунок 1: Пример неправильной экстраполяции частоты слов в корпусе на свойства реальной жизни.

<https://xkcd.com/1007/>

отдельные части текста, затем исчезают, например, признание), и конвергенция в документах с несколькими авторами (тенденция к конвергенции в использовании синтаксических и лексических выборов и семантических фреймов в диалоге и т. д., [Пикеринг и Браниган](#page26) ( [1998 г.](#page26) ); [стали и](#page27) [Леч](#page27) ( [2006 г.](#page27) ); [Пардо](#page26) ( [2006](#page26) ) и многие другие. Есть попытки объяснить вспыльчивость в тематическом моделировании ( [Дойл и Элкан](#page24) , [2009](#page24) ; [Мэдсен и др.](#page26) , [2005](#page26) ), но это не решает основного вопроса: откуда мы знаем, «сколько» текста связано с x? Рассмотрим следующие примеры:

1. Йохан любит кошек.
2. Йохан - кот.
3. Йохан и Катарина любят кошек.
4. Йохан — кот, а Катарина — не кошка.
5. Йохан определенно кот, а Катарина — нет.
6. Йохан — кошка.

Какая часть каждого предложения посвящена кошкам? В экс. а) делается предложение о двух референтах, одним из которых является кошка. В экс. б) делается предикация – и Йохан, и кот относятся к одному и тому же референту. б) тогда «больше» о кошках, потому что у него нет другого референта? Или это меньше о кошках, потому что это утверждение об одной конкретной кошке, а не о виде? Бывший. в) как упр. а), но теперь с другим референтом. Количественно слово «кошки» сократилось с 1/3 до 1/5. Это меньше о кошках, чем а)? Бывший. г) аналогично упр. б), но с другим референтом, не относящимся к кошкам. Количественно о кошках говорится меньше, чем в b) (1/4 слова в b), 2/10=1/5 в d), хотя кошка встречается дважды в d)) и включает в себя кошку в отрицании. Делает ли это «еще меньше» о кошках? Как насчет экс. д), где первый кот выделен, а второй удален? Определенно делает это «больше» о кошках? Бывший. е) не содержит слова кошка, это все-таки про кошек?

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 15 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

Это не риторические вопросы. В сценарии реального исследования, возможно, не придется взвешивать простые и похожие предложения, подобные этим. На практике, однако, аннотирование категорий высокого уровня, таких как области права или описательные отрывки, общеизвестно сложно и требует тщательной разработки руководящих принципов. Как только концепции станут ясными и поддающимися количественной оценке единицами, такими как «предложения, в которых упоминается по крайней мере одно из этих понятий» или «документы, которые содержат по крайней мере пять упоминаний», необходимо принять решение и повторно проверить их. Этот научный процесс утомителен и сложен, но любое моделирование значения — есть причина, по которой формальное моделирование значения является подобластью философии, лингвистики, когнитивной науки и исследования ИИ. Добиться более или менее точной количественной оценки смысла можно, но, конечно же, ее нельзя приблизить путем подсчета слов в сумке вне контекста или встраивания в специфическую для предмета структуру.

4.5 Тематическое моделирование не позволяет сознательно устанавливать научные параметры

Тематическое моделирование в LDA допускает настройку ряда параметров. Помимо очевидного выбора алгоритма, предварительной обработки и количества выполняемых итераций, они также включают порог сходимости, количество тем (если это не определено посредством предварительной обработки с помощью NMF или PCA) и гиперпараметры, которые относятся к аспекты дистрибутива. [16](#page16) Однако ни один из этих параметров не соответствует непосредственно ряду решений, которые требуются при категоризации контента. К ним относятся как минимум следующие параметры:

Гранулярность: сколько тем? Насколько они могут быть абстрактными или конкретными, все ли они относятся к одному и тому же уровню детализации или существует какая-то таксономия?

Явность/неявность: Являются ли темы явно заявленными и, возможно, даже синтаксически определенными (тема: Король/Королева) или они неявны (тема: горе/потеря/изменение/и т. д.)?

Неоднозначность: ясно ли различаются темы, могут ли они пересекаться?

Объем: насколько далеко могут простираться темы или темы в документе? Насколько они локальны?

Номотетическая, идиотетическая и идиосинкразическая категоризация: определяются ли темы относительно полной таксономии или онтологии вещей (номотетически), идиографически, т. е. по отношению к одной ветви вещей (например, Кафка и отчуждение), или идиосинкразически только для один лист в ветви вещей (например, жук и отчуждение)?

Некоторое обсуждение настройки гиперпараметров в LDA представлено в [Как Юнсьон и др.](#page23) ( [2012](#page23) ) и [Паничелла](#page26) ( [2021](#page26) ), но общая практика и рекомендация - поэкспериментировать с темой чисел до тех пор, пока не будет достигнут удовлетворительный уровень семантической согласованности или ясности. Хотя это приемлемо в прикладном контексте, это, очевидно, эпистемологически проблематично, поскольку вызывает предвзятость подтверждения (остановка на модели, которая дает результаты, наиболее соответствующие существующим предположениям, и игнорирование других типов доказательств) и апофению (чтение закономерностей в случайных данных). ). Однако концептуально это также не то же самое, что изоляция перечисленных выше параметров — количество тем является одним измерением, тогда как параметры существуют в нескольких измерениях. Корректировка номеров тем является вмешательством в рамках уменьшения размерности. Трудно сказать, на каком из объединенных входных измерений это действительно было выполнено.

Например, степень детализации существует в континууме — некоторые темы (в жизни) более обширны, другие — более узки, и легко построить тему с промежуточной степенью детализации между большинством деталей такого рода. Однако правдоподобно предположить, что в этом континууме есть несколько более или менее плотных диапазонов, т. е. областей, где более высокая или более низкая согласованность тем или более

1. альфа и бета отражают ограничения на сочетание тем в документе и слов в теме соответственно. Высокий : документы могут содержать много разных тем, а не одну или две темы конкретно; низкий: документы должны быть классифицированы по небольшому количеству тем.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 16 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

ожидается успешное таксономическое картирование. [17](#page17) Однако при тематическом моделировании большее количество тем может соответствовать меньшей степени детализации, меньшей степени неопределенности, более ограниченной локальности или, вероятно, всему вышеперечисленному. На практике представляется невозможным сопоставление большей или меньшей степени детализации по большему или меньшему номеру темы. То же самое верно в отношении двусмысленности, охвата или различных точек зрения, таких как номотетические и идиотетические группы. Невозможно сознательно искать определенные диапазоны в этих разнообразных измерениях, корректируя номера тем.

Поскольку имплицитное значение не заключено в словах, а только прочитывается в (или из) близости слов, никакой параметр даже теоретически не может быть создан для определения того, что ищется. Риск апофении здесь особенно высок. Чтобы экстраполировать наиболее полезное или информативное распределение тем, должна быть возможность проводить эксперименты с гранулярностью, двусмысленностью и т. д., но тематическое моделирование структурно не способно на это.

4.6 Тематическое моделирование недооценивает аналитический потенциал гуманитарных наук

Опираясь на частоту совпадения и различения в сильно сокращенном информационном пространстве, внутренняя работа тематического моделирования оптимизирует распространение контента до наиболее типичной категоризации. Это в лучшем случае — в худшем случае они дают несколько случайные или даже вводящие в заблуждение результаты. [18](#page17) Даже в лучшем случае уже упорядоченная информация затем проходит через человеческое сознание исследователя, что при отсутствии сложной контекстуальной информации в том виде, в каком она существует при внимательном чтении, и без конкретных методологических указаний на интерпретацию информации. Тематическое моделирование еще больше усилит эту прототипичность посредством гештальт-восприятия и когнитивных искажений.

Недавно это было эмпирически показано [Жиллингс и Харди](#page24) ( [готовится к печати](#page24) ), который провел исследование, в котором сравнивал подход «на глаз», основанный на темах, извлеченных с помощью LDA из новостных статей, с подходом к внимательному чтению тех же документов. Из двадцати тем они находят только одну тему, совпадающую между двумя подходами («развитие ветряных электростанций в Шотландии»). Несовпадающие темы включают в себя серьезные расхождения, такие как «сельское хозяйство и сельское хозяйство» (тематическое моделирование) в сравнении с «экологически чистыми покупками» (внимательное чтение) или «сообщество» (тематическое моделирование) в сравнении с «терроризм» (внимательное чтение), а также более постепенные сдвиги, такие как «лейбористы и консерваторы» против «гонки за лидерство лейбористов» или «дикая природа» против «находящейся под угрозой исчезновения или вымершей дикой природы».

Вместо того, чтобы синтезировать наблюдение в знание всей сложности, исследователь ограничивается соединением точек. Оптические иллюзии, например изображения, которые можно воспринимать как утку или кролика, лучше всего работают только с контуром. Добавление очевидных рисунков меха или перьев приведет к меньшему наложению двух возможных форм. Таким неоднозначным контуром может быть тематическая модель – однако, в отличие от оптической иллюзии, изучаемый текст может вовсе не быть двусмысленным. Темы, полученные из тематических моделей, остаются конструкциями, интерпретируемыми, а не считываемыми из комбинации слов, присутствующих в каждой теме. Но так как мало

17 Подобный тип уровня идеальной категории известен как категории базового уровня в психологии восприятия и обучения ( [Rosch et al.](#page27) , [1976](#page27) ; [Гаджибекова](#page25) , [2013](#page25) ; [Эймас и Куинн](#page24) , [1994](#page24) ; [Маркман и Вишневский](#page26) , [1997](#page26) ). Категория базового уровня — это значимая для восприятия, легко обнаруживаемая и запоминаемая категория типа объекта в мире, например автомобиль, а не транспортное средство или Форд; или собаку, а не собаку или бордер-колли. Категории базового уровня также обычно используются для обозначения недавно введенных референтов без дальнейшего контекста («Я у дерева» или «Я у дуба», а не «Я у японского вечнозеленого дуба»). Они могут меняться с опытом, например, владелец собаки может воспринимать бордер-колли как категорию базового уровня, в то время как продавец автомобилей может рассматривать тип Ford как базовый уровень.

1. Это было показано [Шмидт](#page27) ( [201 2](#page27) ) в основанном на тематическом моделировании анализе ссылок на китобойный промысел, который подразумевает связи, которых никогда не существовало. В случае китобойного промысла это относительно легко исправить, но в более неясных темах вводящие в заблуждение результаты тематического моделирования могут привести к совершенно неподходящим моделям истории или литературы.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 17 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

остается информация, особенно из большого корпуса, интерпретация тематической модели может быть похожа на перелистывание изображения. Со временем это недостаточное изучение существующих аналитических способностей накапливает много ненадежной и необработанной информации, хотя существуют другие способы познания, которые могли бы быстро прояснить проблему. Тематическое моделирование эффективно обрывает связь между исследователем и его фактическим объектом исследования, которым является полная коллекция текста, а не его сокращение, как услужливо заметил Фредерик Клавер. Вместо того, чтобы способствовать стабильности эпистемологической структуры и концептуальной ясности предмета, некритическое использование вычислительных методов может эффективно работать как ограничивающий фактор в понимании больших текстовых данных.

V ЗАКРЫТИЕ ПРОБЕЛОВ: К НАИЛУЧШЕЙ ПРАКТИКЕ

В то время как все модели редукционистские по своей природе и абстрагируются от деталей, присутствующих в данных, тематическое моделирование делает это с опорой на неопределенные лингвистические интерфейсы и довольно далеко идущими, непредсказуемыми и до сих пор немоделируемыми способами.

Ясно, что идеи а) неизменяемых семантических пространств и б) обнаружения значений посредством подсчета совпадений строк являются концептуальными упрощениями ради операционализации. Дело не столько в том, что эти упрощения являются эпистемологически проблематичными, сколько в непредсказуемости и плотности этой редукции размерности: совершенно невозможно точно сказать, какая информация была утеряна при сгущении исходного текста до мешков с текстом. слова и, наконец, тематические модели.

Настоящая проблема использования тематического моделирования в научных целях заключается в прояснении объема данных, которые оно предоставляет (ср. [Леонелли](#page25) ( [2019](#page25) )), в разработке передового опыта в области аргументированного внедрения результатов тематического моделирования и в формулировании исследовательской программы, которая приведет к методологической интеграции тематического моделирования с лингвистическими и текстовыми исследованиями до такой степени, что позволяют использовать его как самостоятельный метод исследования.

Тематическое моделирование без массовой проверки не дает лучших, более объективных или более точных данных для большинства исследовательских вопросов, чем внимательное чтение. Следовательно, в центре внимания аргумента должно быть то, как доказательства, которые он предоставляет, на самом деле относятся к исследовательскому вопросу и как его неопределенности влияют на аргумент.

Однако в настоящее время не принято подробно связывать тематические модели с исследовательскими вопросами. Вместо этого его часто представляют как «исследовательские данные» или «новый взгляд на данные». [19](#page18) Это само по себе необычно для эмпирика, и, по-видимому, не так уж часто эти исследования сопровождаются подтверждением, повторением или опровержением в более поздних исследованиях. Это противоречит концепции исследования данных, и в любом случае публикация исследования данных в количественных полях не является общепринятой практикой. Там, где это делается, это происходит в контексте планов или предложений будущих исследований, как случайные результаты в контексте других анализов, основанных на гипотезах, или там, где представлены ресурсы. Было бы крайне необычно опубликовать модель со смешанными эффектами или дисперсионный анализ для эксперимента и вместе с этим добавить все анализы и споры данных для документирования процесса или предположить, что некоторые из случайных графиков обладают высоким потенциалом для дальнейшего анализа, не связывая это. вернуться к аспектам теории. Если результаты моделирования необработанных тем можно воспринимать всерьез как новые, интересные и

1. См. например [Эрлин](#page24) ( [2018](#page24) , 3), [Картер и др.](#page23) ( [2016](#page23) , 1300–1301), или [Роди](#page26) ( [2012](#page26) , 19): «(...) Я полагаю, что тематическое моделирование поэзии работает отчасти из-за ее неудач. Где-то между литературной возможностью, содержащейся в корпусе из тысяч англоязычных стихотворений, и вычислительной строгостью скрытого распределения Дирихле (LDA) есть интерпретационное пространство, столь же важное, как плетение и распутывание на ткацком станке Пенелопы».

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 18 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

соответствующего научного исследования данных, чем что-либо другое, и не требуется никаких дальнейших методологических дебатов.

Если, с другой стороны, методологическая проверка принимается как требование научного исследования, то любой вклад требует аргументированного вложения. Почему и для кого она интересна, актуальна или нова — на какой исследовательский вопрос она отвечает и насколько достоверно?

Хотя сама тематическая модель отражает объективную реальность, а именно результат максимизации функции распределения по словам в корпусе, интерпретация тематической модели не является объективной. Это абдуктивно, что означает, что оно контекстуализировано. Эта контекстуализация отражает систему взглядов отдельного ученого. Чрезвычайно важно сделать лежащую в основе контекстуализацию, модель явной, как посредством работы, основанной на гипотезах, так и путем привязки результатов к теоретическим и концептуальным дебатам в этой области. Представление результатов тематического моделирования как «голых», «исследовательских» результатов подвергает их риску стать троянскими конями, проникающими в научный дискурс в качестве подтвержденных и даже в некоторой степени объективных знаний, а не концептуально очень базовых типов информации, которыми они на самом деле являются. Это влечет за собой проблемы с надежностью, достоверностью и объективностью. Паттерны, даже если они бросаются в глаза и кажутся уместными, не обязательно имеют смысл и не классифицируются как эпистемы. Их актуальность проистекает не из того, что они очевидны, а из того, что они встроены в научный аргумент (ср. [Диксон](#page24) ( [2012](#page24) )).

Принятие результатов исследования тематических моделей в исследовательскую литературу также сопряжено с риском обесценивания усилий по осторожному и явному моделированию знаний в гуманитарных и социальных науках с помощью вычислительных методов. Если результаты тематического моделирования рассматриваются как столь же убедительные или дополняющие гораздо более глубокий и более проработанный анализ в количественных и качественных исследованиях, культура ожидания вокруг сильных заявлений может развиваться легче. В долгосрочной перспективе это нанесет ущерб гуманитарным и общественным наукам и их научной репутации.

Если исследователи текстовых областей настаивают на использовании тематического моделирования в качестве метода исследования, необходимо разработать программу его методологической интеграции. Это должно включать как минимум следующие вопросы:

1. Существуют ли в тексте темы как определяемая и измеримая сущность?
2. Существует ли статистическая корреляция тем с определенной степенью детализации?
3. Насколько похожи различные тематические модели корпуса? Существуют ли более или менее хорошие модели, и можно ли их количественно определить, не полагаясь на массовую человеческую оценку?
4. Насколько точна оценка распространенности тем с помощью тематического моделирования по сравнению с количественной оценкой распространенности, основанной на темах, смоделированных учеными? Как на это влияет длина текста, жанр и другие тексто-лингвистические факторы?
5. На какие исследовательские вопросы действительно может ответить тематическое моделирование, помимо «существует ли хотя бы одно статистически отличительное разделение корпуса на группы, удовлетворяющее определенному требованию»?
6. Каковы эффекты и артефакты языковой динамики на оценку тематического разнообразия?
7. Учитывая массовое сокращение информации, насколько хорошо тематическое моделирование подходит для исследования корпуса? Как это соотносится с внимательным чтением? Как он соотносится с другими алгоритмами, такими как кластеризация k-средних или PCA? Когда и где потеря информации является допустимым упрощением, а где она вводит в заблуждение [( Шмидт](#page27) , [2012](#page27) )? Как мы отличаем одно от другого?

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 19 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

1. Как можно использовать тематическое моделирование в текстовых исследованиях, не слишком полагаясь на его сомнительные результаты? Например, [Вендель и др.](#page27) ( [в печати](#page27) ) используйте его в качестве лексического предварительного фильтра, чтобы найти слова, относящиеся к областям права, не принимая во внимание распространение или распространенность тем, а затем создавайте категории на основе знаний по конкретным предметам. [Тангерлини и Леонард](#page27) ( [2013](#page27) ) предлагают три метода исследования для различных вариантов использования в литературоведении. Как можно синтезировать эти и другие подходы в четко определенную структуру для комбинированного количественного и качественного анализа текста?

В отсутствие такой программы рекомендуется скептически относиться к тематическому моделированию и другим методам извлечения информации в текстовых исследованиях. Передовая практика должна включать уточнение вопроса исследования и сопоставление категорий с предметно-ориентированными моделями, разъяснение гипотез и вариантов интерпретации, хорошо аргументированную причину включения типа доказательств, предоставляемых методом, и чувствительность к его предостережениям. По умолчанию следует не полагаться на результаты автономного тематического моделирования, если для этого нет веских причин в конкретном исследовании. [20](#page20)

Один из способов частично избежать ограничений тематического моделирования, не отказываясь полностью от идеи дистанционного чтения, заключается в одновременном рассмотрении нескольких уровней плотности сложности, иногда называемых мультимасштабным чтением ( [Моа и Росс](#page26) , [2019](#page26) ). Это реализовано, например, в текстовом исследовательском ПО IRaMuTeQ ( [Camargo и Justo](#page23) , [2013](#page23) ; [Соуза](#page27) [и другие.](#page27) , [2018](#page27) ), которая поддерживает связь между конкретной реальностью вовлеченных текстов и его абстракции с помощью статистической классификации и других алгоритмов кластеризации. Важно отметить, что основное внимание здесь уделяется качественному анализу. Это позволяет избежать многих проблем, обсуждавшихся ранее, поскольку не подразумевается никаких нежелательных предположений о скалярности, линейности или стационарности. Однако если все тексты не рассматривать как через призму близкого, так и дистанционного чтения, остается относительно высокая степень неопределенности, а если так, то необходимость дистанционного чтения ставится под вопрос.

Если количественный анализ текста с помощью мер извлечения информации оговаривается как неизбежный в каком-то контексте, способ избежать наиболее проблематичных аспектов тематического моделирования заключается в рассмотрении алгоритмов, которые не полагаются на подходы «мешка слов» и/ или независимость от темы. [Ламирель и др.](#page25) ( [2020](#page25) ) представляют такой подход в кластеризации слов с максимизацией характеристик в статьях, относящихся к области науки о науке в Китае. При максимизации признаков слова моделируются как признаки, а комбинации признаков сравниваются с точки зрения сходства или расстояния. Этот подход, в отличие от моделирования темы, не предполагает независимости темы, а вместо этого позволяет моделировать отношения между кластерами, которые можно визуализировать в виде так называемых контрастных графов. Это позволяет получить некоторое представление о степени детализации и подключении кластера, а поскольку алгоритм полностью не контролируется и не зависит от настройки гиперпараметров, он менее подвержен непреднамеренной настройке данных.

Подходы кластеризации такого рода обеспечивают явные преимущества с точки зрения согласования между концептуальной и математической моделью в некоторых отношениях и могут давать лучшие результаты с точки зрения качества извлеченного контента. [Ламирель и др.](#page25) ( [2020](#page25) ) показывают это в сравнении с темами того же корпуса, извлеченными с помощью LDA. Однако кластеризация не решает основных лингвистических проблем, связанных с интерпретацией значений совпадений слов вне непосредственного контекста. Это означает, что многие из обсуждавшихся ранее ловушек остаются, в том числе тот факт, что не существует лингвистической модели, которая сопоставляла бы совпадения слов или их частоту с

20 Одним из таких примеров является исследование [Блок и Ньюман](#page23) ( [2011](#page23) ). Он показывает сохраняющееся разнообразие тем как контрдоказательство утверждения о том, что исторические журналы ограничивают свой охват темами женской истории. Для этого конкретного исследовательского вопроса достаточно просто свидетельства разнообразия, чтобы опровергнуть это утверждение, но это не типично для большинства исследований.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 20 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

специфические смысловые единицы, такие как темы. Хотя в своем исследовании Ламирель и его коллеги обнаруживают высокую сходимость между извлеченными кластерами и экспертными описаниями истории науки в Китае, это нельзя рассматривать как окончательное доказательство способности алгоритма раскрывать объективные реальности, лежащие в основе текста. Кластеры слов по-прежнему подлежат интерпретации, они естественно не соответствуют определенным темам или областям, а конструируются как таковые через эксперта, что в постфактумной интерпретации может быть затронуто апофенией и предвзятостью подтверждения.

Чтобы быть уверенным в том, что данные подтверждают гипотезу, группы слов, которые могут быть интерпретированы как индикаторы определенной темы, должны быть описаны до анализа. В Ламиреле и соавт. это может быть не так, потому что эксперт (профессор в этой области) также будет хорошо знаком с включенными текстами, таким образом, фактически не полагаясь на дистанционное чтение - в биографическом смысле это можно даже рассматривать как мультимасштабное чтение. В отличие от этого дистанционное чтение часто предлагается как альтернатива внимательному чтению больших коллекций текстов и фактически как единственный способ приблизиться к тексту в эпоху больших данных. Конкретный исследовательский вопрос может оправдать использование конкретных методов извлечения информации, таких как кластеризация с максимизацией признаков. Но вообще говоря, степень эпистемологической и лингвистической неопределенности остается высокой везде, где извлечение значения опирается почти исключительно на слова, а не на более глубокие аннотации, описывающие смысловые структуры более высокого порядка или онтологии, относящиеся к данной области.

VI ЗАКЛЮЧЕНИЕ

Обсуждение в этой статье показало, что доказательства, полученные с помощью тематического моделирования, концептуально слабы по ряду причин. Даже для статистически оптимальной модели интерфейс с лингвистикой в значительной степени не определен. Поскольку проверка изнутри в значительной степени невозможна, данные тематического моделирования сами по себе имеют очень ограниченный вес.

Везде, где в игру вступают статистика, комбинаторика и сложность, здравых объяснений редко бывает достаточно. Легко отличить кита от попугая, но правильно идентифицировать разные подвиды попугаев гораздо сложнее. Точно так же очевидно, что в корпусе, в котором в каждом втором тексте упоминается слово «судебный», по сравнению с корпусом, в котором во всех остальных текстах упоминается слово «корм для кроликов» и не упоминается «судебный», разные темы в любом смысле этого слова скорее всего, будут в игре. Тем не менее, это не эквивалентно скалярной или постепенной модели примерности на основе статистических характеристик текста.

Частота (совместного) появления не является достаточным маркером для границ категории. Это связано с различной степенью детализации тем, но в большей степени из-за специфики распределения лексем. Предположение о неизменности частоты (совместного) появления по теме и максимизации дискретности между темами может быть еще более проблематичным, поскольку оно смещает тематические модели к наиболее прототипическим образцам и систематически противодействует распознаванию более тонких различий и динамических или меняющихся границы. Это делает модель структурно неспособной исчерпать аналитический потенциал социальных или гуманитарных наук. Это также создает впечатление объективности («это согласуется с общим консенсусом в данной области») и очевидной согласованности («эти слова явно связаны»), что может больше опираться на человеческую предвзятость к восприятию прототипов, апофении. , и предвзятость подтверждения, чем «фактическая», значимая согласованность.

Проблема с тематическим моделированием заключается не в том, что оно не может дать последовательных или даже полезных результатов, а в том, что эти результаты являются редукционистскими относительно непредсказуемыми способами и что нет очевидного способа интегрировать такие результаты как предметно-специфические знания без высокой степени. методологических и теоретических усилий. Слепо принимать его результаты — значит овеществлять их.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 21 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

как знания, хотя на самом деле они представляют собой просто в значительной степени немодулированный тип информации. Если тематическое моделирование и другие вероятностные методы интеллектуального анализа текста должны серьезно восприниматься в продвижении знаний в социальных и цифровых гуманитарных науках, они должны подвергнуться такому же строгому анализу и проверке, как и другие методы.

Это также подчеркивает важность четких исследовательских вопросов и систематического подхода к выбору, применению и проверке методов. Определить область доказательности метода означает прояснить исследовательские вопросы, на которые он может с уверенностью ответить относительно типа данных: «С помощью этого метода можно ответить на вопрос X при условии, что есть доступ к данным типа Y, и ответ может быть интегрирован в заданную теоретическую модель способами a, b или c».

Исследователи в области естественных наук располагают набором инструментов, состоящим в основном из четко определенных методов, которые можно использовать для четко определенных целей: они знают, когда использовать рН-тест, а когда использовать тахометр, и они, скорее всего, не будут измерять кислотность жидкости и сообщать о ней. или скорость вращающегося объекта в исследовательских целях, если только они не провели серию тестов, которые включали кислотность или скорость вращения среди других аспектов. Эта методологическая уверенность возникла не за одну ночь, а развивалась в течение столетий научных исследований. Аналогичные усилия по систематизации необходимы в компьютерно-ориентированных областях текстовых исследований.

Без такого встраивания почти неизбежно включение ложной или случайной информации посредством тематического моделирования. Хотя это может иметь небольшое значение в прикладных контекстах, поскольку ложные результаты могут быть отфильтрованы в цикле обратной связи приложения, это не относится к социальным наукам или цифровым гуманитарным наукам. Если бы мы доверяли машине, мы были бы вынуждены принять любое обнаруженное скрытое измерение как объективно существующее — и, следовательно, если бы мы серьезно относились к текущим исследованиям, нам пришлось бы рассматривать доказательства, предоставляемые машиной во всех аспектах. будущие исследования. Слепое принятие сгенерированных машиной тематических распределений в наше понимание предмета похоже на оснащение библиотек книгами с семантически связными названиями (как определено оценщиками-людьми и машиной), но в основном с зачерненными страницами и обращением к этой библиотеке как к библиотеке наших знаний. предмета.

Сами по себе тематические модели не обеспечивают лучших, более объективных или более точных доказательств, чем другие типы построения категорий на основе данных. Темы не являются однозначно определенным понятием и не могут быть однозначно определены для любого текста. На самом деле они даже не являются языковым понятием как таковым, а конструируются из слов текста, его контекста и даже пространства между словами текста. Решение о том, действительно ли слова в тематической модели темы соответствуют признанной или вероятной новой категории в их дисциплине, остается и должно оставаться за ученым. Если темы, темы и т. д. задуманы как научные категории, они требуют построения и синтеза процесса дебатов, тем самым создавая научное знание.

Широкий спектр методологической литературы показывает, что к прояснению аргументативной логики и построения категорий в качественных социальных и гуманитарных науках предъявляются высокие требования. То же самое должно относиться к а) количественным текстовым исследованиям в целом и б) компьютерным методам в частности.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 22 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

использованная литература

Арора С., Ге Р., Халперн Ю., Мимно Д., Мойтра А., Зонтаг Д., Ву Ю. и Чжу М. (2013). Практический алгоритм тематического моделирования с доказуемыми гарантиями. На Международной конференции по машинному обучению, стр. 280–288.

Арора С., Ге Р. и Мойтра А. (2014). Новые алгоритмы изучения бессвязных и переполных словарей.

В конференции по теории обучения, страницы 779–806.

Аскари, М. (2014). Опасности дистанционного чтения: переоценка подхода Моретти к литературным жанрам. Жанр:

Формы дискурса и культуры, 47 (1): 1–19.

Асгари, Э., Гассеми, М., и Финлейсон, Массачусетс (2013). Подтверждение тематики и интерпретативного единства газельской поэзии с использованием тематических моделей. В семинаре по системам обработки нейронной информации (NIPS) для тематических моделей.

Асунсьон, А., Веллинг, М., Смит, П., и Тех, Ю.В. (2012). О сглаживании и выводе для тематических моделей. Препринт arXiv arXiv: 1205.2662.

Асунсьон, Х.У., Асунсьон, А.С., и Тейлор, Р.Н. (2010). Отслеживаемость программного обеспечения с тематическим моделированием. В 2010 г. 32-я Международная конференция ACM / IEEE по программной инженерии, том 1, страницы 95–104. IEEE.

Баайен, Р. Х. (1994). Продуктивность в языковом производстве. Язык и когнитивные процессы, 9(3):447–469.

Баайен, Р. Х. (2002). Распределение частот слов, том 18. Springer Science & Business Media.

Бакши Р.К., Каур Н., Каур Р. и Каур Г. (2016). Сбор мнений и анализ настроений. В 2016 г. 3-я международная конференция по вычислительной технике для устойчивого глобального развития (INDIACom), страницы 452–455. IEEE.

Бао С., Сюй С., Чжан Л., Ян Р., Су З., Хан Д. и Ю Ю. (2009). Совместное моделирование эмоций и тем для анализа социального аффективного текста. В 2009 г. Девятая международная конференция IEEE по интеллектуальному анализу данных, страницы 699–704. IEEE.

Барони М., Бернарди Р., Зампарелли Р. и др. (2014). Фреге в космосе: программа композиционно-распределительной семантики. Лингвистические проблемы языковых технологий, 9 (6): 5–110.

Берталан, В.Г. и Руис, ЕЭС (2019). Использование тематического моделирования для поиска основных тем для обсуждения на бразильских политических веб-сайтах. В материалах 25-го бразильского симпозиума по мультимедиа и Интернету, страницы 245–248.

Бертрам Р., Лайне М. и Карвинен К. (1999). Взаимодействие типа словообразования, аффиксальной омонимии и продуктивности лексической обработки: свидетельство морфологически богатого языка. Журнал психолингвистических исследований, 28 (3): 213–226.

Бхатия, С., Лау, Дж. Х., и Болдуин, Т. (2018). Внедрение темы для автоматической оценки модели темы. В материалах конференции 2018 г. по эмпирическим методам обработки естественного языка, страницы 844–849.

Блей, Д., Карин, Л., и Дансон, Д. (2010). Вероятностные тематические модели. Журнал обработки сигналов IEEE, 27 (6): 55–65.

Блей, Д. и Лафферти, Дж. (2006). Коррелированные тематические модели. Достижения в области нейронных систем обработки информации, 18:147.

Блей, Д.М. (2012). Вероятностные тематические модели. Сообщения ACM, 55 (4): 77–84.

Блей, Д.М., Нг, А.Ю. и Джордан, М.И. (2003). Скрытое распределение Дирихле. Журнал исследований машинного обучения, 3 (январь): 993–1022.

Блок, С. и Ньюман, Д. (2011). Что, где, когда, а иногда и почему: анализ данных за два десятилетия истории женщин. Журнал женской истории, 23 (1): 81–109.

Бок, Дж. К. (1986). Значение, звук и синтаксис: лексическая подготовка в построении предложений. Журнал Экспериментальный

Психология: обучение, память и познание, 12 (4): 575.

Boyd-Graber, J., Hu, Y., Mimno, D., et al. (2017). Приложения тематических моделей. Основы и тенденции R в информационном поиске, 11 (2-3): 143–296.

Бойд-Грабер, Дж., Мимно, Д., и Ньюман, Д. (2014). Уход и кормление тематических моделей: проблемы, диагностика, доработки. Справочник по моделям смешанного членства и их приложениям, 225255.

Бруни, Э., Тран, Н.-К., и Барони, М. (2014). Мультимодальная дистрибутивная семантика. Журнал исследований искусственного интеллекта, 49:1–47.

Камарго, Б.В. и Хусто, А.М. (2013). IRAMUTEQ: бесплатное программное обеспечение для анализа текстовых данных. Temas em psikologia, 21(2):513–518.

Картер, Д.Дж., Браун, Дж., и Рахмани, А. (2016). Чтение Верховного суда на расстоянии: Тема моделирования юридического предмета и судебной деятельности Высокого суда Австралии, 1903-2015 гг. UNSWLJ, 39:1300–1354.

Люстра М., Стейкардт А., Матевет Р., Диверси С. и Хименес О. (2018). Контент-анализ освещения в газетах реколонизации волков во Франции с использованием структурного тематического моделирования. Биологическая консервация, 220:254–261.

Чанг, Дж., Герриш, С., Ван, К., Бойд-Грабер, Дж. Л., и Блей, Д. М. (2009). Чтение чайных листьев: как люди интерпретируют тематические модели. В Достижениях в области нейронных систем обработки информации, страницы 288–296.

Чен, X., Цзоу, Д., Ченг, Г., и Се, Х. (2020). Выявление скрытых тем и трендов в образовательных технологиях

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 23 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

более четырех десятилетий с использованием структурного тематического моделирования: ретроспектива всех томов по компьютерам и образованию.

Компьютеры и образование, стр. 103855.

Чен Ю., Чжан Х., Лю Р., Е З. и Линь Дж. (2019). Экспериментальные исследования по добыче коротких текстов между схемами на основе LDA и NMF. Системы, основанные на знаниях, 163:1–13.

Чуанг Дж., Гупта С., Мэннинг С. и Хир Дж. (2013). Диагностика тематической модели: оценка релевантности предметной области посредством тематического выравнивания. В Международной конференции по машинному обучению, страницы 612–620.

Чуанг, Дж., Робертс, М.Е., Стюарт, Б.М., Вайс, Р., Тингли, Д., Гриммер, Дж., и Хир, Дж. (2015). TopicCheck: интерактивное выравнивание для оценки стабильности тематической модели. В материалах конференции 2015 года Североамериканского отделения Ассоциации вычислительной лингвистики: технологии человеческого языка, страницы 175–184.

Даллоф М. и Берглунд К. (2019). Лица, драки и семьи: моделирование темы и гендерные темы в двух корпусах шведской прозы. В DHN 2019, 4th Digital Humanities in the Nordic Country 2019, Копенгагенский университет, Копенгаген, Дания, 6–8 марта 2019 г., страницы 92–111.

де Пауло Фалейрос, Т. и де Андраде Лопес, А. (2016). Об эквивалентности алгоритмов неотрицательной матричной факторизации и скрытого размещения Дирихле. В ЕСАНН.

Дебовски, Л. (2018). Является ли естественный язык периграфическим процессом? Повторение теоремы о фактах и словах.

Энтропия, 20:85.

Дирвестер, С., Дюме, С.Т., Фернас, Г.В., Ландауэр, Т.К., и Харшман, Р. (1990). Индексирование с помощью скрытого семантического анализа. Журнал Американского общества информатики, 41(6):391–407.

Девлин Дж., Чанг М.-В., Ли К. и Тутанова К. (2018). Берт: Предварительная подготовка глубоких двунаправленных преобразователей для понимания языка. Препринт arXiv arXiv: 1810.04805.

Ди Маджио, П., Наг, М., и Блей, Д. (2013). Использование сходства между моделированием темы и социологическим взглядом на культуру: применение к газетному освещению финансирования искусства правительством США. Поэтика, 41(6):570–606.

Дин, К., Ли, Т., и Пэн, В. (2006). Неотрицательная матричная факторизация и вероятностное скрытое семантическое индексирование:

Статистика эквивалентности хи-квадрат и гибридный метод. В AAAI, том 42, страницы 137–43.

Диксон, Д. (2012). Инструмент анализа или методология исследования: существует ли эпистемология закономерностей? В книге «Понимание цифровых гуманитарных наук», страницы 191–209. Спрингер.

Дойл, Г. и Элкан, К. (2009). Учет взрывчатости в тематических моделях. В материалах 26-й ежегодной международной конференции по машинному обучению, ICML '09, стр. 281–288, Нью-Йорк, штат Нью-Йорк, США. Ассоциация вычислительной техники.

Друкер, Дж. (2017). Почему не дистанционное чтение. ПМЛА, 132(3):628–635.

Эймас, П.Д. и Куинн, ПК (1994). Исследования формирования перцептивных категорий базового уровня у младенцев раннего возраста. Развитие ребенка, 65 (3): 903–917.

Эрлин, М. (2018). Тематическое моделирование, эпистемология и английский и немецкий роман. СоцАрХив.

Этаярадж, К. (2019). Насколько контекстуальными являются контекстуализированные представления слов? Сравнение геометрии вложений BERT, ELMo и GPT-2. Препринт arXiv arXiv: 1909.00512.

Фабр, К. и Ленчи, А. (2015). Распределительная семантика сегодня: введение в специальный выпуск. Revue TAL, ATALA (Association pour le Traitement Automatique des Langues), Semantique´ Distributionnelle, 56(2):7–20.

Фергюсон, Калифорния (1997). Стандартизация как форма распространения языка. В Структуралистских исследованиях арабского языкознания, страницы 69–80. Брилл.

Фрид, Д. и Дух, К. (2014). Включение как дистрибутивной, так и реляционной семантики в представления слов.

Препринт arXiv arXiv: 1412.4369.

Жиллингс, М. и Харди, А. (ожидается). Интерпретация тематических моделей для научного анализа: оценка и критика современной практики.

Джиролами, М. и Кабан, А. (2003). Об эквивалентности между PLSI и LDA. В материалах 26-й ежегодной международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска, стр. 433–434.

Голдстоун, А. и Андервуд, Т. (2014). Тихие трансформации литературоведения: что могли бы нам сказать тринадцать тысяч ученых. Новая литературная история, 45 (3): 359–384.

Гутсос, Д. (1997). Моделирование темы дискурса: последовательные отношения и стратегии в пояснительном тексте, том 59.

Издательская группа Гринвуд.

Гранехейм, У. Х. и Лундман, Б. (2004). Качественный контент-анализ в исследованиях сестринского дела: концепции, процедуры и меры для достижения достоверности. Медицинское образование сегодня, 24 (2): 105–112.

Грин, Д. и Кросс, Дж. П. (2016). Изучение политической повестки дня европейского парламента с использованием подхода к динамическому тематическому моделированию. Препринт arXiv arXiv: 1607.03055.

Грис, ST (2005). Синтаксическое прайминг: подход, основанный на корпусе. Журнал психолингвистических исследований, 34(4):365–399.

Гриммер, Дж. (2010). Байесовская иерархическая тематическая модель для политических текстов: измерение выраженных программ в

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 24 | [http://jdmdh.episc iences.org](http://jdmdh.episciences.org) |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

Пресс-релизы Сената. Политический анализ, 18 (1): 1–35.

Гаджибекова, Л. (2013). Категории базового уровня: обзор. Журнал информатики, 39 (5): 676–687. Хасан, Р. (1967). Лингвистика и изучение художественных текстов. Этюды лингвистических аппликаций, 5: 106–121. Гербелот, А. (2013). Что есть в тексте, чего нет и какое это имеет отношение к лексической семантике. В производстве

10-й Международной конференции по вычислительной семантике (IWCS 2013) - Краткие статьи, страницы 321–327. Гербелот, А. и Ганесалингам, М. (2013). Измерение семантического содержания в векторах распределения. В материалах 51-го ежегодного собрания Ассоциации компьютерной лингвистики (Том 2: Краткие статьи), стр.

440–445.

Хьёрланд, Б. (1998). Теория и метатеория информатики: новая интерпретация. Журнал документации.

Хоуи, М. (2012). Лексическая подготовка: новая теория слов и языка.

Хогенрад, Р., Маккензи, Д.П., и Пеладо, Н. (2003). Сила и влияние в контент-анализе: производство новых социальных знаний. Качество и количество, 37(3):221–238.

Джейкобс, AM (2018). Корпус английской поэзии Гутенберга: образцовый количественный анализ повествования. Границы цифровых гуманитарных наук, 5:5.

Янике,¨ С., Францини, Г., Чима, М.Ф., и Шойерманн, Г. (2015). О близком и дистанционном чтении в цифровых гуманитарных науках: обзор и будущие задачи. В EuroVis (STAR), страницы 83–103.

Джокерс, М.Л. и Мимно, Д. (2013). Важные темы в литературе 19 века. Поэтика, 41 (6): 750–769. Келер, А. (2004). Темы дискурса, темы предложений и связность. Теоретическая лингвистика, 30 (2-3): 227–240. Кимберли, А.Н. (2011). Руководство по контент-анализу, 2-е издание. SAGE PUBLICATIONS Incorporated. Корфиатис, Н., Стамолампрос, П., Курутанассис, П., и Сагиадинос, В. (2019). Измерение качества обслуживания от

неструктурированные данные: приложение для тематического моделирования онлайн-отзывов авиапассажиров. Экспертные системы с приложениями, 116:472–486.

Криппендорф, К. (2004a). Контент-анализ: введение в его методологию (второе издание). Публикации шалфея.

Криппендорф, К. (2004b). Надежность в контент-анализе: некоторые распространенные заблуждения и рекомендации.

Исследования человеческого общения, 30(3):411–433.

Куанг, Д., Чу, Дж., и Парк, Х. (2015). Неотрицательная матричная факторизация для интерактивного тематического моделирования и кластеризации документов. В алгоритмах частичной кластеризации, страницы 215–243. Спрингер.

Кун, К.Д. (2018). Использование моделирования структурных тем для выявления скрытых тем и тенденций в отчетах об авиационных происшествиях.

Транспортные исследования, часть C: Новые технологии, 87: 105–122.

Лайтинен, Л. (2004). Грамматизация и стандартизация. Типологические исследования языка, 59: 247–262. Ламирель, Дж.-К., Чен, Ю., Кушак, П., Аль-Шехаби, С., Дуге, Н., и Лю, З. (2020). Обзор истории

Наука о науке в Китае на основе использования библиографических данных и данных цитирования: новый метод анализа, основанный на кластеризации с максимизацией признаков и контрастными графами. Наукометрия, 125:2971–2999.

Лампах, Н. и Дьевре, А. (2018). Обратите внимание на международные суды: метод анализа текста. Доступно по адресу SSRN 3251186.

Лау, Дж. Х., Ньюман, Д., и Болдуин, Т. (2014). Машинное чтение чайных листьев: автоматическая оценка согласованности темы и качества модели темы. В материалах 14-й конференции Европейского отделения Ассоциации компьютерной лингвистики, стр. 530–539.

Лаубихлер, доктор медицины, Майеншайн, Дж., и Ренн, Дж. (2019). Вычислительная история знаний: вызовы и возможности. Исида, 110 (3): 502–512.

Закон, Д.С. (2016). Конституционные архетипы. Техас Л. Rev., 95:153.

Ленгерих, Б.Дж., Маас, А.Л., и Поттс, К. (2017). Модернизация распределенных вложений для графов знаний с функциональными отношениями. Препринт arXiv arXiv: 1708.00112.

Леонелли, С. (2019). Что отличает данные от моделей? Европейский журнал философии науки, 9(2):22. Леви, О. и Голдберг, Ю. (2014). Вложения слов на основе зависимостей. В материалах 52-го ежегодного

Собрание Ассоциации компьютерной лингвистики (Том 2: Краткие статьи), страницы 302–308.

Льюис, М. и Стидман, М. (2013). Комбинированная дистрибутивная и логическая семантика. Труды Ассоциации компьютерной лингвистики, 1: 179–192.

Линдстедт, Северная Каролина (2019). Структурное тематическое моделирование для социологов: краткое тематическое исследование с литературой по изучению социальных движений, 2005–2017 гг. Социальные течения, 6 (4): 307–318.

Лю Дж., Вейнерт А. и Амин С. (2019). Семантический анализ данных дорожных камер: извлечение тематического сигнала и обнаружение аномальных событий. Препринт arXiv arXiv: 1905.07332.

Лю Л., Тан Л., Донг В., Яо С. и Чжоу В. (2016). Обзор тематического моделирования и его текущих приложений в биоинформатике. SpringerPlus, 5(1):1608.

Лю, X. и Джин, М. (2020). Классификационный анализ романов Кодзи Уно с использованием тематической модели. Поведенческая метрика, 47(1):189–212.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 25 | [http://jdmdh. episciences.org](http://jdmdh.episciences.org) |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

Лю Ю., Лю З., Чуа Т.-С. и Сунь М. (2015). Тематические вложения слов. В материалах конференции AAAI по искусственному интеллекту, том 29.

Ливермор, Массачусетс, Ридделл, А., и Рокмор, Д. (2016). Формирование повестки дня и Верховный суд США: подход тематической модели. Обзор закона Аризоны, 1 (2).

Макнамара, Дж. Р. и соавт. (2005). Анализ медиа-контента: его использование, преимущества и передовая методология. Азиатско-Тихоокеанский журнал по связям с общественностью, 6 (1): 1.

Мэдсен, Р.Э., Каучак, Д., и Элкан, К. (2005). Моделирование всплеска слов с использованием распределения Дирихле. В материалах 22-й международной конференции по машинному обучению, стр. 545–552.

Маркман, А.Б. и Вишневски, Э.Дж. (1997). Похожие и разные: дифференциация категорий базового уровня.

Журнал экспериментальной психологии: обучение, память и познание, 23 (1): 54.

Мэйринг, П. (2004). Качественный контент-анализ. У фон Кардорфа Э., Флика У. и Стейнке И., редакторов, A Companion to Quality Research, стр. 159–176.

Макфарланд, Д.А., Рэймидж, Д., Чуанг, Дж., Хир, Дж., Мэннинг, К.Д., и Джурафски, Д. (2013). Дифференциация использования языка с помощью тематических моделей. Поэтика, 41(6):607 – 625. Тематические модели и культурология.

Медхат, В., Хассан, А., и Кораши, Х. (2014). Алгоритмы и приложения анализа настроений: обзор. Инженерный журнал Айн Шамс, 5(4):1093–1113.

Митрофанова О. и Седова А. (2017). Моделирование темы в параллельных и сопоставимых художественных текстах (на примере английской и русской прозы). В материалах международной конференции IMS-2017, стр. 175–180.

Моа, Б. и Росс, С. (2019). Аналитика больших данных для мультимасштабного чтения. В книге «Больше цифровых гуманитарных наук», страницы 199–236. Рутледж.

Моретти, Ф. (2013). Дистанционное чтение. Версо Книги.

Моллерс,¨ К., Шадрова, А., и Вендел, Л. (2021). БВерфГЭ-Корпус, v.1.0.

https://doi.org/10.5281/zenodo.4551408.

Наварро-Колорадо, Б. (2018). О моделировании поэтической темы: извлечение тем и мотивов из корпуса испанской поэзии. Границы цифровых гуманитарных наук, 5:15.

Панагис, Ю., Кристенсен, М.Л., и Садл, У. (2016). В дополнение к темам: Использование тематического моделирования для изучения динамического прецедентного права международных судов. В JURIX, страницы 161–166.

Паничелла, А. (2021). Систематическое сравнение основанных на поиске подходов к настройке гиперпараметров LDA.

Информационные и программные технологии, 130:106411.

Пардо, Дж. С. (2006). О фонетической конвергенции при разговорном общении. Журнал Акустического общества Америки, 119 (4): 2382–2393.

Питерс, М.Е., Нойманн, М., Ийер, М., Гарднер, М., Кларк, К., Ли, К., и Зеттлемойер, Л. (2018). Глубокие контекстуализированные представления слов. Препринт arXiv arXiv: 1802.05365.

Петтерссон, Б. (2002). Семь тенденций в последних темах и тематическое исследование. In Louwerse, M. и van Peer, W., редакторы, Thematics. Междисциплинарные исследования, страницы 237–252. Издательство Джона Бенджаминса.

Пиантадоси, ST (2014). Закон частотности слов Ципфа в естественном языке: критический обзор и дальнейшие направления.

Психономический бюллетень и обзор, 21 (5): 1112–1130.

Пикеринг, М.Дж. и Браниган, Х.П. (1998). Представление глаголов: данные синтаксического прайминга в языковом производстве. Журнал памяти и языка, 39 (4): 633–651.

Пьерумбер, Дж. Б. (2012). Порывистость глаголов и производных существительных. В «Сыграем ли мы в Festschrift Game?», стр. 99–115. Спрингер.

Писаревская А., Леви Н., Шолтен П. и Янсен Дж. (2020). Картирование миграционных исследований: эмпирический анализ становления области исследований. Миграционные исследования, 8(3):455–481.

Рэймидж, Д., Мэннинг, К.Д., и Дюме, С. (2011). Частично помеченные тематические модели для интерпретируемого анализа текста. В материалах 17-й международной конференции ACM SIGKDD по открытию знаний и интеллектуальному анализу данных, страницы 457–465.

Рамирес, Э. Х., Брена, Р., Магатти, Д., и Стелла, Ф. (2012). Проверка тематической модели. Нейрокомпьютинг, 76(1):125–133.

Ребер, У. (2019). Преодоление языковых барьеров: оценка потенциала машинного перевода и тематического моделирования для сравнительного анализа многоязычных корпусов текстов. Методы и меры коммуникации, 13(2):102–125.

Роди, Л.М. (2012). Тематическое моделирование и образный язык. Журнал цифровых гуманитарных наук, 2 (1): 19–35.

Райс, Д. Р. (2017). Подразделения по вопросам и принятие решений Верховным судом США. Журнал политики, 79 (1): 210–222.

Риммон-Кенан, С. (1995). Что такое тема и как мы к ней подходим? В Bremond, C., Landy, J. и Pavel, T., редакторы, Thematics: New Approaches, страницы 9–20. СУНИ Пресс.

Робертс, М. Е., Стюарт, Б. М., и Тингли, Д. (2019). stm: пакет R для структурных тематических моделей. Журнал статистического программного обеспечения, 91 (2): 1–40.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 26 | [http://jdmdh.episc iences.org](http://jdmdh.episciences.org) |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

Робертс, М.Е., Стюарт, Б.М., Тингли, Д., Лукас, К., Ледер-Луис, Дж., Гадарян, С.К., Альбертсон, Б., и Рэнд, Д.Г. (2014). Структурные тематические модели для открытых ответов на опросы. Американский журнал политических наук, 58(4):1064–1082.

Рош, Э., Мервис, С.Б., Грей, В.Д., Джонсон, Д.М., и Бойс-Брэм, П. (1976). Основные объекты в естественных категориях. Когнитивная психология, 8(3):382–439.

Шлобински, П. и Шутце¨-Кобурн, С. (1992). По теме темы и преемственности темы. Лингвистика, 30(1):89–122.

Шмидлин, Р. (2011). Die Vielfalt des Deutschen: Standard und Variation: Gebrauch, Einschatzung¨ und Kodi-fizierung einer plurizentrischen Sprache, том 106. Вальтер де Грюйтер.

Шмидт, Б.М. (2012). Одни слова: Разбор тематических моделей в гуманитарных науках. Журнал цифровых гуманитарных наук, 2 (1): 49–65.

Шох, ¨ С. (2017). Жанр тематического моделирования: исследование французской классической и просветительской драмы. ДКВ:

Цифровые гуманитарные науки Ежеквартально, 11 (2).

Шадрова, А. (2020). Измерение ограничений совместного выбора в учебных корпусах: подход на основе графов. Кандидатская диссертация, Университет имени Гумбольдта в Берлине.

Шадрова А. (в печати). Дело может быть в структуре, а не в сочетаниях: графовые метрики как альтернатива статистическим показателям в корпусно-лингвистических исследованиях. В графических технологиях в гуманитарных науках 2020.

Си, Дж., Ли, К., Цянь, Т., и Дэн, X. (2014). Группировка пользователей по интересам из онлайн-обзоров на основе частоты и порядка тем. Всемирная паутина, 17(6):1321–1342.

Зиг, К. (2019). Тематическое моделирование фон Fallgeschichten. Zeitschrift fur¨ Literaturwissenschaft und Linguistik, 49 (4): 653–671.

Соллорс, В. (2002). Тематика сегодня. В Louwerse, M. и van Peer, W., редакторы, Тематика: междисциплинарные исследования, том 3, страницы 217–235. Издательство Джона Бенджаминса.

Соуза, МАР д., Уолл, М.Л., Тулер, А.С. MC, Лоуэн, IMV, и Перес, AM (2018). Использование программного обеспечения IRAMUTEQ для анализа данных в качественных исследованиях. Revista da Escola de Enfermagem da USP, 52.

Спир, Р. и Лоури-Дуда, Дж. (2017). Conceptnet at semeval-2017 Задача 2: Расширение встраивания слов с помощью многоязычных реляционных знаний. Препринт arXiv arXiv: 1704.03560.

Стилс, Л. и Леч, М. (2006). Перспективное выравнивание на пространственном языке. Препринт arXiv cs/0605012. Стерлинг Дж., Йост Дж. Т. и Хардин К. Д. (2019). Либеральные и консервативные представления о хорошем обществе: A

(социальный) структурный подход к тематическому моделированию. Sage Open, 9 (2): 2158244019846211.

Тангерлини, Т.Р. и Леонард, П. (2013). Траление в море Великого Непрочитанного: Подкорпусное тематическое моделирование и гуманитарные исследования. Поэтика, 41 (6): 725–749.

Тома С., Реттингер А. и Бот Ф. (2017). На пути к целостному представлению понятий: внедрение реляционных знаний, визуальных атрибутов и семантики распределенных слов. В Международной семантической веб-конференции, страницы 694–710. Спрингер.

Андервуд, Т. (2014). Теоретизируя исследовательскую практику, мы забыли теоретизировать двадцать лет назад. Представления, 127 (1): 64–72.

Андервуд, Т. (2017). Родословная дистанционного чтения. DHQ: Digital Humanities Quarterly, 11 (2).

ван Дер Хофт, Дж.Дж., Ванди, Дж., Барретт, член парламента, Берджесс, К.Э., и Роджерс, С. (2016). Тематическое моделирование для нецелевого исследования подструктуры в метаболомике. Труды Национальной академии наук, 113 (48): 13738–13743.

Ван Пир, В. (2002). Откуда берутся литературные темы? В Louwerse, M. и van Peer, W., редакторы Thematics:

Междисциплинарные исследования, том 3, страницы 253–263. Издательство Джона Бенджаминса.

Ван, К. и Блей, Д.М. (2011). Совместное тематическое моделирование для рекомендации научных статей. В материалах 17-й международной конференции ACM SIGKDD по открытию знаний и интеллектуальному анализу данных, страницы 448–456.

Ван, Ю., Бауэрс, А.Дж., и Фикис, Д.Дж. (2017). Автоматизированный анализ интеллектуального анализа текстовых данных пятидесятилетней литературы по исследованию лидерства в сфере образования: вероятностное тематическое моделирование статей EAQ с 1965 по 2014 год. Ежеквартальное управление образованием, 53 (2): 289–323.

Вейтин, Т. и Хергет, К. (2017). Фалькентопикс. Zeitschrift fur¨ Literaturwissenschaft und Linguistik, 47 (1): 29–48. Вендель Л., Шадрова А. и Тишбирек А. (в печати). От смоделированных тем к областям права: сравнительный

анализ видов производства в Федеральном конституционном суде Германии. Немецкий юридический журнал.

Весслен, Р. (2018). Компьютерный анализ текста для социальных наук: тематические модели и не только. Препринт arXiv arXiv: 1803.11045.

Ян, Т.-И., Торгет, А., и Михалча, Р. (2011). Моделирование темы на исторических газетах. В материалах 5-го семинара ACL-HLT по языковым технологиям для культурного наследия, социальных и гуманитарных наук, страницы 96–104.

Янг, Д. Т. (2012). Как вы измеряете конституционный момент: использование алгоритмического тематического моделирования для оценки

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 27 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |

Теория конституционных изменений Брюса Акермана. Йельский ЖЖ, 122:1990.

Янг, Р.Л. (2019). Язык и риторика позитивных действий: анализ модели структурной темы сводок amicus Верховного суда. Кандидатская диссертация, Университет Айовы.

Ю Л., Чжан К., Шао Ю. и Цуй Б. (2017). LDA\*: надежная и крупномасштабная система тематического моделирования. Труды Фонда VLDB, 10 (11): 1406–1417.

Чжун Ю., Чжу К. и Чжан Л. (2015). Классификация сцен на основе многофункциональной вероятностной тематической модели слияния для изображений дистанционного зондирования с высоким пространственным разрешением. IEEE Transactions on Geoscience and Remote Sensing, 53(11):6207–6222.

Ципф, ГК (1965). Психобиология языка. Введение в динамическую филологию. 1935. Кембридж, Массачусетс: MIT Press.

|  |  |  |
| --- | --- | --- |
| Журнал интеллектуального анализа данных и цифровых гуманитарных наук | 28 | <http://jdmdh.episciences.org> |
| ISSN 2416-5999, журнал открытого доступа. |  |  |