ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет Санкт-Петербургская школа физико-математических и компьютерных наук НИУ ВШЭ

Титова Ксения Максимовна

ОБУЧЕНИЕ РАЗРЕЖЕННЫХ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ ДЛЯ ЗАДАЧИ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Выпускная квалификационная работа

по направлению подготовки 01.04.02 Прикладная математика и информатика
образовательная программа «Машинное обучение и анализ данных»

Рецензент
Старший NLP разработчик, MTS AI

И.В. Кротова

Руководитель
к. ф-м. н., доцент депратамента информатики, НИУ ВШЭ – Санкт-Петербург

М.С. Мухин

Консультант
к. т. н., доцент факультета программной инженерии, университет ИТМО

А.В. Платонов

Санкт-Петербург, 2023 год

Оглавление

Аннотация

Тематическое моделирование - инструмент для обнаружения скрытых тематик в коллекциях документов. Тематика определяется набором слов, наиболее характерных для документов, которые к ней относятся. Недавние исследования показали эффективность применения векторных представлений слов и документов для задачи тематического моделирования, однако при таком подходе остается нерешенной проблема интерпретируемости. В данной работе представляется модель SpaRT, в которой в векторное пространство представлений документов внедряются свойства разреженности и независимости компонент, согласующиеся с интуицией задачи тематического моделирования. Регуляризация осуществляется во время обучения с многоцелевой функцией потерь при помощи мер негауссовости и l1-регуляризации. Для спецификации тематик разработан метод class-weighted-TF-IDF, выделяющий из представления тематики наиболее коррелирующие с ней слова. Для построенной тематической модели были подтверждены гипотезы о связи индуцированных свойств независимости и разреженности с метриками тематического моделирования. Модель показала качественные результаты, сравнимые с известными тематическими моделями.

Ключевые слова: тематическое моделирование, векторное представление, семантическое пространство, BERT, разреженность, негэнтропия.

Topic modeling is a tool for discovering hidden topics in document collections. The topic is determined by a set of words that are most characteristic of related documents. Recent studies have shown the effectiveness of word and document embeddings for topic modeling, however the problem of its interpretability remains unsolved. We present the SpaRT

model, in which the properties of sparsity and component independence are introduced into the vector space of document representations, confirming the intuition of the topic modeling. Regularization is introduced during training with a multi-objective loss function using measures of non-Gaussianity and l1-regularization. To specify topics, the class-weighted-TF-IDF method has been developed. It extracts the most correlated words from the topic representation. For the constructed topic model, the hypotheses about the relationship between the induced properties of independence and the metrics of topic modeling were confirmed. The model showed qualitative results comparable to known topic models.

Keywords: topic modeling, word embedding, semantic space, BERT, sparsity, negentropy.

Введение

Тематическое моделирование – аппарат для анализа коллекции текстовых документов, который позволяет выделить набор представленных в текстах тем, и помогает определить, к каким из них относится каждый из документов. Каждая тема, далее также называемая тематикой, определяется набором наиболее для неё характерных слов.

Многие тематические модели основаны на статистических методах – к ним, например, относятся модели латентного размещения Дирихле [1]. В таких моделях каждая тематика описывается модельным распределением по словам словаря, а документ представляется как смесь тематик.

Вероятностные тематические модели обычно просты в реализации и широко используются, но имеют ряд технических недостатков. Во-первых, хоть они и учитывают статистические свойства естественного языка, они не принимают в расчет лингвистические свойства текста - не учитывают порядок слов, семантическую связь между ними и их смысловую составляющую. Из-за этого результат не всегда получается интуитивным, – смоделированные темы могут быть слишком похожими на другие темы, казаться несвязными и непонятными. Во-вторых, они плохо работают с большими словарями, что становится важной проблемой с закономерным увеличением в мире количества данных и объема датасетов. С ней борятся удалением из мешка слов слишком редких или слишком частых термов, из-за чего в представлениях документов теряется смысловая часть, особенно если мы работаем с корпусами узкоспециализированных текстов.

Одна из идей борьбы с указанными проблемами использует методы семантического моделирования, в частности – построение и использо-

вание векторных представлений слов и документов вместо вычисления статистик. Для построения векторных представлений слов в основном используют нейронные сети. С появлением в 2017 году механизма внимания и архитектуры трансформера [2], векторные представления текстов научились искать внутренние лингвистические закономерности, что помогло достичь более качественных результатов в различных задачах обработки естественного языка, в том числе в задаче тематического моделирования. Главным недостатком векторных представлений, построенных с помощью нейронных сетей, является их плохая интерпретируемость.

В данной работе предлагаются и исследуются методы регуляризации векторного пространства эмбеддингов во время обучения модели с целью внедрения в него свойств, повышающих интерпретируемость представлений документов. Основной целью данной исследовательской работы было построение тематической модели, использующей разреженные векторные представления документов с выполнением свойства независимости компонент.

Чтобы достичь поставленной цели необходимо выполнить следующие задачи.

1. Изучение существующих способов построения векторных семантических пространств и способов их регуляризации.

2. Разработка архитектуры нейронной сети для построения векторных представлений документов.

3. Исследование меры свойств разреженности и независимости - задача включает в себя построение эвристических алгоритмов и поиск наилучших приближений используемых в них численных величин; изучение способов их внедрения в качестве регуляризаторов на этапе обучения модели.

4. Разработка способа для спецификации тематики из векторного представления документа.

5. Оценка качества построенной тематической модели:

   a) подбор адекватных метрик;

   b) проведение экспериментов для проверки гипотез, связанных с выполняемостью свойства независимости;

   c) сравнение результатов с существующими моделями на задаче тематического моделирования.

Глава 1. Семантические векторные пространства

1.1. Векторные представления слов и документов

Вместе с базовым алгоритмом тематического моделирования LDA возникла идея векторных представлений слов, называемых также эмбеддингами. Исследования в этой области начались с нейронной языковой модели Бенджио [3], опубликованной в том же году и журнале, что и статья о LDA [1]. Переходя к векторным представлениям слов, мы отказываемся от "one-hot"представлений - такие представления являются векторами длины словаря, состоящими из нулей и единицы в единственной позиции соответствующего слова. Векторные представления индуцируют распределенное представление, в котором слова с похожими значениями находятся близко друг к другу в пространстве более низкой размерности [4].

Идея использования векторных представлений слов стала важна во многих приложениях обработки естественного языка, а также была расширена на иные типы данных. Наиболее известным примером векторных представлений слов является word2vec [5].

В 2019 году авторы статьи [6] предложили объединить идеи LDA и векторных представлений слов и представили модель ETM (Embedding Topic Model).

Идею векторных представлений слов можно продолжить и прийти к векторным представлениям документов. В пространстве таких векторов каждому документу сопоставляется элемент в векторном пространстве, – это позволяет оценивать семантическую близость документов как близость соответствующих векторов, а также кластеризовать похожие

документы. Построенное семантическое пространство позволяет решать задачи, для которых не требуется обучение с учителем: это могут быть задачи кластеризации, поиск похожих документов или, непосредственно, тематическое моделирование.

Модель doc2vec предложили в статье [7] - её используют для задачи информационного поиска, и оценивают релевантность запроса и документа скалярным произведением их представлений. Для задачи тематического моделирования авторы представили в [8] алгоритм top2vec, использующий представления doc2vec и переводящий их в пространство меньшей размерности и кластеризующий их для определения тематик через группы обособленных документов.

В статье [9] предлагают объединять эмбеддинги слов с представлениями параграфов и подавать на вход модеи сконкатенированный вектор. Эксперименты показали, что такой подход улучшает семантические свойства эмбеддинга и повышает качество работы модели на unsupervised-задачах (поиск похожих документов).

1.2. Датасеты тематического моделирования

Основной целью тематического моделирования является определение тематической кластерной структуры текстовой коллекции, – поиск и описание содержащихся в ней тем. Такая задача имеет множество приложений, среди которых, например, поиск в электронных библиотеках, выявление и отслеживание новостей, поиск тематического контента в социальных сетях. Датасеты, которые используются при построении и оценке тематических моделей, обычно состоят из соответствующих таким подзадачам данных: это могут быть заголовки новостей или блоги пользователей.

При изучении статей о существующих подходах к построению и сравнению тематических моделей, было замечено что датасеты для это-

го используются, как правило, разные, и не существует общепринятого бенчмарка для оценки тематических моделей. Это связано в том числе с тем что задача тематического моделирования является unsupervised-задачей, то есть задачей, для решения которой не требуется меток. Поэтому авторы тематических моделей используют любые датасеты, согласующиеся с интуицией приложений тематического моделирования, – важно лишь корректно выбирать метрики для оценки качества.

Наиболее известные датасеты, использующиеся для оценки тематических моделей:

- 20NewsGroups: Датасет содержит 18,831 новостей с соответствующими группами - каждому документу соответствуют одна из двадцати тематик. Зачастую такие метки игнорируют ([8, 10]) и требуют от модели построить другие тематики, так что каждому документу не обязательно будет соответствовать лишь одна тема.

- Reuters: Содержит 21,578 неразмеченных новостных заголовков. Этот датасет будет использован для оценки нашей модели.

- Research Articles 2.0: Датасет с научно-популярными статьями объема 20,006. Каждой статье соотвествуют метки, описывающие научные разделы. На рисунке (1.1) показано распределение количества меток на документ – можно сделать вывод о малом количестве меток на каждый документ, не превышающем семи.

1.3. Разреженность

Задание размерности семантического пространства - простейший способ регулирования его свойств. Чем меньше берется размерность пространства, тем легче с ним становится работать: пространства меньшей размерности легче интерпретировать и визуализировать, работа с ними на конечном устройстве будет быстрее. С другой стороны, чем меньше

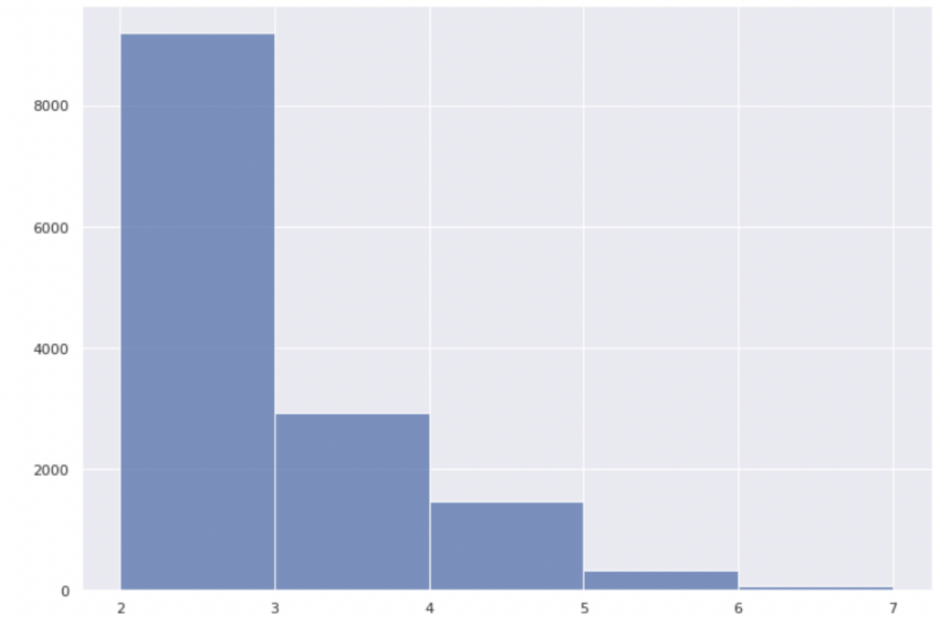

Рис. 1.1. Гистограмма количества тематик в Research Articles 2.0

размерность, тем более "сжатыми"будут представления - с уменьшением количества признаков у представлений увеличивается их смешанность и зависимость, что затрудняет отделение непохожих документов друг от друга.

В векторных пространствах высокой размерности представления будут содержать больше информации, но работать с ними будет менее эффективно. Переход к разреженным представлениям - представлениям с преимущественно нулевыми элементами - решает эту проблему.

Подобный подход был описан в статье [11], где авторы обучали многомерное пространство представлений документов для задачи информационного поиска и накладывали на него требование разреженности. Это свойство позволяет построить инвертированный индекс для каждого признака-столбца в векторном пространстве. В дальнейшем такой подход помогает искать релевантные к запросу документы за константную временную сложность, используя инвертированные индексы как хэш-таблицы.

Мы также предлагаем использовать для построения векторных представлений документов пространство большой размерности, при этом мы хотим индуцировать в нем разреженность. Для задачи тематического моделирования использование инвертированных индексов в общем случае не пригодится, однако все еще может быть полезным инструментом для анализа выделенных моделью тематик и соответствующим им документов. Разреженность семантического пространства для нашей задачи обусловлена интуивным пониманием тематик - как было показано в предыдущем разделе, каждому документу обычно соответствует малое количество тем.

1.4. Независимость признаков

Добавление разреженности в векторные представления документов - не единственная эвристика, согласующаяся с задачей тематического моделирования. Зачастую для проверки адекватности тематической модели от неё требуют способность выделять тематики, различные по смыслу. Рассматривая тематики как индексы компонентов векторов из многомерного разреженного пространства, мы хотим чтобы они были независимыми.

Чтобы определить понятие независимости, рассмотрим две случайные величины y_1 и y_2. Пусть $p(y_1, y_2)$ - совместная плотность распределения этих случайных величин, тогда положим что $p_1(y_1) = \int p(y_1, y_2)dy_2$ - маргинальная плотность y_1 (для y_2 - по аналогии). Тогда определим что y_1 и y_2 независимы, когда совместная плотность вероятности факторизуется:

$$p(y_1, y_2) = p_1(y_1)p_2(y_2) \tag{1.1}$$

Чтобы оценивать независимость в [12] предлагается использовать меры негауссовости. По центральной предельной теореме распределение суммы независимых случайных величин стремится к нормальному рас-

пределению при выполнении ряда условий. Отсюда следует эвристика: сумма двух независимых случайных величин имеет распределение более близкое к нормальному, чем любое у двух исходных. Поэтому с максимизацией негауссовости увеличивается независимость.

Для оценки негауссовости вектора y рассмотрим несколько численных мер. Первая, коэффициент эксцесса, определяется формулой (1.2).

$$kurt(y) = E\left\{y^4\right\} - 3(E\left\{y^2\right\})^2 \tag{1.2}$$

Несложно показать, что если $y \sim N(0,1)$, то $kurt(y) = 0$. Для большинства негауссовых случайных величин коэффициент эксцесса ненулевой, а значит максимизация его абсолютной величины коррелирует с задачей максимизации негауссовости.

Второй оценкой может служить негэнтропия. Она основана на информационно-теоретическом свойстве дифференциальной энтропии. Для случайной величины y энтропия задается формулой (1.3), если y имеет дискретное распределение, и (1.4) - если абсолютное.

$$H(Y) = -\sum_i P(Y = a_i) \log P(Y = a_i) \tag{1.3}$$

$$H(y) = -\int f(y) \log f(y) dy \tag{1.4}$$

Энтропия мала для тех распределений, значения которых концентрированы в определенных значениях. Согласно [13], в классе распределений с одинаковой дисперсией наибольшую энтропию имеет случайная величина с нормальным распределением. Чтобы получить неотрицальную меру негауссовости равную нулю для случайной величины, распределенной нормально, определим негэнтропию:

$$J(y) = H(y_{gauss}) - H(y) \tag{1.5}$$

С точки зрения результатов теории статистики, использование негэнтропии более обоснованно для оценки негауссовости, чем использование коэффициента эксцесса, однако напрямую получать её вычислительно сложно. В разделе (2.3.2) мы обсудим возможные численные приближения негэнтропии.

Выводы и результаты по главе

В данной главе был проведен обзор литературы и были описаны базовые понятия, такие как векторные представления слов и документов, а так же способы их использования для построения тематической модели. Была показана обоснованность свойств разреженности и независимости компонент и показана их связь с задачей тематического моделирования. Описан способ внедрения независимости с помощью эвристики с мерами негауссовости и приведены их примеры.

Глава 2. Построение модели и её обучение

В данной главе описывается модель (далее называемая **SpaRT**: **Spa**rse **R**epresentations from **T**ransformers), спроектированная нами для построения разреженных векторных представлений, включая её архитектуру, методы обучения и их математическое обоснование.

2.1. Математическая постановка задачи

Пусть есть коллекция документов $S = \{x_1, x_2, ..., x_K\}$, где $x_i \in \mathbb{R}^n$, а K - количество документов. Необходимо построить функцию $F : \mathbb{R}^N \longrightarrow \mathbb{R}^M$, которая принимает исходное представление документа и переводит его в пространство высокой размерности $M >> N$, при этом должен выполняться ряд условий:

1. $\frac{1}{M} \sum_{i=1}^{M} |F(x_k)_i| \longrightarrow 0$
2. $y = F(x_k); P(y) \longrightarrow \prod_{i=1}^{M} P(y_i)$

Сконкатенировав по новой оси представления из $\mathbb{R}^M$ для всех документов, мы получим матрицу A, в которой каждой строке i соответствует документ, каждому столбцу j - тематика, а элементу матрицы a_{ij} соответсвует вес тематики для соответствующего документа. Таким образом, о матрице A можно думать как о наборе вектор-строк документов в базисе из тематик, а можно как о наборе вектор-столбцов тематик в базисе всей коллекции документов. Такие представления будут индуцировать семантическое пространство, на которое при обучении будут налагаться ограничения.

В данной постановке исходным представлением документа может считаться его токенизированное представление, а функция F определяется прямым проходом нейронной сети, которая обучается под нашу за-

дачу. Первое условие налагает ограничение на разреженность пространства, а второе на независимость тематик - в разделе 2.3 мы обсудим как их учесть в функции потерь.

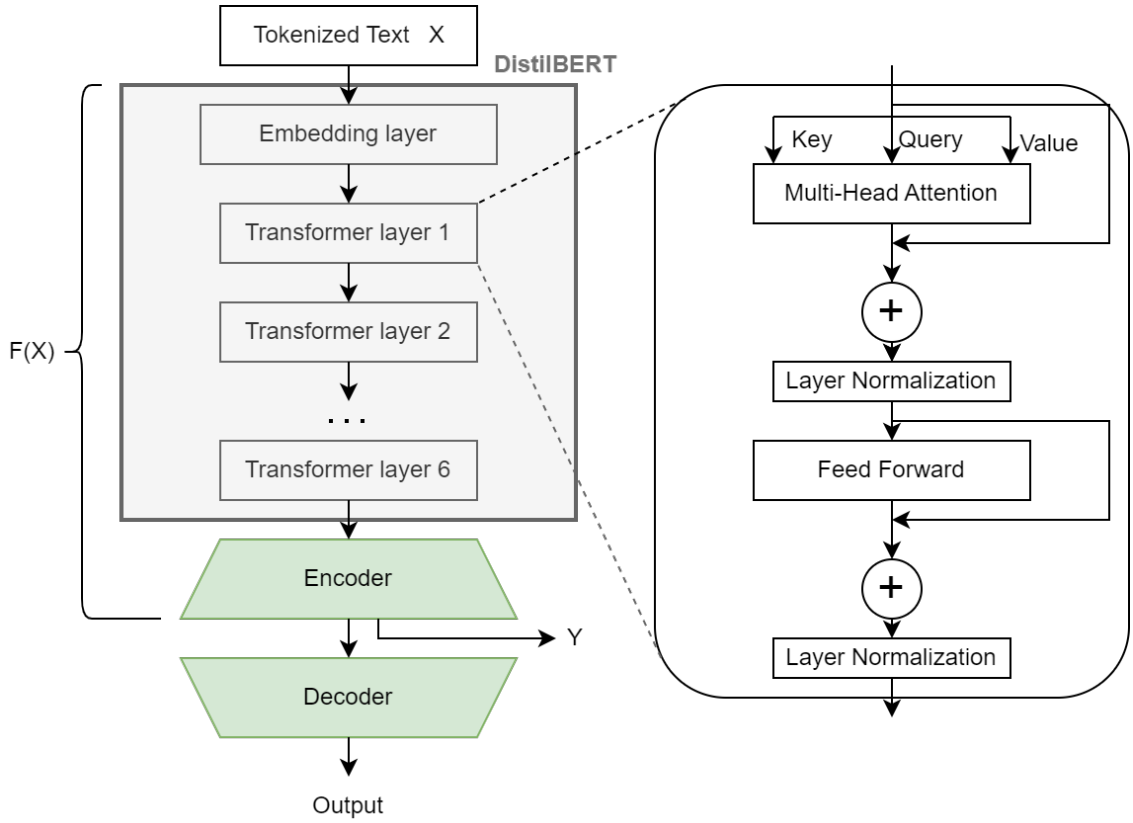

Рис. 2.1. Архитектура модели SpaRT

2.2. Архитектура модели

Для получения векторных представлений документов было решено использовать архитектуру трансформера - механизм self-attention поможет найти внутренние связи между словами в документе, что будет полезно в том числе для задачи тематического моделирования. За основу в качестве базовой модели мы возьмем архитектуру DistilBERT, состоящую из 6 блоков-трансформеров (см. рис. 1).

К архитектуре базовой модели добавляется блок состоящий из кодировщика и декодировщика. Кодировщик состоит из линейного слоя и активации ReLU – он будет переводить векторное представление доку-

мента из базовой модели в многомерное разреженное представление, а декодировщик - сжимать его обратно до размерности выхода DistilBERT.

Веса для базовой модели были взяты из репозитория [14] hugginface - их вариант модели ColBERT был обучен на датасете MSMARCO-Passage-Ranking - модель также училась строить представления документов, хоть и на задачах отличных от тематического моделирования. Эти параметры были заморожены во время обучения.

2.3. Функция потерь

Пусть Θ – обучаемые параметры нейронной сети. Главная составляющая функции потерь определяется членом реконструкции (2.1).

$$R_{rec}(\Theta) = \frac{\sum (F_\Theta(x) - x)^2}{K} \tag{2.1}$$

Мы требуем от нашей модели, чтобы представления возвращаемые кодировщиком не теряли информацию от предобученной модели и декодировщик мог их восстановить.

Ограничения, перечисленные в математической постановке задачи, также должны учитываться при обучении разреженных векторных представлений. Для этого в функцию потерь необходимо добавить два регуляризационных члена $R_{Sp}(\Theta)$ и $R_{Ind}(\Theta)$, отвечающих ограничениям на разреженность и независимость признаков.

Итак, функция потерь определяется следующим выражением:

$$J(\Theta) = \alpha_1 R_{rec}(\Theta) + \alpha_2 R_{Sp}(\Theta) + \alpha_3 R_{Ind}(\Theta) \tag{2.2}$$

2.3.1. L0 и L1-регуляризация

Для выполнения свойства разреженности необходимо научить модель занулять компоненты в векторе $y = F(x_k)$. Для этого добавим в функцию потерь регуляризационный член.

Регуляризация с минимизацией l_0-нормы (2.3) по определению штрафует модель за количество ненулевых компонент. В силу невыпуклости l_0-нормы, оптимизационная задача с её минимазацией не может быть эффективно решена вычислительно, и в целом является NP-трудной.

$$J(\Theta) = R(\Theta) + \lambda||\Theta||_0 = R(\Theta) + \lambda\#(i|\Theta_i \neq 0),\ \lambda > 0 \tag{2.3}$$

Накладывая штраф на целевую функцию с использованием l_1-нормы, мы получим (2.5) - метод, также именуемый LASSO-регуляризацией. Не напрямую, как в случае с l_0-нормой, но он также индуцирует разреженность, поскольку в отличие от минимизации $l_q, q > 1$ норм, l_1-норма всегда уменьшается на константное значение, вне зависимости от её величины.

$$J(\Theta) = R(\Theta) + \lambda||\Theta||_1 = R(\Theta) + \lambda\sum_{i=1}^{m}|\Theta_i|,\ \lambda > 0 \tag{2.4}$$

С учетом приведенных соображений первый регуляризационный член из функции потерь выглядит следующим образом:

$$R_{Sp}(\Theta) = I\left(\#(i|\Theta_i \neq 0) > t_{min}\right)\lambda_1\sum_{i=1}^{M}|\Theta_i| \tag{2.5}$$

Параметр t_{min} отвечает за минимальное количество ненулевых признаков в разреженном представлении документа.

2.3.2. Приближения мер негауссовости

Увеличение независимости признаков (компонент вектора $y = F(x_k)$) будет происходить посредством минимизации мер негауссовости.

1. Первая мера негауссовости - коэффициент эксцесса, который можно вычислить напрямую через формулу (1.2).

2. Классический метод аппроксимации негэнтропии [15] использует моменты высокого порядка:

$$\mathrm{Neg}(y) \approx \frac{1}{12} E\left\{y^3\right\}^2 + \frac{1}{48} kurt(y)^2 \tag{2.6}$$

3. Использование коэффициента эксцесса в формуле (2.6) отрицательно влияет на робастность модели [16], поэтому в [17] предлагают (2.7) новый класс приближений негэнтропии.

$$\begin{gathered} \mathrm{Neg}(y) = H(\nu) - H(x) \approx \\ k_1(E\left\{\bar{G}_1(X)\right\})^2 + k_2(E\left\{\bar{G}_2(X)\right\} - E\left\{\bar{G}_2(\nu)\right\})^2 \end{gathered} \tag{2.7}$$

Ниже приведены два практических примера функций $\hat{G}_i$ и коэффициентов k_1, k_2.

a) $\hat{G}_1(x) = x\exp(-x^2/2)$; $\hat{G}_{2a}(x) = |x|$

$k_1 = 36/(8\sqrt{3} - 9)$; $k_{2a} = 1/(2 - 6/\pi)$

b) $\hat{G}_1(x) = x\exp(-x^2/2)$; $\hat{G}_{2b}(x) = \exp(-x^2/2)$

$k_1 = 36/(8\sqrt{3} - 9)$; $k_{2b} = 24/(16\sqrt{3} - 27)$

Введем обозначения для четырех мер негауссовости – среди них коэффициент эксцесса NG_{kurt} (1), классическое приближение негэнтропии NG^{C}_{neg} (2) и два приближения из пункта (3): $\mathrm{NG}^{G_1}_{neg}$ и $\mathrm{NG}^{G_2}_{neg}$. Для увеличения негауссовости негэнтропию нужно максимизировать, поэтому в качестве коэффициента регуляризации $R_{Ind}(\Theta)$ мы будем брать обратные к перечисленным приближениям величины.

Выводы и результаты по главе

В главе была описана математическая постановка задачи и детали построения модели **SpaRT**. Описана её архитектура и многоцелевая функция потерь, а так же обоснованы приближения и регуляризаторы, используемые в ней.

Глава 3. Извлечение тематик из векторных представлений

Получаемые с помощью нейронных сетей векторные представления плохо интерпретируются, что не позволяет объяснить определяющие их признаки. В следующей главе описываются идеи интерпретации компонентов вектора разреженного представления, соответствующих тематикам, и как впоследствии мы будем оценивать построенную модель для задачи тематического моделирования.

3.1. CW-TF-IDF

В задаче тематического моделирования нам необходимо уметь извлекать из векторных представлений тематик (столбцов матрицы, полученной конкатенацией представлений документов) наиболее коррелирующие с ними слова. Для этой задачи можно использовать статистические подходы.

Вспомним суть метода TF-IDF: в нем наиболее значимые для текста слова ищутся как слова с высокой частотностью в нем, и низкой во всех остальных. Вес слова $t \in \text{Dict}$ для документа $d \in D$ определяется следующей формулой:

$$\textbf{TF-IDF}(t, d, D) = \textbf{TF}(t, d) \times \textbf{IDF}(t, D) \tag{3.1}$$

В выражении $\textbf{TF}(t, d) = \frac{n_t^d}{\sum\limits_{k \in \text{Dict}} n_k^d}$ (n_t^d - количество вхождений слова t в документе d) оценивается важность слова для конкретного документа, а в выражении $\textbf{IDF}(t, D) = \log \frac{|D|}{\{d_i \in D | t \in d_i\}}$ рассчитывается важность данного слова t по совокупности документов D.

Распространим идею использования TF-IDF на наши разреженные представления. Каждой тематике соответствует вектор-столбец весов для каждого документа из коллекции (рис. 2). Значит для каждой тематики T_j можно собрать набор документов с ненулевыми весами $D^T = \{d_i \in D | a_{ij} \neq 0\}$, причем больше вес у документа, тем больше он коррелирует с тематикой.

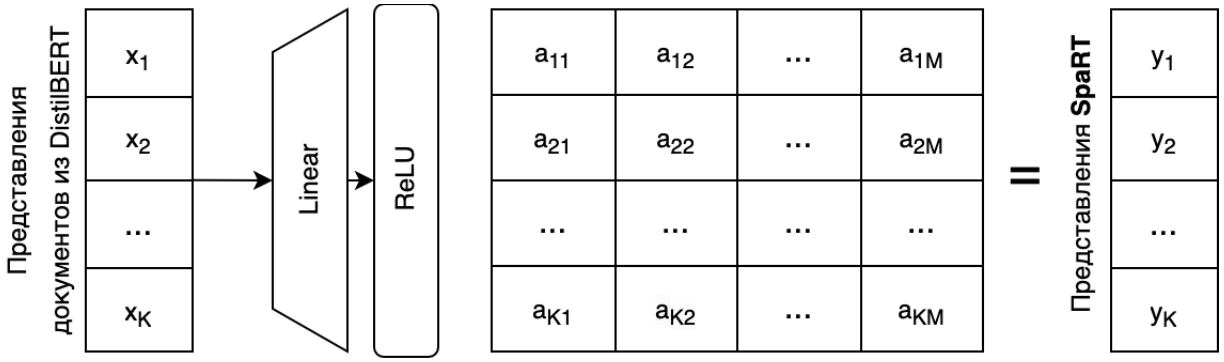

Рис. 3.1. Схематичное представление работы кодировщика

В методе **CW-TF-IDF** (class-based weighted TF-IDF) считается вес слова не для отдельных документов, а для тематики, при этом учитываются коэффициенты при соответствующих тематике документах. Множители в (3.1) принимают следующий вид:

$$\mathbf{TF}_{CW}(t, T_j) = \frac{\sum\limits_{i=0}^{|D|} a_{ij} n_t}{\sum\limits_{k \in \text{Dict}} \sum\limits_{i=0}^{|D|} a_{ij} n_k} \tag{3.2}$$

$$\mathbf{IDF}_{CW}(t, D) = \log\left(1 + \sum_{j=0}^{m} \frac{\overline{T}}{\sum\limits_{i=0}^{|D|} a_{ij} n_t}\right) \tag{3.3}$$

В формуле (3.3) величину $\overline{T} = \frac{\sum\limits_{k \in \text{Dict}} \sum\limits_{i=0}^{|D|} a_{ij} n_k}{m}$ можно считать взвешенным средним количеством слов в тематике.

Похожий метод предлагается в [10], – авторы также распространяют идею TF-IDF на классы документов. Наш метод также использует коэффициенты из матрицы представлений документов как их веса.

3.2. Интеграция с библиотекой OCTIS

Для проведения экспериментов удобно прибегнуть к помощи инструмента, позволяющего сравнивать построенную модель с известными тематическими моделями. Было решено воспользоваться Python-библиотекой OCTIS, в которой реализованы классические методы и предоставлен функционал для их оценивания с помощью различных state-of-the-art метрик.

Модель **SpaRT** была имплементирована через класс модели OCTIS – от него, как от тематической модели, требуются методы предобработки данных, обучения модели и извлечения наиболее важных слов для каждой выделенной тематики. Под предобработкой данных для **SpaRT** подразумевается токенизация текста с помощью токенизатора DistilBERT; обучение может проводиться как на коллекции документов, из которых требуется извлечь тематики, так и на некотором большом корпусе документов; извлечение важных слов производится с помощью метода, описанного в разделе 3.1.

Стоит отметить, что датасеты представленные в библиотеке, были заранее предобработаны авторами под статистические методы – в них опускаются знаки препинания, слова лемматизируются и фильтруются. Для нейросетевых моделей с архитектурой трансформера такая предобработка не требуется и скорее отрицательно скажется на результате предсказаний, поэтому датасеты для оценки были взяты из других источников.

Выводы и результаты по главе

В данной главе была описана проблематика спецификации тематик, построенных с помощью нейронных сетей. Был разработан метод **CW-TF-IDF** для извлечения наиболее важных для тематики слов из её векторного представления. Так же описаны детали интеграции с Python-библиотекой OCTIS, необходимой для дальнейшей оценки тематической модели.

Глава 4. Оценка качества

4.1. Метрики тематического моделирования

Оценивание тематических моделей - нетривиальная задача, поскольку в парадигме обучения без учителя у входных данных нет меток, на которые можно было бы ориентироваться для оценки качества предсказаний модели. Из-за этого может возникнуть субъективность интерпретации выходов модели. При выборе метрик мы руководствуемся соображениями о том, что они должны быть легко интерпретируемыми и должны качественно подражать человеческим суждениям.

Для оценки качества тематических моделей было выбрано две метрики: согласованность тематик (topic coherence) и их разнообразие (topic diversity).

Согласованность тематик - это метрика, которая измеряет, насколько связанными являются слова, отнесенные к одной теме в тематической модели. Она будет численно оцениваться с помощью нормализованной поточечной взаимной информации (normalized pointwise mutual information, [18]). Такая оценка лежит в интервале $[-1, 1]$, где единице соответствует идеальная согласованность, то есть чем выше такая метрика, тем более семантически похожи между собой слова, определяющие тематику.

Разнообразие тематик определяется долей уникальных слов для каждой тематики, то есть эта метрика измеряет, насколько разнообразны темы в тематической модели. Она предполагает, что более разнообразные темы обеспечивают лучшее покрытие тем в текстовом корпусе и более полное представление контента. Оценка лежит в интервале $[0, 1]$,

где нулю соответствуют тематики, повторяющие друг друга по смыслу, а единице - разнородные тематики.

4.2. Проверка выполнения свойств модели

Все последующие эксперименты были проведены на данных, собранных из новостных статей агенства Reuters. Пример данных из датасета состоит из заголовка и содержания статьи на английском языке. Как правило для сравнения работы тематических моделей используют малое количество тематик, но поскольку мы хотим проверить свойства разреженности и независимости, положим количество тематик равным тысяче. В приведенных далее экспериментах модели обучались с различными параметрами α_i (коэффициенты регуляризации) – с точными их значениями для каждого эксперимента можно ознакомиться в приложении.

Веса параметров инициализировались как веса модели $\mathbf{SpaRT_{BASE}}$, заранее обученной на задачу реконструкции представлений DistilBERT – таким образом, представления тематик на данном этапе уже можно считать построенными, однако для них пока не выполняются желаемые свойства. В экспериментах (4-5) мы вычислим метрики тематического моделирования в том числе для этой базовой модели.

Прежде чем формулировать и проверять гипотезы о связи свойств разреженности и независимости с метриками тематического моделирования, нужно убедиться в том что эти свойства вообще удалось индуцировать. Начнем с проверки свойства разреженности - запустим обучение модели **SpaRT** с различными значениями параметра α_2 (α_3 равно нулю). Изучив рис. 4.4, мы можем убедиться что добавление члена l1-регуляризации действительно зануляет компоненты векторного пред-

ставления, причем чем больше коэффициент α_2, тем больше в эмбеддинге нулевых компонент.

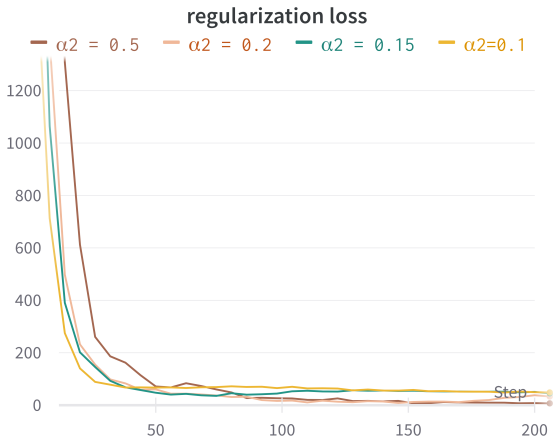

Рис. 4.1. Зависимость $R_{Sp}(\Theta)$ от итерации обучения

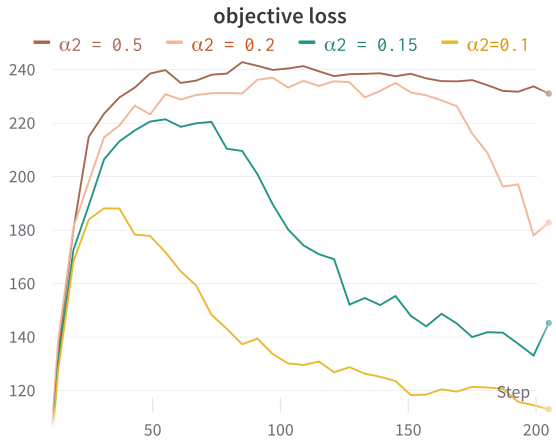

Рис. 4.2. Зависимость $R_{Rec}(\Theta)$ от итерации обучения

Коэффициент регуляризации α_2	Среднее количество ненулевых компонент
$\alpha_2 = 0.1$	9.67
$\alpha_2 = 0.15$	4.86
$\alpha_2 = 0.2$	3.18
$\alpha_2 = 0.5$	1.13

Рис. 4.3. Количество ненулевых компонент в представлении документа по завершению обучения

Рис. 4.4. Эксперимент №1: Разреженность

Было замечено что несмотря на значение параметра $t_{min} = 1$, количество ненулевых компонент для различных α_2 перестает уменьшаться (например, для $\alpha_2 = 0.15$ количество ненулевых компонент останавливается на пяти), таким образом модель сама определяет оптимальную степень разреженности в соответствии с заданным набором гиперпараметров.

Аналогичный эксперимент был поставлен для проверки изменения мер негауссовости. Графики с рис.4.7 демонстрируют поведение функции потерь, состоящей только из членов $R_{rec}(\Theta)$ и $R_{Ind}(\Theta)$. Были сделаны выводы о несостоятельности коэффициента эксцесса как меры регуляризации независимости. Согласно [16], коэффициент эксцесса сильно зависит от выбросов в данных, и его значение может зависеть лишь от нескольких наблюдений в хвостах распределения.

Интересным является то как изменение каждого из приближений влияет на количество ненулевых компонент - из самого факта независимости тематик не обязательно следует разреженность, но это именно то свойство которое нам хочется индуцировать, поэтому. Из графика видно что третье приближение более других влияет на количество ненулевых элементов.

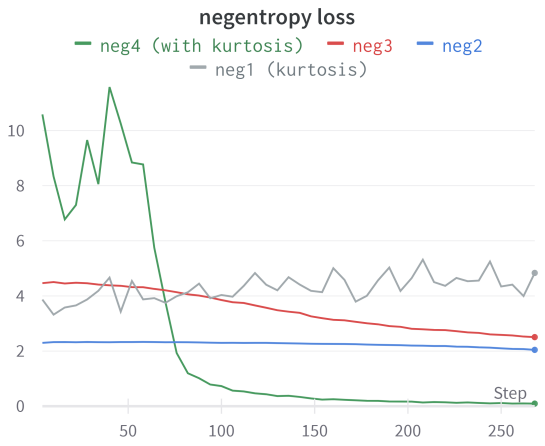

Рис. 4.5. Зависимость $R_{Ind}(\Theta)$ от итерации обучения

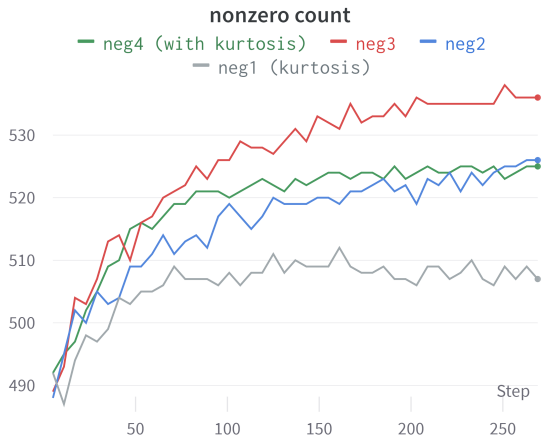

Рис. 4.6. Зависимость количества ненулевых компонент от итерации обучения

Рис. 4.7. Эксперимент №2: Независимость

График4.8 демонстрирует поведение функции потерь и её составляющих при обучении со всеми регуляризационными членами. Коэффициенты α_i подбирались так чтобы каждое слагаемое было одного порядка. Наблюдается что у модели **SpaRT** с трудом получается одновременно

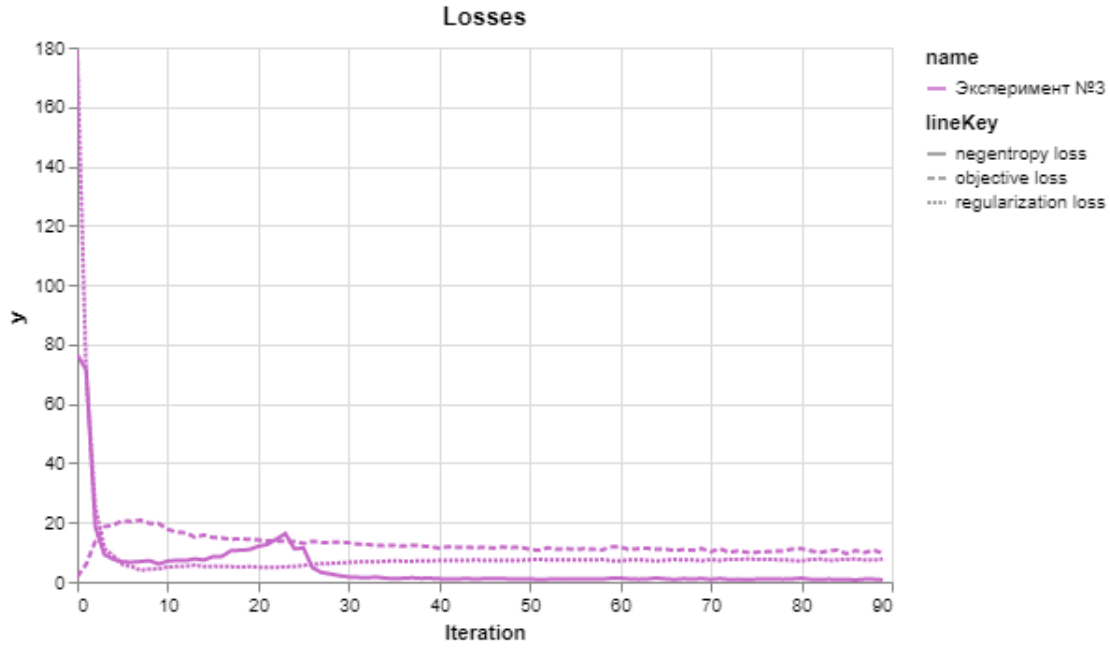

Рис. 4.8. Члены функции потерь при обучении модели SpaRT

минимизировать оба члена регуляризации $R_{Sp}(\Theta)$ и $R_{Ind}(\Theta)$, хотя согласно эвристике соответствующие им свосства должны положительно коррелировать.

Далее мы проверим, коррелируют ли свойства построенных векторных представлений с метриками тематического моделирования.

Первую выдвигаемую гипотезу можно сформулировать следующим образом: *при обучении модели с функцией потерь, включающей член минимизирующий взаимную информацию – максимизирующий приближение негэнтропии – метрика, соответствующая разнообразию тематик, должна также расти.* Она опирается на тот эмпирический факт, что чем более независимы тематики, тем более они разнообразны. Для проверки этой гипотезы был проведен эксперимент, в котором $\mathbf{SpaRT_{BASE}}$ сравнивается с моделью, обученной с членом максимизирующим негэнтропию. Было обучено четыре модели, с каждой из мер негауссовости, описанной в разделе (2.3.2). Результаты экперимента описываются в таблице 4.1.

Мера негауссовости	Topic Coherence (NPMI)	Topic Diversity
Без NG ($\mathbf{SpaRT_{BASE}}$)	0.079	0.609
NG_{kurt}	0.076	0.610
NG^{C}_{neg}	0.078	0.618
$\mathrm{NG}^{G_1}_{neg}$	0.074	0.630
$\mathrm{NG}^{G_2}_{neg}$ ($\mathbf{SpaRT_{NEG}}$)	0.073	0.639

Таблица 4.1. Эксперимент №4. Значения метрик при обучении с различными мерами негауссовости

Результаты эксперимента показывают, что за счет обучения модели с максимизацией негауссовости достигается независимость компонент и в интуитивном понимании: слова из тематик, определяемых вектор-столбцами в векторном пространстве эмбеддингов, действительно отличаются по смыслу и описывают разные идеи. Модель с наивысшим значением метрики TD обозначим $\mathbf{SpaRT_{NEG}}$.

Следующая гипотеза звучит следующим образом: *метрика, описывающая согласованность тематик, неявно должна расти при одновременном увеличении их независимости и разреженности.* С одной стороны, чем больше отличаются тематики между собой, тем более конкретизированными они должны быть. С другой – чем меньше документов соответствует тематике, тем более она конкретизирована и в меньшей степени описывает мотивы, встречающиеся в большом количестве документов. В таблице 4.2 приведены результаты эксперимента, поставленного для проверки этой гипотезы - в нем модель $\mathbf{SpaRT_{BASE}}$ дообучалась с обоими регуляризационными членами.

Заметим, что метрика Topic Coherence значительно выросла при обучении с обоими регуляризационными членами по сравнению с базовой моделью $\mathbf{SpaRT_{BASE}}$, однако Topic Diversity, напротив, уменьшилась.

Мера негауссовости	Topic Coherence (NPMI)	Topic Diversity
Без NG ($\mathbf{SpaRT_{BASE}}$)	0.079	0.609
NG^{C}_{neg}	0.085	0.578
$\mathrm{NG}^{G_1}_{neg}$	0.083	0.581
$\mathrm{NG}^{G_2}_{neg}$	0.083	0.588

Таблица 4.2. Эксперимент №5. Значения метрик при обучении с l1-регуляризацией и различными мерами негауссовости

Таким образом, из экспериментов 4.1 и 4.2, следует вывод о том что при дообучении модели $\mathbf{SpaRT_{BASE}}$ не получится одновременно увеличить обе метрики TC и TD.

4.3. Сравнение с другими тематическими моделями

Для оценки модели **SpaRT** она была обучена с обоими регуляризационными членами $R_{Sp}(\Theta)$ и $R_{Ind}(\Theta)$ с соответствующими коэффициентами $\alpha_2 = 0.01$ и $\alpha_3 = 10.0$. Сперва проверим качество работы модели на предложенных в разделе 4.1 метриках и сравним результаты с другими тематическими моделями. Результаты приведены в таблице 4.5 – модель **SpaRT** показывает результаты, сравнимые с **CTM**.

Для оценки тематической модели было решено в полной мере не проверять согласованность смысловой составляющей документов и выделенных тематик – такая задача требует наличия меток в датасете, и даже в случае их наличия они бы не давали точного представления о качестве выделенных тематик. Однако беглый анализ примера выделенных тематик показал адекватность модели: в таблице 4.9 приведены примеры выделенных с помощью **SpaRT** слов для одной из тематик, и документы, соответствующие ей с наибольшими коэффициентами.

Модель	Topic Coherence (NPMI)	Topic Diversity
SpaRT$_{\mathbf{BASE}}$	0.079	0.609
SpaRT$_{\mathbf{NEG}}$	0.073	**0.639**
SpaRT	**0.085**	0.578
LDA	-0.029	0.340
ETM	0.062	0.221
BERTopic	**0.089**	0.595

Таблица 4.3. Значения метрик для различных тематических моделей

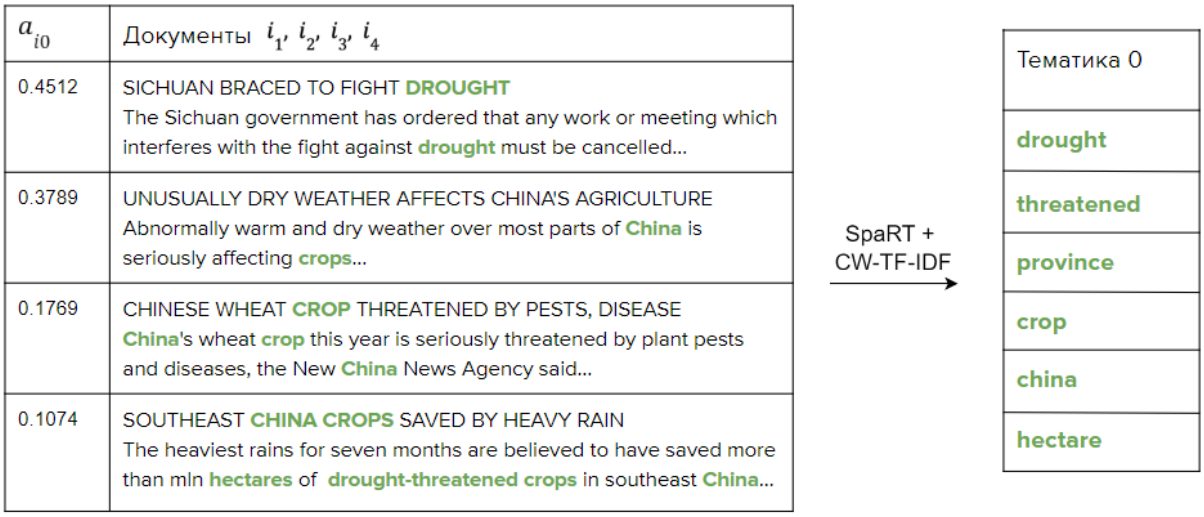

Рис. 4.9. Пример тематики и документов, соответствующих ей с наибольшими коэффициентами

Выводы и результаты по главе

Итак, для модели **SpaRT** была проверена выполняемость свойств разреженности и независимости. Были выбраны метрики тематического моделирования, с помощью которых былм проверены и подтверждены гипотезы о связи их с упомянутыми ранее свойствами. Модель была сравнена с другими тематическими моделями и показала качественные результаты.

Заключение

4.4. Результаты

В рамках данной работы были изучены существующие методы и особенности построения векторных пространств слов и документов, а так же исследованы способы их применения для задачи тематического моделирования.

Для получения векторных представлений документов и тематик была разработана модель **SpaRT** на основе нейронной сети с использованием архитектуры трансформера. Удалось построить функцию потерь, которая с помощью регуляризации весов и приближений негэнтропии неявно индуцирует свойства разреженности и независимости компонент в пространстве векторных представлений, а так же использовать её для обучения модели.

Идеи классических методов статистического анализа документов помогли сформулировать метод **CW-TF-IDF** для извлечения наиболее коррелирующих с тематикой слов из её векторного представления. С его помощью построенную модель можно тестировать и оценивать как тематическую.

Были проведены эксперименты для проверки свойств семантического пространства построенных векторных представлений. Так же были проверены гипотезы о связи свойств разреженности и независимости признаков с улучшением выбранных метрик тематического моделирования, что показало состоятельность модели для этой задачи. Проведен сравнительный анализ результатов построенной модели и других тематических моделей, таких как LDA, ETM и BERTopic.

4.5. Выводы и направления дальнейших исследований

Выбранные для функции потерь приближения оказались хорошим регуляризаторами свойств разреженности и независимости. Удалось показать, что математические меры этих свойств коррелируют с метриками тематического моделирования.

На задаче тематического моделирования построенная модель показала приемлимые результаты, и качественно сравнима с другими моделями, использующими векторные представления документов.

Свойства независимости и разреженности – не единственное, что можно пытаться индуцировать в векторном пространстве представлений. От него можно так же требовать выполнения геометрических свойств – например, линейность преобразования $F(x)$. В данной работе также проводились эксперименты с добавлением члена регуляризации:

$$R_{Lin}(\Theta) = \frac{1}{|s||s-1|}\sum_{i=1}^{|s|}\sum_{k=1}^{i}(||x_i - x_k||_2 - ||f(x_i) - f(x_k)||_2)^2$$

Обучение с ним оказалось слишком длительным и неэффективным.

Стоит отметить, что построенную модель можно рассматривать как модель информационного поиска. Опираясь подход, описанный в статье [11], мы можем построить разреженные представления документов, среди которых будет осуществляться поиск, и сформировать словарь инвертированных индексов. Впоследствии запрос также переводится в разреженное пространство, и похожесть запроса на документ считается как скалярное произведение их представлений. Наличие инвертированных индексов позволяет значительно ускорить подсчет произведения эмбеддингов.

Библиографический список

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

Vaswani, A. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin // Advances in neural information processing systems. — 2017. — Т. 30.

Bengio, Y. A neural probabilistic language model / Y. Bengio, R. Ducharme, P. Vincent // Advances in neural information processing systems. — 2000. — Т. 13.

Rumelhart, D. E. A model for analogical reasoning / D. E. Rumelhart, A. A. Abrahamson // Cognitive Psychology. — 1973. — Т. 5, № 1. — С. 1—28.

Mikolov, T. Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean // arXiv preprint arXiv:1301.3781. — 2013.

Dieng, A. B. Topic modeling in embedding spaces / A. B. Dieng, F. J. Ruiz, D. M. Blei // Transactions of the Association for Computational Linguistics. — 2020. — Т. 8. — С. 439—453.

Le, Q. Distributed representations of sentences and documents / Q. Le, T. Mikolov // International conference on machine learning. — PMLR. 2014. — С. 1188—1196.

Angelov, D. Top2Vec: Distributed Representations of Topics / D. Angelov. — 2020.

Dai, A. M. Document embedding with paragraph vectors / A. M. Dai, C. Olah, Q. V. Le // arXiv preprint arXiv:1507.07998. — 2015.

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

Zamani, H. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing / H. Zamani, M. Dehghani, W. B. Croft, E. Learned-Miller, J. Kamps // Proceedings of the 27th ACM international conference on information and knowledge management. — 2018. — С. 497—506.

Hyvarinen, A. Independent component analysis, a tutorial / A. Hyvarinen // http://www. cs. helsinki. fi/u/ahyvarin/papers/NN00new. pdf. — 1999.

Cover, T. M. Information theory and statistics / T. M. Cover, J. A. Thomas // Elements of information theory. — 1991. — Т. 1, № 1. — С. 279—335.

Hofstätter, S. Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation / S. Hofstätter, S. Althammer, M. Schröder, M. Sertkan, A. Hanbury. — 2020.

Jones, M. C. What is projection pursuit? / M. C. Jones, R. Sibson // Journal of the Royal Statistical Society: Series A (General). — 1987. — Т. 150, № 1. — С. 1—18.

Huber, P. J. Projection pursuit / P. J. Huber // The annals of Statistics. — 1985. — С. 435—475.

Hyvärinen, A. New approximations of differential entropy for independent component analysis and projection pursuit / A. Hyvärinen // Advances in neural information processing systems. — 1997. — Т. 10.

Bouma, G. Normalized (pointwise) mutual information in collocation extraction / G. Bouma // Proceedings of GSCL. — 2009. — Т. 30. — С. 31—40.

Приложение

В приложенной таблице описаны параметры основных вариантов модели, с которыми проводились эксперименты из раздела 4.2.

$\mathbf{SpaRT_{BASE}}$	$\mathbf{SpaRT_{NEG}}$	**SpaRT**
FROM: SCRATCH	FROM: $\mathbf{SpaRT_{BASE}}$	FROM: $\mathbf{SpaRT_{BASE}}$
MIN_TOPICS (t_{min})=1	MIN_TOPICS (t_{min})=1	MIN_TOPICS (t_{min})=2
NUM_TOPICS=1000	NUM_TOPICS=1000	NUM_TOPICS=1000
LR=10^{-3}	LR=10^{-4}	LR=10^{-4}
BATCH_SIZE=64	BATCH_SIZE=64	BATCH_SIZE=64
EPOCHS=10	EPOCHS=5	EPOCHS=5
α_1=1.0	α_1=1.0	α_1=1.0
α_2=0.0	α_2=0.0	α_2=0.1
α_3=0.0	α_3=100.0	α_3=100.0

Таблица 4.4. Значения параметров при обучении вариантов модели

frost	gasoline	housing	display	antibiotic
temperature	rule	bill	redeploy	salmonella
wheat	crisis	rural	tentiative	victim
cold	energy	insurance	tv	resistant
survive	oil	authorization	video	infection
minus	tax	urban	accounted	researcher

Таблица 4.5. Примеры тематик, выделенных моделью из датасета новостей Reuters