Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
«Прикладная математика и информатика»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
Программный проект на тему
«Построение системы тематического моделирования на основе аннотаций научных статей»
« Building a mathematical modeling system based on abstracts of scientific articles»

Выполнила студентка группы БПМИ188, 4 курса,
Квиндт Ева Сергеевна

Руководитель ВКР:
Младший научный сотрудник МЛ ИССА,
Паринов Андрей Андреевич

Москва, 2022

Annotation

We make a detailed review of the algorithms of some modern approaches to topic modeling: LSA, LDA, Top2Vec, BERTopic, lda2vec, Contextualized Contextualized Topic Models & Kitty. The simplicity of the implementation of all methods and their results were compared on the dataset of abstracts of scientific articles from arxiv.org. It is proposed to use the results of mathematical modeling in the future.

Аннотация

Мы делаем подробный обзор алгоритмов некоторых существующих методов тематического моделирования: LSA, LDA, Top2Vec, BERTopic, lda2vec, Contextualized Topic Models & Kitty. Уделяется внимание скорости работы, количеству и качеству выделяемых тем, а также осуществлена попытка рассчитать метрики для некоторых из них. Сравнение моделей было произведено на наборе данных, содержащем аннотации научных статей с arxiv.org. Предложено использование результатов тематического моделирования в дальнейшем.

Ключевые слова

Эмбеддинги; тематическое моделирование; анализ текста; кластеризация; рекомендательная система

Содержание

1. Введение

Текстовая аналитика находится в активном развитии. К задачам этой сферы относится проблема разбиения корпуса текстов на подгруппы (классы) для дальнейшей работы с ними; выявление тем, упомянутых в текстах. Результаты такой работы могут быть использованы при построении рекомендательной системы, в продуктовой аналитике с нестандартными объектами и создания собственной таксономии.

Методы тематического моделирования также могут быть использованы для решения проблемы классификации текстов. Они определяют какие темы представлены в документе и не требуют знания об имеющихся темах, в отличие от задач классификации. Соответственно, данные методы относятся к классу задач с неконтролируемым обучением. Мотивацией к выбору тематического моделирования вместо тематической классификации является дальнейшее желание работать с корпусами текстов без знания об имеющихся классах, на которые можно разбить данные, а также предположении о том, что текст содержит в себе не одну тему, а представлен набором разных с разной степенью содержания.

Алгоритмы тематического моделирования создают коллекции выражений и слов, которые, по их мнению, связаны, позволяя выяснить, что означают эти отношения. Включает подсчет слов и группировку похожих шаблонов слов для определения тем в неструктурированных данных.

Целью данной работы является наиболее эффективное решение задачи тематического моделирования на основе аннотаций научных статей с ресурса arxiv.org. Во внимание берутся время работы модели, количество выделяемых тем и возможность корректирования этого числа, показатели известных метрик для оценки качества смоделированных тем и наличие встроенных функций и методов для удобства визуальной оценки результатов. В первой части работы описывается используемый датасет и его предобработка. Далее рассматриваются имеющиеся современные подходы и методы решения данной задачи, сравнение

их устройства и результатов на основе выбранных данных. Затем предлагаются методы использования полученных результатов и, в частности, проводится обзор существующих подходов к построению рекомендательных систем. В конце сделан вывод о полученных результатах и описаны дальнейшие пути использования моделей для различных сфер.

2. Обзор литературы

На данный момент существует множество работ [2] [3] [4] [8] [10], посвященных тематическому моделированию. Чаще всего это статьи о разработке нового метода и колоссальных результатах, которые он показывает в сравнении с уже известными подходами. Проблема заключается в том, что для задачи тематического моделирования не существует универсального размеченного датасета, на котором можно было бы оценить любую модель. Качество результата всегда зависит от предложенных данных и предметной области.

Важной работой в этой сфере является публикация 2013-ого года Воронцова К.В. [7], из которой было взято множество теоретических идей о тематическом моделировании и понятий. В ней описывается задача тематического моделирования, предложены гипотезы, на основе которых строятся вероятностные модели.

В процессе обзора методов тематического моделирования также были использованы интернет-источники, описывающие математическую логику, скрывающуюся за архитектурой каждого метода, а также примеры их имплементации для различных данных. Например, работа 2003-его года от создателей LDA [2] или более новая публикация от 2020-ого года об устройстве Top2Vec [4]. Уделялось внимание также статьям со сравнениями разных моделей [3]. Подробнее они будут рассмотрены в дальнейших главах.

3. Данные

Для сравнения методов тематического анализа текстов использовался датасет с одного из соревнований на Kaggle, содержащий более 2 миллионов аннотаций научных статей на английском языке с сайта arXiv.org. Датасет содержит в себе авторов статьи, заголовок, аннотацию, doi, информацию о принадлежности к тематике по таксономии arXiv.

Перед применением методов тематического моделирования, датасет был очищен от пунктуации, чисел и стоп-слов. Имеет место применение лемматизации к словам для обучения вероятностных тематических моделей, которые рассматривают текст как мешок слов. Так как мы имеем дело с корпусом на английском языке, подойдет и стемминг – упрощенная лемматизация, отбрасывающая концы слов. Был выбран PorterStemmer, как один из распространенных. Регистр в данной задаче не имеет значения и может отразиться на результате в худшую сторону, поэтому все слова переведены в нижний регистр. Для реализации вышеперечисленных предобработок были использованы стандартные средства библиотеки pandas.

В результате получен предобработанный датасет. Элементы словаря теперь могут называться «терминами», так как это не всегда осмысленные слова, но в том числе и их части, формы.

Для более быстрых вычислений в Google Collab выбрано 20000 рандомных строк.

4. Тематическое моделирование

Все методы тематического моделирования можно разделить на несколько категорий, в зависимости от их архитектуры. Одни используют представление документа как мешка слов с искомым распределением слов в теме и тем в документе. Другие, так называемые нейросетевые тематические модели,

используют современные трансформеры, векторное представление слов и документов. Есть также и комбинации этих двух классов.

Далее рассматриваются одни из самых известных методов тематического моделирования, описывается их архитектура, настройка гиперпараметров. Также приведены результаты тестирования на выбранных данных.

4.1. Latent Semantic Analysis

LSA основан на дистрибутивной гипотезе, которая предполагает, что семантика двух слов будет схожей, если они встречаются в схожих контекстах [3]. Метод подсчитывает частоту каждого слова в пределах одного документа и в пределах корпуса, и, используя TF-IDF, строит матрицу документ - слово. TF-IDF для терма (слова) x в документе y будет вычисляться по формуле (1).

$$W_{x,y} = tf_{x,y} * log \frac{N}{df_x}$$

где $tf_{x,y}$ – частота слова x в документе y, N – общее число документов, df_x – количество документов, содержащих x. Используя SVD полученная матрица раскладывается в произведение трех: матрица U документ - слово, диагональная матрица S и матрицу V слово-тема.

Устанавливая t в качестве гиперпараметра и перемножив первые t столбцов U и первые t строк V с наибольшим значением t сингулярным значением S мы получаем t наиболее часто встречающихся найденных тем.

LSA предполагает, что аналогичные документы будут иметь примерно одинаковое распределение частот слов. Однако, данный метод утрачивает информацию о порядке слов в предложении и множественности смысла слов. Здесь используется гипотеза о «мешке слов» [7], основанная на предположении, что тематику документа можно узнать даже при произвольной перестановке слов в документе. То есть в этом предположении тематика документа определяется не столько смысловыми взаимосвязями между предложениями и словами, как долей содержания тех или иных терминов в тексте.

4.2. Latent Dirichlet Allocation

LDA впервые представлен в 2003 году в статье [2]. Данный метод предполагает, что распределение тем в документе и распределение слов в темах являются распределениями Дирихле. Каждому документу корпуса сопоставляется набор тем, которые охватывают большую часть слов в этом документе.

Из распределения Дирихле $Dir(\alpha)$ мы рисуем случайную выборку, представляющую распределение тем θ или смесь тем конкретного документа. Из θ мы выбираем конкретную тему Z на основе распределения. Затем, из другого распределения Дирихле $Dir(\beta)$, мы выбираем случайную выборку, представляющую распределение слов φ темы Z. Из φ мы выбираем слово w. Про устройство LDA можно почитать подробнее в статье [2].

Как и предыдущий метод, LDA представляет документы в виде мешка слов, что влечет за собой игнорирование синтаксической информации, порядка слов и многозначности термов.

Результат алгоритма представляет собой вектор, содержащий охват каждой темы для моделируемого документа, где i-ое значение показывает охват документом i-ой темы.

Чтобы получить представления о сходстве документов или распределении тем по корпусу документов можно использовать, например, косинусное сходство.

4.3. Top2Vec

Помимо алгоритмов тематического моделирования стали разрабатываться готовые пакеты для данной задачи. Под их архитектурой зачастую скрывается готовый пайплайн, включающий предобработку датасета, применение современных языковых моделей, кластеризация полученных векторов и попытка построить вероятностное распределение тем и терминов.

Например, в статье от августа 2020 года «Top2Vec: Distributed Representations of Topics» [4] описывается новая техника тематического моделирования, которая, используя известные модели эмбеддинга слов, генерирует совместно встроенные векторы темы, документа и слова.

После применения Doc2Vec и уменьшения размерности пространства при помощи UMAP, данный метод кластеризует набор документов, используя HDBSCAN. Top2Vec пытается найти скопления документов, а затем определить, какие слова объединяют эти документы вместе. Каждая плотная область – это тема, ее центр тяжести документов - вектор темы, а n ближайших векторов слов к нему – слова темы.

Top2Vec прост в использовании. Модель не требует предобработки текста, осуществляя ее самостоятельно и автоматически находит количество тем, а также имеет много встроенных функций для визуализации. Результирующие векторы тем совместно встроены в векторы документа и слова с расстоянием между ними, представляющим семантическое сходство. Недостаток этого метода заключается в том, что остаются элементы корпуса, не относящиеся к темам.

4.4. BERTopic

BERTopic – сравнительно новая техника тематического моделирования, разработанная в 2020 году [5]. В процессе обзора выяснилось, что она очень схожа по своему устройству с Top2Vec.

На первом этапе BERTopic создает векторное представление документов, используя предварительно обученные языковые модели на основе трансформеров. Затем происходит снижение размерности пространства при помощи UMAP и кластеризация семантически похожих кластеров с HDBSCAN. В конце извлекаются термины для репрезентации тем на основе расчёта class-based TF-IDF. Для класса i и слова t это значение будет равно:

$$c - TF - IDF_i = \frac{t_i}{w_i} * \log\left(1 + \frac{m}{\sum_j^n t_j}\right)$$

где t_i частота слова t в теме i, w_i – общее количество слов, m – среднее число слов в теме. Данное значение «важности» для каждого термина в кластере используется для создания темы.

BERTopic позволяет сокращать количество выделенных тем через вычисление матрицы c-TF-IDF документов и итеративного объединения наименее часто встречающейся темы с наиболее похожей на основе их матриц c-TF-IDF. Сгенерированные темы могут быть иерархически упорядочены и визуализированы, что также помогает сократить количество тем до релевантного значения.

BERTopic поддерживает управляемое, полуконтролируемое и динамическое тематическое моделирование, о которых можно прочитать подробнее в статье [5]. Недостатки метода – возможное наличие документов, не привязанных ни к одной из тем при использовании HDBScan кластеризации и сведение задачи тематического моделирования к задаче кластеризации.

4.5. Lda2vec

lda2vec – расширение word2vec и LDA, которое совместно обучает векторы слова, документа и темы.

lda2vec строится поверх skip-gram word2vec для генерации векторов слов, но для прогнозирования используется контекстный вектор, которые вычисляется как сумма двух других: вектор слова и вектор документа. Первый создается при помощи word2vec, второй – взвешенная комбинация вектора веса документа, представляющего “веса” (которые позже будут преобразованы в проценты) каждой темы в документе, и матрицы тем, представляющей каждую тему и ее соответствующее векторное вложение. Подробнее архитектуру lda2vec можно рассмотреть на Рисунке 1.

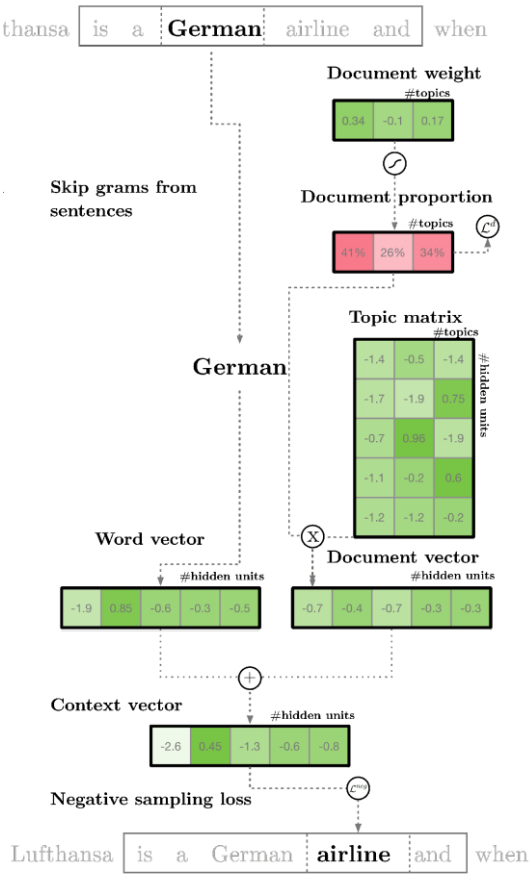


Рисунок 1. Архитектура Lda2vec

Сила lda2vec заключается в том, что он одновременно обучает представления тем и представления документов.

Этот метод был представлен в 2016 году [6] и несмотря на любопытную попытку смешанного представления документа и темы даже автор модели советует подходить к выбору lda2vec с осторожностью: «…если вы хотите переработать свои собственные тематические модели, которые совместно соотносят темы статьи с голосами или прогнозируют темы для пользователей, тогда вас может заинтересовать lda2vec […] Требует огромного объема вычислений, и поэтому я бы на самом деле не стал применять его без графических процессоров. Кроме того, я не измерял производительность lda2vec по сравнению с базовыми показателями LDA и word2vec».

Поэтому в практической части, lda2vec не рассматривался из-за своей ненадежности и вычислительной сложности.

4.6. Combined Topic Model

В 2020 году группой разработчиков из Италии было представлено семейство тематических моделей Contextualized Topic Models (CTMs) [8], сочетающих в себе предобученные представления естественного языка и мешок слов.

Данное семейство представлено двумя моделями: Combined Topic Model (CombinedTM) и ZeroShot Topic Model (ZeroShotTM) – и модулем Kitty для быстрой классификации документов и создания именованных кластеров. Вторая модель, использует только контекстуальные эмбеддинги, умеет обрабатывать слова, которые не представлены на этапе обучения, а также может быть использована для межъязыкового тематического моделирования. Например, дан маленький корпус на языке А в определенной предметной области, для которого нужно решить задачу тематического моделирования. Одновременно мы имеем корпус большего размера для распространенного языка Б (например, английский) в этой же предметной области. ZeroShotTM позволяет провести обучение для более репрезентативного и объемного набора данных на языке Б и перенести результат на язык А.

В данной работе использовалась тематическая модель CombinedTM. Она сочетает в себе нейронную тематическую модель ProdLDA и эмбеддинг SBERT. Работает одновременно с контекстуальным эмбеддингами и представлением текста в виде мешка слов. Детальней об архитектуре данного метода можно прочитать [10].

Накладывается ограничение на длину документа – она не должна превышать установленной в SBERT длины предложения. Иначе остальная часть документа будет отрезана.

Преимущество данного семейства моделей – возможность использовать различные эмбеддинги. Однако, разработчики упоминают, что модели работают лучше, когда размер словаря в представлении текста в виде мешка слов ограничен количеством в 2000 элементов. Это связано с количеством параметров

встроенной нейросетевой модели, обрабатывающей мешок слов. Данное ограничение не является строгим и при правильной предобработке данных, можно его превысить.

5. Оценка методов тематического моделирования

Сложность задач тематического моделирования заключается в оценивании качества полученных моделей, так как зачастую имеем дело с классом задач с неконтролируемым обучением.

В данном случае за целевую переменную могла быть взята характеристика category. Однако мы задались целью научиться применять методы тематического моделирования к задачам с неразмеченными данными. Это может быть полезно для случая «холодного старта», когда появляется новый объект, не отнесенный ни к одной из категорий. Другое применение – использование тематических моделей для сервисов, на которых объект может относится к нескольким темам одновременно и выбор тем задается автоматически, без вмешательства пользователя. Также такая постановка задачи может быть полезна для построения своей собственной таксономии для неразмеченного набора текстов.

Для создания рекомендательных систем не на основе тегов, а по смысловому содержанию объекта, полезно научиться улавливать более тонкие взаимосвязи между объектами, помимо принадлежности к категории по заданной таксономии, в чем также поможет тематических анализ текста.

В случае с неразмеченными данными нет точного решения, все зависит от желаемого результата и представлении о задаче. Нет такого списка тем, подходящего для любого корпуса. Однако, метрики могут дать представление о том, как могут изменяться темы в зависимости от выбора алгоритма и значения параметров, предоставляя разработчику самостоятельно определять подходящую конфигурацию параметров для поставленной задачи.

Естественный язык наполнен множественными смыслами и включает в себя субъективные интерпретации, поэтому если мы работаем с корпусом, основанном на маленьком словаре и включающем в себя структурированные тексты (например, юридические), то оценить модель проще. Существует несколько подходов и метрик для оценки качества тематических моделей. Рассмотрим некоторые из них и попробуем выбрать наиболее релевантные для данной задачи.

5.1. Визуальная оценка

Самый очевидный способ оценить модель – визуальная оценка результатов. Обратить внимание нужно на количество выделенных тем, список терминов, представляющих каждую тему и вероятность их принадлежности к данной теме.

В качественной модели темы будут представлены осмысленными терминами, семантически связанными друг с другом в пределах одной темы и имеющими высокие вероятности отношения к ней. При этом между темами термины должны быть семантически далекими. Также стоит обратить внимание на распределение и размеры тем.

Очевидный недостаток данного метода – отсутствие какого-либо математического обоснования, то есть субъективность оценки. Чтобы придать формальности можно оценивать интерпретируемость темы, то есть насколько сложно эксперту по 10 топ-словам темы дать ей подходящее название. Однако, привлечение экспертов для данной задачи затруднительно, так как тексты относятся широкому спектру областей.

Визуальная оценка имеет место быть, так как позволяет определить заведомо некачественную или неподходящую для поставленной цели модель, не прибегая к вычислению метрик.

5.2. Перплексия

Весь датасет делится на обучающую и тестовую выборки. Модель представлена матрицей тем и гиперпараметром альфа для распределения тема-

документ. Рассчитывается логарифмическая вероятность нерассмотренных документов из тестового набора.

$$L(w) = \log p(w\,|\Phi, \alpha) = \sum_{d} \log p(w_d|\Phi, \alpha)$$

Мерой для оценки тематической модели будет являться величина

$$perplexity(test\ set\ w) = \exp\{\frac{-L(w)}{count\ of\ tokens}\}$$

Более низкое значение данной меры подразумевает лучшее качество модели.

Данная метрика подходит для оценки вероятностных тематических моделей (например, LDA, LSA), так как использует вероятностное распределение текст-тема. Например, перплексию затруднительно рассчитать для модели Top2Vec, так как в ней не задано распределение, а есть только ранжирование слов и расстояния.

5.3. Мера согласованности

В лингвистике существует понятие когезии (связности) текста с помощью повторов, синонимов, кореферентности компонентов. Меры согласованности (когерентность) темы основаны на этом понятии и оценивают каждую тему путем измерения степени семантического сходства между словами, получившими высокие оценки в теме, а также частоту встречаемости терминов вблизи друг друга. Данная оценка позволяет выявлять среди всех тем, полученных путем статистического вывода, семантически интерпретируемые темы, то есть те, которые действительно несут в себе информацию о естественном языке. Качественной считается та модель, для которой мера согласованности тем высока, то есть слова, репрезентирующие тему, образуют кластеры внутри текста.

Однако, простая оценка совстречаемости топ-слов темы в тексте имеет недостаток. Топ-слова покрывают лишь малую часть документа (1-2%) и при рассмотрении темы, как списка из 3-10 слов, теряется качественная информация.

В одной из работ меры согласованности разделяют [12] на когерентносоть по топ-словам и внутритекстовую когерентность. Во второй учитывается вышеописанный недостаток и распределение темы в тексте за счет образования однородных фрагментов.

Есть несколько мер согласованности, основанных на вышеописанной идее. О них можно почитать подробнее на ресурсе [9].

5.4 Проблемы оценки тематических моделей

Как говорилось ранее, оценить тематические модели сложно, так как нет четкого представление о качественном разделении документов и результаты зависят от поставленных целей и выбранной предметной области. Первые несколько терминов не всегда лучшим образом описывают всю тему, а сама тема может быть плохо интерпретируема с точки зрения естественного языка, но быть качественной с точки зрения статистического вывода.

Исследования также показали, что зачастую оценка качества модели при помощи метрик не коррелирует с оценками визуального анализа тем и эмпирическим подбором параметров. Зачастую разработчикам приходится подбирать значения исходя из своего опыта и субъективного представления о качественных результатах. Поэтому, не стоит пренебрегать визуальной оценкой качества построенных моделей, какой бы необоснованной она не казалась.

Задачу тематического моделирования можно также свести к задаче кластеризации, если перейти от распределения тем на документе к кластерам, взяв за искомый кластер наиболее вероятную тему для документа. В таком случае можно было бы использовать более точные метрики для алгоритмов кластеризации, такие как Индекс Дэвиса-Болдина, Индекс Данна и Коэффициент

Силуэта. Однако, такую постановку задачи мы не рассматривали и остановились именно на тематическом моделировании.

6. Результаты моделей

Использование всех моделей для использования на данных не имело смысла, так как на этапе обзора устройства некоторых из них было понятно, что метод слишком прост, недоработан или не подходит для наших целей.

При сравнении тематических моделей обращалось внимание на количество выделенных тем, качество разделения (качество представления тем через N-top слов), затраченное время и удобство реализации (количество встроенных методов, возможность объединения похожих тем и визуализация).

	LSA	LDA	Top2Vec	BERTopic	CTM
Затраченное время, мин	~ 0.08	~ 11	~ 8	~ 5	~ 5
Выделено тем	5 / 27	27	148	130	20 / 100
Coherence CV	0.44 / 0.28	0.62	–	–	0.65 / 0.64
Coherence NPMI	0.02 / -0.01	0.02	–	–	0.07 / 0.07
Coherence U_MASS	-1.92 / -2.29	-5.61	–	–	-2.55 / -2.45

Таблица 1. Сравнение результатов моделей

Как и говорилось выше, метрика согласованности (когерентность) не всегда коррелирует с выводами человека о смоделированных темах. Хотя и удалось добиться достаточно высоких значений метрики (Таблица 1), осмысленных тем не получилось ни для одной из вероятностных моделей. Более интерпретируемые темы получились для моделей, основанных на эмбеддингах и смешанных подходах.

6.1. LSA

Не удалось добиться хороших результатов при использовании LSA. Была попытка максимизировать когерентность через подбор гиперпараметра. Однако при хорошем значении метрики (0,436 для 5 кластеров), осмысленных тем не получилось.

```
0.352*"model" + 0.269*"use" + 0.182*"method" + 0.178*"data" + 0.166*"propos" + 0.156*"result" + 0.145*"gener"
-0.373*"model" + -0.251*"learn" + -0.194*"data" + 0.168*"quantum" + 0.167*"system" + 0.167*"field" + -0.167*"
-0.770*"model" + 0.222*"method" + 0.194*"algorithm" + 0.176*"network" + 0.158*"problem" + 0.142*"propos" + 0.
0.288*"graph" + 0.271*"model" + 0.238*"problem" + -0.230*"data" + 0.223*"n" + -0.220*"use" + 0.202*"algorithm
-0.487*"system" + 0.445*"data" + -0.266*"quantum" + -0.170*"state" + -0.153*"network" + 0.152*"galaxi" + -0.1
```

Рисунок 2 Распределение терминов для модели LSA (num_topic = 5)

Увеличение количества тем до 30 также не дало качественных результатов. Меры согласованности снижаются, интерпретируемость тем не становится лучше, вероятности терминов низкие.

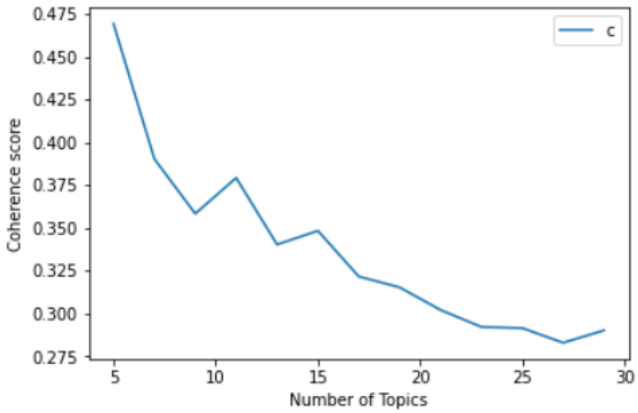


Рисунок 3. График зависимости согласованности тем от их количества для LSA

```
0.352*"model" + 0.269*"use" + 0.182*"method" + 0.178*"data" + 0.166*"propos" + 0.156
0.373*"model" + 0.251*"learn" + 0.194*"data" + -0.168*"quantum" + -0.167*"system" +
-0.770*"model" + 0.222*"method" + 0.194*"algorithm" + 0.176*"network" + 0.159*"probl
0.288*"graph" + 0.271*"model" + 0.237*"problem" + -0.230*"data" + 0.222*"n" + -0.220
-0.487*"system" + 0.445*"data" + -0.265*"quantum" + -0.171*"state" + -0.153*"network
0.344*"network" + -0.314*"data" + 0.294*"imag" + -0.266*"system" + -0.254*"algorithm
-0.571*"method" + 0.467*"data" + 0.269*"system" + 0.212*"graph" + 0.195*"network" +
0.366*"network" + -0.353*"data" + -0.294*"quantum" + -0.292*"gener" + 0.238*"graph"
-0.554*"system" + 0.441*"quantum" + 0.308*"algorithm" + -0.211*"gener" + 0.189*"stat
-0.675*"use" + 0.261*"learn" + -0.213*"graph" + 0.204*"problem" + -0.198*"quantum" +
```

Рисунок 4. Репрезентация первых 10 тем через термины для LSA (num_topic=30)

Добавим, что эксперименты также проводились на датасете, включающем биграммы (часто встречающиеся словосочетания из 2-3 слов) и это не дало также никаких улучшений (когерентность 0,380).

6.2. LDA

Для LDA также не удалось добиться хорошей репрезентации тем через термины. Данная модель лучше улавливает взаимосвязь между терминами, чем LSA и некоторым темам можно предположительно дать хорошие названия, однако на визуализации видно, что качественной модели не получилось. Темы неравномерно распределены, как и термины. Изменение количества тем отразилось на когерентности в лучшую сторону, однако осмысленности темам не придало.

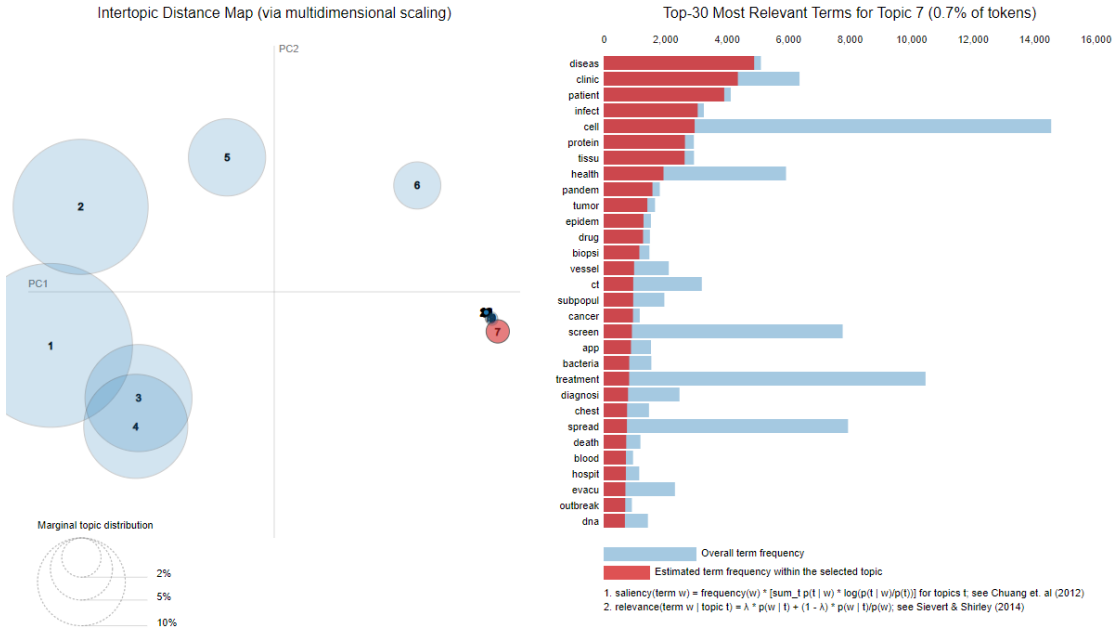


Рисунок 5. Репрезентация тем через термины для LDA

6.3. Top2Vec

Top2Vec с высокой степенью достоверности определяет сходство документов, их принадлежность к теме и строит высококачественное представление тем. Однако, затруднительно оценить модель, кроме проведения визуального анализа каждой темы. На первый взгляд, темам несложно дать названия, термины с высокой вероятностью качественно репрезентируют их, распределение тем по документам более равномерное по сравнению с другими моделями.

```
Topic 18
 ['diagnosis' 'ct' 'cancer' 'breast' 'mri' 'clinical' 'patients' 'lung'
 'tissue' 'lesions' 'segmentation' 'medical' 'tumor' 'tomography'
 'diseases' 'patient' 'imaging' 'scans' 'cardiac' 'ct mri']
Scores of first five terms:  [0.6172989  0.6071648  0.5647254  0.564204   0.56343895]

Topic 19
 ['monocular' 'scene' 'pose' 'stereo' 'cameras' 'object pose' 'object'
 'camera' 'camera pose' 'lidar' 'depth' 'scenes' 'rgb' 'monocular depth'
 'pose estimation' 'image' 'cues' 'dense depth' 'images' 'depth map']
Scores of first five terms:  [0.66800785 0.59900624 0.56286395 0.5498804  0.5470988 ]

Topic 20
 ['gpus' 'gpu' 'cpu' 'workloads' 'hardware' 'execution' 'accelerators'
 'gpu memory' 'hpc' 'cpu gpu' 'speedup' 'hardware platforms'
 'parallel processing' 'units gpus' 'hardware software' 'implementations'
 'libraries' 'latency' 'hardware accelerators' 'implementation']
Scores of first five terms:  [0.6109548 0.5863665 0.5838635 0.5429244 0.5361176]
```

Рисунок 6. Репрезентация некоторых тем через термины для Top2Vec

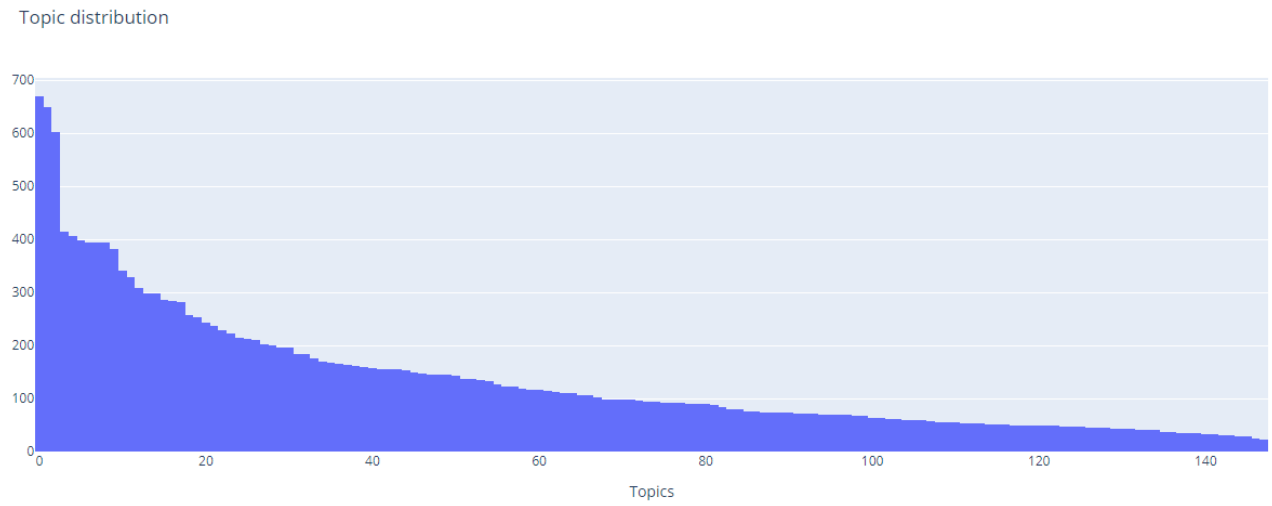
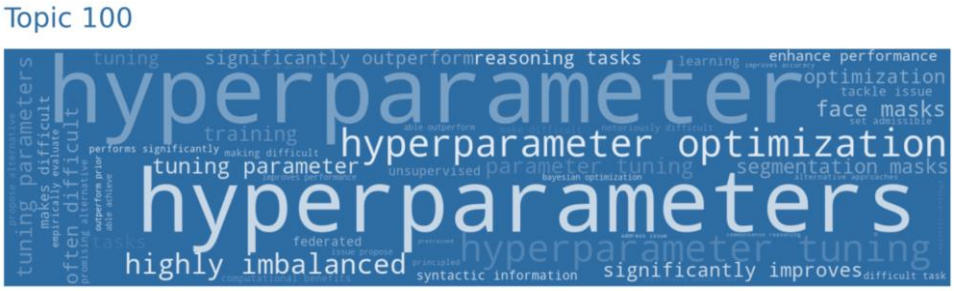


Рисунок 7. Распределение топиков по документам

Метрику когерентности рассчитать невозможно, так как модель не предоставляет матрицу распределений, а считает расстояние в векторном пространстве.

На примере Темы №100 можно видеть, что облако слов состоит из терминов, принадлежащей к одной области, они связаны по смыслу. При этом в найденных документах, относящихся к Теме №100, прослеживается данная область, а документы принадлежат к теме с высокой вероятностью (уверенность более 70%, Рисунок 8).

Также Top2Vec позволяет находить ближайшие документы к термину из словаря, темы по ключевому слову, ближайшие термины по заданному.




```
Document: 4515, Score: 0.7826732993125916
-----------
  A typical assumption in supervised machine learning is that the train
(source) and test (target) datasets follow completely the same distribution.
This assumption is, however, often violated in uncertain real-world
applications, which motivates the study of learning under covariate shift. In
this setting, the naive use of adaptive hyperparameter optimization methods
such as Bayesian optimization does not work as desired since it does not
address the distributional shift among different datasets. In this work, we
consider a novel hyperparameter optimization problem under the multi-source
covariate shift whose goal is to find the optimal hyperparameters for a target
task of interest using only unlabeled data in a target task and labeled data in
multiple source tasks. To conduct efficient hyperparameter optimization for the
target task, it is essential to estimate the target objective using only the
available information. To this end, we construct the variance reduced estimator
that unbiasedly approximates the target objective with a desirable variance
property. Building on the proposed estimator, we provide a general and
tractable hyperparameter optimization procedure, which works preferably in our
setting with a no-regret guarantee. The experiments demonstrate that the
proposed framework broadens the applications of automated hyperparameter
optimization.


Document: 19798, Score: 0.7717650532722473
-----------
  Machine learning training methods depend plentifully and intricately on
hyperparameters, motivating automated strategies for their optimisation. Many
existing algorithms restart training for each new hyperparameter choice, at
considerable computational cost. Some hypergradient-based one-pass methods
exist, but these either cannot be applied to arbitrary optimiser
hyperparameters (such as learning rates and momenta) or take several times
longer to train than their base models. We extend these existing methods to
develop an approximate hypergradient-based hyperparameter optimiser which is
applicable to any continuous hyperparameter appearing in a differentiable model
weight update, yet requires only one training episode, with no restarts. We
also provide a motivating argument for convergence to the true hypergradient,
and perform tractable gradient-based optimisation of independent learning rates
for each model parameter. Our method performs competitively from varied random
hyperparameter initialisations on several UCI datasets and Fashion-MNIST (using
a one-layer MLP), Penn Treebank (using an LSTM) and CIFAR-10 (using a
ResNet-18), in time only 2-3x greater than vanilla training.
```

Рисунок 8. Репрезентация одной из темы для Top2Vec

6.4. BERTopic

Обучение BERTopic было проведено на предобработанных данных без стемминга, для качественного встраивания эмбеддингов, и нескольких разных трансформерах (Таблица 2).

	all-MiniLM-L6-v2	albert-base-v2	all-roberta-large-v1	all-distilroberta-v1	USE	paraphrase-MiniLM-L12-v2
Время обучения	2,15 мин	4,65 мин	18,5 мин	5 мин	3,7 мин	3,1 мин
Количество выделенных тем	182	176	158	164	116	187

Таблица 2. Результаты BERTopic на разных трансформерах

Значительного различия между разными вариациями модели не наблюдается, не считая времени работы, поэтому было принято решение взять среднюю по времени обучения и количеству выделенных тем модель и попробовать улучшить ее. Далее обучение проходило на *'all-MiniLM-L12-v2'* (затраченное время обучения 166 секунд, выделено 262 темы).

Топик «-1» отвечает за документы, не отнесенные ни к одной из тем. То есть BERTopic сводит задачу тематического моделирования скорее к задаче кластеризации. Оставшиеся темы выглядят осмысленными, можно даже попробовать дать им названия. Значительного улучшения удалось добиться при использовании биграмм и триграмм. Вероятно, устойчивые словосочетания лучше репрезентируют тему, чем отдельные слова. Но в таком случае количество документов, не отнесенных ни к одной из тем, возрастает до 50%.

	Topic	Count	Name
0	-1	7909	-1_model_data_show_results
1	0	802	0_black_gravity_theory_black hole
2	1	366	1_solutions_equation_equations_solution
3	2	341	2_random_stochastic_brownian_processes
4	3	253	3_language_word_languages_translation
5	4	208	4_spaces_space_operators_banach
6	5	205	5_segmentation_images_image_medical
7	6	172	6_forecasting_time series_data_series
8	7	162	7_magnetic_spin_magnetization_antiferromagnetic
9	8	159	8_speech_speaker_audio_asr

Рисунок 9. Распределение первых 10 тем для BERTopic

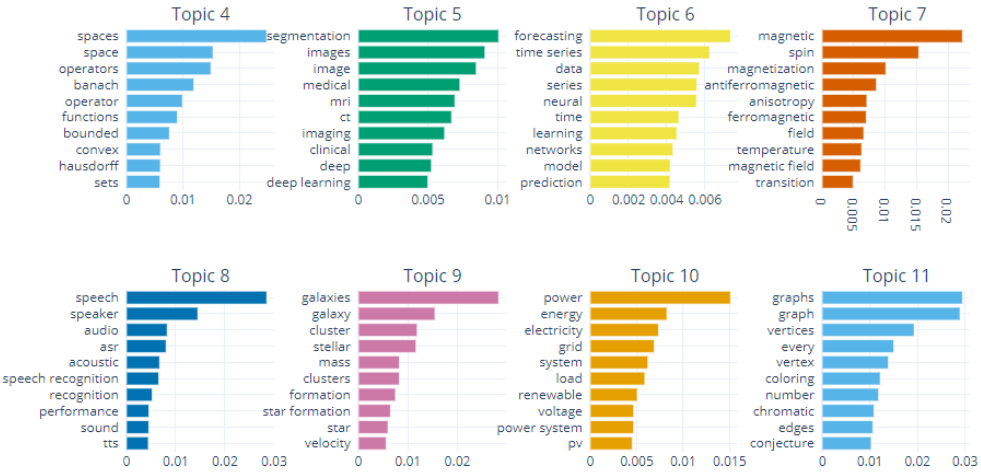


Рисунок 10. Репрезентация восьми тем для BERTopic

При обучении модели использовалось 15 топ слов для репрезентации модели. Стоит задаться вопросом, какое оптимальное число необходимо, чтобы с легкостью дать название теме по заданному количеству терминов. Однозначного ответа быть не может, так как, с одной стороны, языковые модели определяют слова, как близкие, если они однородные (это можно заметить по репрезентации тем 20 и 21). И тогда первой пары терминов будет недостаточно, нужно смотреть на большую выборку. С другой стороны, темы представлены рядом терминов, начинающихся с наиболее репрезентативного по оценке c-TF-IDF. Чем выше оценка, тем более репрезентативно слово для данной темы. Поскольку слова темы сортируются по их баллам c-TF-IDF, баллы медленно снижаются с каждым добавленным словом, т.е. в какой-то момент добавление терминов к представлению темы лишь незначительно увеличивает общий балл c-TF-IDF и не приносит пользы для его представления, темы становятся похожими между собой.

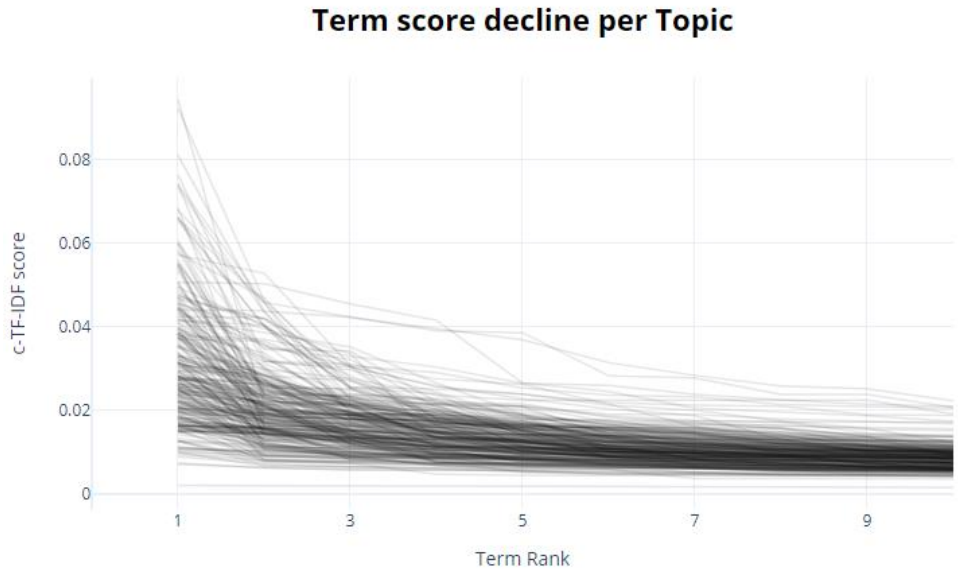


Рисунок 11. Зависимость ранга термина от его вклада в значимость темы

Для определения оптимального количества терминов для репрезентации темы на Рисунке 6.4.3. используется метод локтя. То есть в данном случае примерно первые 5-7 терминов несут в себе значимую информацию о теме, а последующие дают лишь линейный прирост.

По сравнению с другими моделями, BERTopic выделил очень много тем. Несмотря на то, что первые 20 тем легко интерпретируются, стоит посмотреть на корреляцию между оставшимися и попытаться объединить редко встречающиеся. BERTopic умеет строить матрицу корреляций, основываясь на матрице косинусного подобия между векторами тем. На ней (Рисунок 12) видно, что некоторые темы очень схожи, значит некоторые из них можно объединить. Так же на взаимосвязи между темами может указать иерархическая кластеризация тем (Рисунок 13). Исходя из визуализаций, количество тем уменьшено вдвое.

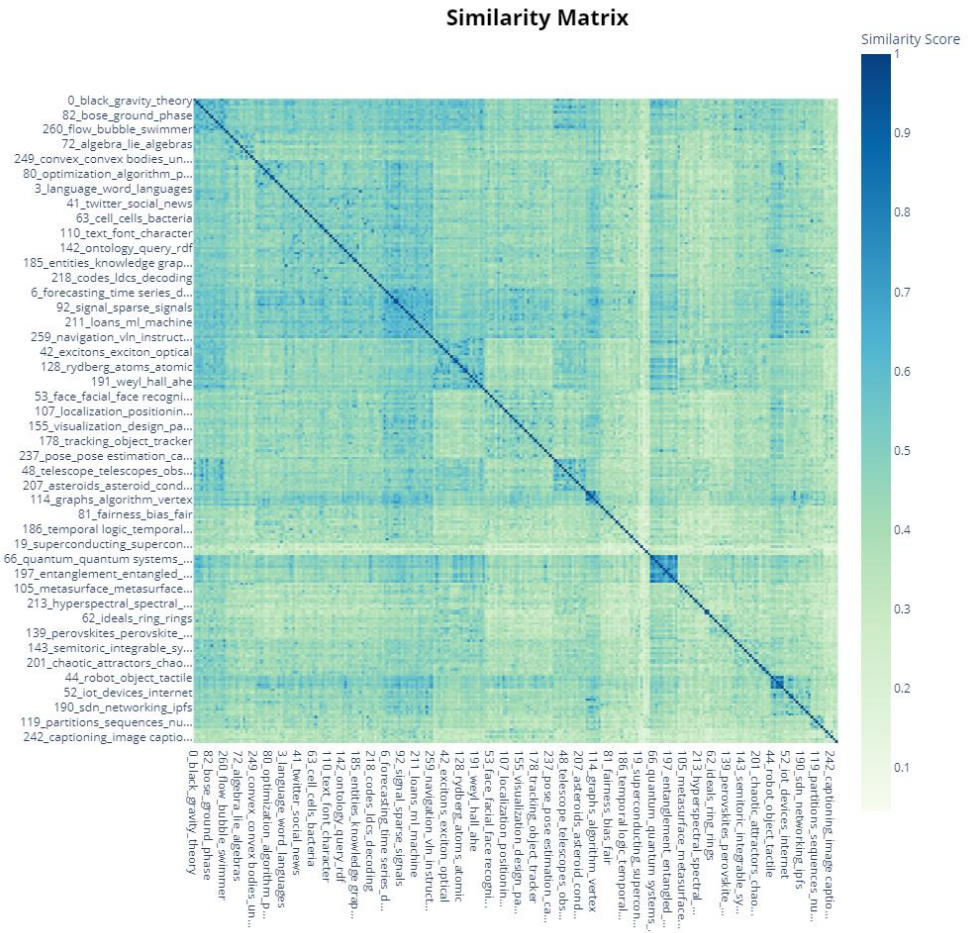


Рисунок 12. Корреляция между смоделированными темами

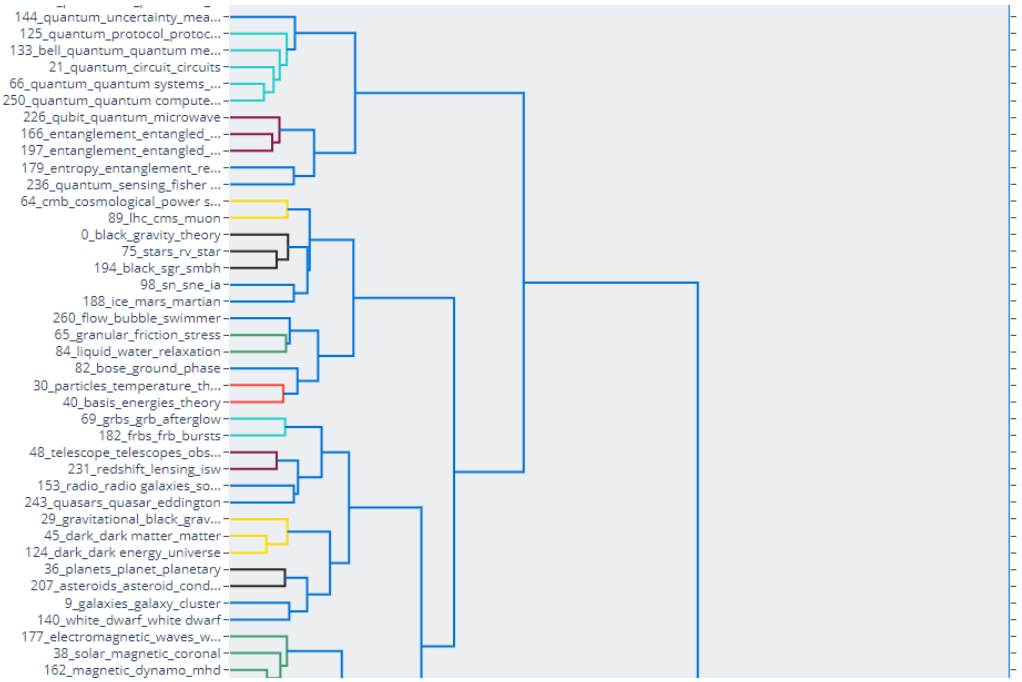


Рисунок 13. Часть иерархической кластеризации тем для BERTopic

6.5. Combined Topic Model

При обучении CombinedTM использовался предобработанный датасет без стемминга и необработанный датасет, что не дало значительных различий (значения когерентности c_v 0,64 и 0,61 соответственно). На первом варианте данных по когерентности тем было подобрано два количества тем – 20 и 100. Метрики показали на этих количествах одинаковые результаты.

Данная тематическая модель равномерно распределяет документы по темам и строит практически равномерное распределение на терминах.

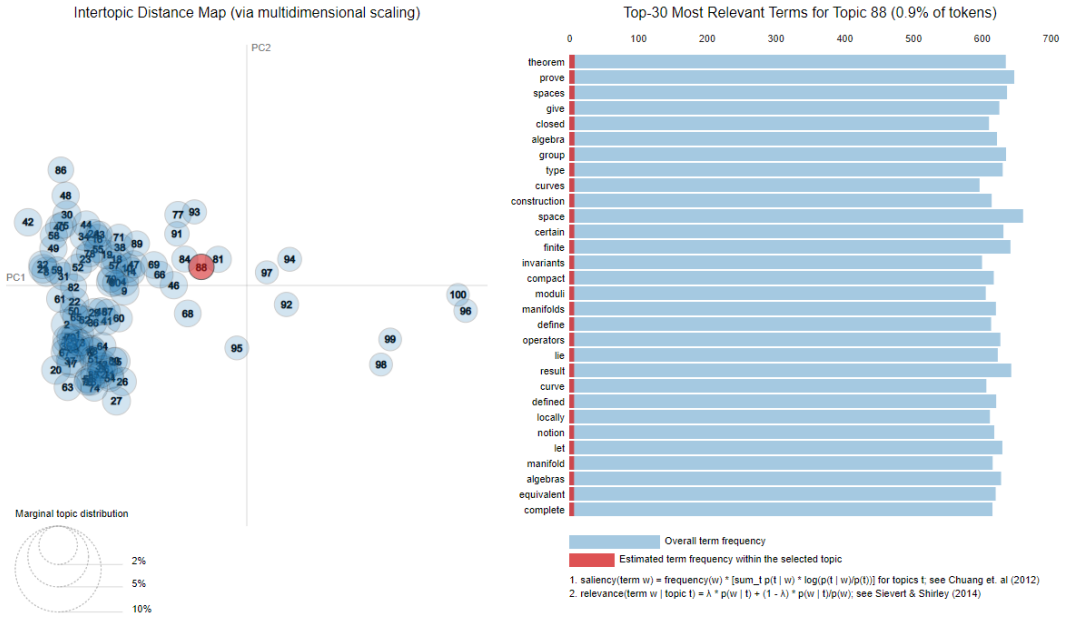


Рисунок 14. Распределение терминов для CombinedTM

7. Выводы

Было рассмотрено 6 тематических моделей и протестировано 5 из них. Существующие подходы можно поделить на 3 вида: вероятностные, основанные на эмбеддингах и смешанные. На первых двух вероятностных моделях (LSA и LDA) не удалось добиться качественных результатов. Увеличение меры согласованности тем за счет изменения количества выделяемых тем не дало улучшений. Для LSA темы плохо интерпретируются. LDA выделяя большое количество тем, также не делает их понятными человеку, распределение получается сильно неравномерным.

Тематические модели, основанные на эмбеддингах и смешанных подходах показали лучшие результаты. Но для них оказывается затруднительным подбор параметров, в том числе из-за отсутствие матрицы вероятностей и расчета метрик. Также модели подобного вида зачастую приближают или сводят задачу тематического моделирования к кластеризации.

Top2Vec отлично справляется с сохранением семантической и синтаксической связей из текста при моделировании тем. Документы с высокой вероятностью содержат в себе выделенные темы, термины с высокой вероятностью репрезентируют тему. При этом при визуальной оценке тем можно легко дать каждой из них название, что говорит о высокой согласованности темы. Однако, Top2Vec приближает задачу тематического моделирования к задаче кластеризации за счет уменьшения пространства совместно встроенных векторов и нахождению центров скопления документов с помощью HDBSCAN. Данная архитектура пренебрегает предположением о том, что документ может быть представлен несколькими темами с разной вероятностью.

BERTopic, показывает качественные результаты, моделирует легко интерпретируемые темы, но имеет два главных недостатка по сравнению с вышеупомянутой моделью. Первый заключается в наличии документах, которым не присвоено темы. Авторы ссылаются на то, что такой метод повышает качество моделирования тем, и несвязные, бессмысленные документы остаются изолированными. Однако, доля неразмеченных документов высока в проведенных экспериментах (30-50%), а топ-слова данной шумовой темы можно однозначно отнести к остальным смоделированным. Второй недостаток – сведение задачи тематического моделирования к задаче кластеризации. Авторы предлагают использовать матрицу вероятностей HDBSCAN как распределение тем по документам, но такое решение все еще не учитывает, что документы могут содержать несколько тем во время обучения BERTopic.

Для CombinedTM не найдено подходящего количества тем. Мера согласованности слабо меняется с изменением гиперпараметра. Тестирование

проводилось на 20 и 100 темах. В обоих случаях репрезентация тем через термины выходит смешанной, многим темам несложно дать название.

Также в экспериментах можно наглядно видеть недостатки меры согласованности для оценки качества тематических моделей – когерентность редко коррелирует с представлениями человека о качественной теме. Можно выдвинуть предположение, что рассмотрение текста как мешка слов игнорирует синтаксические и семантические отношения между элементами текста, а топ-слова, представляющие тему, не покрывают даже большей части текста. Модели, основанные на мешке слов, изначально представляют данные некогерентным способом.

Тематические модели на основе языковых моделей сохраняют основные характеристики осмысленного текста – синтаксическая и семантическая взаимосвязи между его элементами – но могут уделять слишком большое внимание синонимам и однокоренным словам. Здесь может помочь использование биграмм и триграмм. Предположительно, часто встречающиеся словосочетания лучше репрезентируют тему, чем отдельно взятые слова. Положительная сторона моделей, основанных на языковых моделях – они могут оставаться релевантными и совершенствоваться с течением времени при создании новых языковых моделей с большей производительностью и качеством эмбеддингов.

В итоге, наиболее релевантной моделью для тематического моделирования на основе аннотаций научных статей считаем Top2Vec.

8. Дальнейшее развитие

Один из вариантов использования результатов тематического моделирования – построение рекомендательной системы в случае наличия текстовых данных для объектов. Например, рецензия на фильм или краткое содержание произведений. В летней практической работе (Приложение 2)

мною проводилась работа над построением рекомендательной системы для сервиса по поиску киноплощадок на основе их текстового описания. Далее рассмотрим несколько способов создания dashboard’ов и кратко дадим описания существующим подходам к построению рекомендательных систем.

8.1. Дэшборды

Дэшборд – документ (чаще всего в электронном виде), представляющий из себя набор графиков, схем, таблиц. Содержит статистические данные, показывает основные показатели для дальнейшей аналитики или визуализации полученных результатов.

Так как планируемые рекомендательные системы пока что не встроены в какой-либо сервис, дэшборд помогает смоделировать рекомендательную систему и предоставляет возможности интерактивного взаимодействия с ней.

Есть несколько фреймворков для построения подобных интерфейсов визуализации данных. Для данной работы был выбран бесплатный фреймворк Dash, так как он включает в себя технологии Flask, React.js и Plotly.js, поддерживает язык Python и требует лишь базовых знаний HTML и CSS.

8.2. Рекомендательные системы

Разнообразие рекомендательных систем можно наблюдать, например, на современных сервисах по подбору музыки. Композиции рекомендуются на основе жанра, исполнителя, предпочтений пользователя, его профайлу или предыдущих оценках. Можно вручную указывать, какой объект будет рекомендоваться пользователю, но, конечно, наиболее распространенный метод основывается на анализе данных и машинном обучении, так как позволяет улавливать неявные взаимосвязи между объектом и пользователем.

Рекомендательные системы, основанные на математических моделях, разделяются на коллаборативную фильтрацию и фильтрацию на основе содержания. Первая группа использует информацию о прошлых оценках пользователя, прогнозирует оценку для нового объекта. Недостатком подхода

является «холодный старт» - отсутствие начальной информации для обучения для нового объекта или пользователя. Второй тип рекомендательных систем группирует объекты по характеристикам и выдает пользователю готовые подборки или варианты схожие с ранее просмотренными. Недостаток такого типа рекомендаций – зависимость от структуры объекта и предметной области, то есть специфичность построенной модели. К тому же такие модели получаются менее персонализированными, в отличие от моделей коллаборативной фильтрации.

8.3. Имеющиеся наработки

В практической работе для датасета описаний киноплощадок (Приложение 2) было разработано три рекомендательные системы: по ключевым словам, на основе результатов эмбеддинга Doc2Vec и на основе кластеризации. Ключевые слова генерировались из часто встречающихся существительных текстов описаний. Часть речи определялась при помощи функционала MyStem. Рекомендация на основе Doc2Vec строится по ближайшему вектору описания площадки в сравнении с ранее просмотренным пользователем. Рекомендательная система на основе кластеризации объединяет близкие векторы Doc2Vec в кластеры и выдает сразу группу похожих киноплощадок на ранее просмотренную.

8.4. Перенос результатов

Вышеописанные наработки можно перенести на результаты тематического моделирования на аннотациях статей. Рекомендательную систему на основе ключевых слов построить на топ-словах тем, которые выделил Top2Vec. В рекомендательной системе на основе кластеризации использовать результаты тематического моделирования, так как в одной статье (а, следовательно, и в ее аннотации) затрагивается несколько научных областей.

Данные идеи планируется имплементировать в имеющуюся работу и распространить на продуктовую аналитику.

9. Заключение

Было рассмотрено пять методов тематического моделирования на датасете аннотаций научных статей. Модели дали различные результаты, и каждая показала особенности в работе и имплементации. При выборе метода необходимо в первую очередь опираться на специфику предметной области и поставленные цели. Некоторые методы улавливают глобальные темы и могут выделить большие темы за короткий промежуток времени. Другие пригодятся для моделирования большего количества тем и требуют больше времени для обработки. При выявлении релевантной модели не стоит пренебрегать визуальной оценкой результатов и при необходимости и возможности привлекать экспертов.

Результаты тематического моделирования планируется использовать при построении рекомендательных систем для объектов с текстовым описанием.

10. Список источников

[1] ArXiv Public Datasets, Kaggle, 2019 по наши дни [Электронный ресурс]. URL: https://www.kaggle.com/Cornell-University/arxiv

[2] Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation //Journal of machine Learning research. – 2003. – Т. 3. – №. Jan. – С. 993-1022. [Электронный ресурс]. URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[3] J. Xu, “Topic Modeling with LSA, PLSA, LDA & lda2Vec”, Medium.com. URL: https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05 (accessed Feb, 2022)

[4] Angelov D. Top2vec: Distributed representations of topics //arXiv preprint arXiv:2008.09470. – 2020. [Электронный ресурс]. URL: https://arxiv.org/abs/2008.09470

[5] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure //arXiv preprint arXiv:2203.05794. – 2022. [Электронный ресурс]. URL: https://maartengr.github.io/BERTopic/index.html

[6] Moody C., “Introducing our Hybrid lda2vec Algorithm”, Сан-Франциско, CA, 2016. [Электронный ресурс]. URL: https://multithreaded.stitchfix.com/blog/2016/05/27/lda2vec/#topic=38&lambda=1&term=

[7] Воронцов К. В. Вероятностное тематическое моделирование //Москва. – 2013. [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

[8] Bianchi F. et al. Cross-lingual contextualized topic models with zero-shot learning //arXiv preprint arXiv:2004.07737. – 2020. [Электронный ресурс] URL: https://aclanthology.org/2021.eacl-main.143.pdf

[9] S. Kapadia, “Evaluate Topic Models: Latent Dirichlet Allocation (LDA)”, 2019. [Электронный ресурс] URL: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[10] Bianchi F., Terragni S., Hovy D. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence //arXiv preprint arXiv:2004.03974. – 2020. [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/images/archive/0/04/20180627210520%21Alekseev2018BSThesis.pdf

11. Приложения

[1] ВКР: https://github.com/KvindtEva/GRADUATE-WORK

[2] Курсовая работа за 3-ий курс: https://github.com/KvindtEva/Course