Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Стенин Сергей Сергеевич

Мультиграммные аддитивно регуляризованные тематические модели

010990 — Интеллектуальный анализ данных

Выпускная квалификационная работа магистра

Научный руководитель:
д. ф.-м. н.
Воронцов Константин Вячеславович

Москва
2015

Содержание

1 Введение **4**

2 Общая теория для близких работ **7**

2.1 Униграммные модели 7

2.2 Мультиграммные модели 8

2.3 Биграммная тематическая модель 10

2.4 Словосочетания в латентном размещении Дирихле 11

2.5 Мультиграмная тематическая модель 13

2.6 Пути улучшения существующих моделей 15

3 Постановка задачи **16**

4 Предлагаемая модель **16**

4.1 Алгоритм поиска локальных оптимумов 19

4.2 Связь с моделью TNG 20

4.3 Биграммная модель как мультимодальная 20

5 Эксперимент **22**

5.1 Данные 22

5.2 Сравнение интерпретируемости униграммной TM и Bigram-ARTM . . . 22

5.3 Сглаживающий регуляризатор для выделения фоновых тем 24

6 Заключение **26**

Список литературы **27**

Аннотация

В работе рассматривается задача построения мультиграммных аддитивно регуляризованных тематических моделей. Задача тематического моделирования является некорректно поставленной, а униграммные тематические модели, основанные на гипотезе мешка слов, плохо интерпретируются. Предлагается тематическая модель Bigram-ARTM, которая при построении учитывает как слова, так и биграммы, а также использует аддитивную регуляризацию при решении задачи тематического моделирования. Предложен эффективный алгоритм, находящий приближение оптимального решения в Bigram-ARTM. Показаны результаты применения модели к коллекции тезисов конференций ИОИ-ММРО за 2007-2013 годы.

Ключевые слова: *тематические модели, аддитивная регуляризация, языковые модели, мультиграммы, PLSA.*

1 Введение

Вероятностные тематические модели (далее ВТМ или просто ТМ) — одно из современных приложений машинного обучения к анализу коллекций текстовых документов [1]. ТМ предназначены для определения того, какие темы встречаются в заданном документе и какие слова образуют между собой темы. Темы рассматриваются как дискретные вероятностные распределения над множеством слов коллекции, а документы — как дискретные вероятностные распределения надо множеством тем. Самыми известными тематическими моделями являются метод вероятностного латентного семантического анализа (Probabilistic Latent Semantic Analisys, PLSA) [2] и латентное размещение Дирихле (Latent Dirichlet Allocation, LDA) [1].

Обе модели можно также рассматривать как инструмент решения задачи неотрицательного матричного разложения [3, 4]. А именно, разложения матрицы частот слов в документах в произведение двух стохастических матриц — вероятностей появления слов в темах, и вероятностей появления тем в документах. Но такая задача является некорректно поставленной, поэтому при ее решении используется регуляризация. Существует несколько подходов к регуляризации.

Первый подход основывается на наложении априорных распределений на столбцы матриц распределения слов в темах и тем в документах с последующим байесовским выводом апостериорных вероятностей. Такой подход сейчас уже является классическим. Первый раз он был предложен в модели LDA, где в качестве априорного распределения было выбрано распределение Дирихле. В основном из-за того, что это распределение, из которого получаются случайные вектора заданной размерности с неотрицательными компонентами, нормированными на единицу в норме $l1$. Позднее этот подход использовался в большей части работ по тематическому моделированию. Приведем ссылки лишь на некоторые из них [5–9]. Заметим, что байесовский подход к регуляризации требует использования плотностей некоторых априорных распределений, что накладывает ограничения неотрицательности и нормированности.

Второй подход появился сравнительно недавно и впервые встречается в работе [10]. Этот подход отказывается от байесовского вывода в пользу регуляризации по Тихонову [11] и предлагает добавлять к целевой функции оптимизационной задачи — к логарифму правдоподобия коллекции текстовых документов — функции, назы-

ваемые *регуляризаторами*, наличие каждого из которых обеспечивает определенные свойства решения задачи тематического моделирования. Из-за того, что к целевой функции можно добавлять сразу несколько регуляризаторов независимо друг от друга, подход получил название АРТМ — аддитивная регуляризация тематических моделей. Обзор различных типов регуляризаторов можно найти в работах [10, 12]. В данной работе при построении ВТМ будет использоваться метод АРТМ.

Второй важной особенность, которая часто присутствует в ТМ, является использование так называемой гипотезы мешка слов («Bag of words», BoW). Гипотеза состоит в том, что тему в документе можно определить, даже если слова в нем переставить случайным образом. Понятно, что такая гипотеза неправдоподобна — человеку будет практически невозможно понять, о чем текст, если в нем перемешать слова. Поэтому с 2006 года начали появляться тематические модели, которые отказывались от гипотезы мешка слов [5, 13, 14]. Эти работы показывают, что отказ от гипотезы мешка слов позволяет выделять из коллекций текстовых документов темы с такими распределениями слов и словосочетаний, что наиболее вероятными в каждой из выделенных тем является набор слов, который воспринимается человеком как тематически связанное множество.

Целями данной работы является переформулировка известных байесовских мультиграммных моделей в терминах АРТМ, их систематизация и обобщение, а также построение вычислительно эффективной мультиграммной аддитивно регуляризованной тематической модели.

В работе предлагается тематическая модель Bigram-ARTM, которая одновременно использует гипотезу мешка слов и гипотезу мешка фраз (биграмм), а для регуляризации использует метод АРТМ, основанный на классической регуляризации. Показана связь модели Bigram-ARTM и модели TNG. Также установлено, что предлагаемая модель — частный случай мультимодальной модели с регуляризацией. Работа Bigram-ARTM проиллюстрирована на коллекции текстов тезисов конференции ИОИ-ММРО за 2007-2013 года.

Работа организована следующим образом. В разделе 2 строится общая теория существующих мультиграммных ТМ и описываются способы их усовершенствования. В разделе 3 описывается постановка задачи. В разделе 4 описывается предлагае-

мая модель Bigram-ARTM, доказывается теорема об условиях, выполненных в точке локального оптимума, показывается связь Bigram-ARTM с моделью TNG и с мультимодальной TM. В разделе 5 приводятся результаты вычислительного эксперимента на коллекции тезисов конференций ИОИ и ММРО.

2 Общая теория для близких работ

2.1 Униграммные модели

Обозначим через D множество заданных текстовых документов. Это множество также будем называть *коллекцией*. Через W обозначим множество слов, которые встречаются в коллекции D. W также будем называть *словарем*. Тогда документ $d \in D$ представляет последовательность слов или терминов $(w_1, \ldots, w_{n_d})$, где n_d — число слов в документе, слова могут встречаться в документе несколько раз. Через n_{wd} будем обозначать, сколько раз слово w встретилось в документе d. Матрицу счетчиков n_{wd} слов в документах будем обозначать через $N = \Big(n_{wd}\Big)_{W\times D}$, а матрицу N, в который каждый из столбцов нормирован на единицу в норме $l1$ будем называть матрицей частот слов в документах и обозначать через $P = \Big(p(w|d)\Big)_{W\times D}$.

В вероятностном тематическом моделировании предполагается, что появление слова w в документе d связано с неизвестной скрытой темой t. Коллекция D рассматривается как набор троек (d, w, t), которые получены независимо из распределения $p(d, w, t)$. Распределение $p(d, w, t)$ задано на конченом множестве $D \times W \times T$, где T — множество всех рассматриваемых скрытых тем. Гипотеза независимости троек (d, w, t) также называется гипотезой мешка слов (*Bag of words, BoW*), так как согласно этой гипотезе порядок терминов в документе не важен для определения тематики документа, его можно восстановить, даже если слова в документе переставить в случайном порядке.

Под построением тематической модели при заданном числе тем $|T|$ понимают нахождение скрытых условных распределений слов в темах $p(w|t)$ и тем в документах $p(t|d)$ по наблюдаемым распределениям частот слов в документах $p(w|d)$, полученных нормированием матрицы N. При работе с BTM также принято предполагать, что распределение слов в теме не зависит от рассматриваемого документа, то есть для всех документов одинаково. Это предположение называется *гипотезой условной независимости*. Гипотеза условной независимости может быть выражена в виде

$$p(w|d, t) = p(w|t). \tag{1}$$

Из гипотезы условной независимости можно получить разложение условной вероятности слова в документе

$$p(w|d) = \sum_{t \in T} p(w, t|d) = \sum_{t \in T} p(w|t, d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d). \tag{2}$$

Договоримся об обозначениях. Условные вероятности слов в скрытых темах будем обозначать $p(w|t) = \phi_{wt}$ и представлять в виде матрицы $\Phi = \big(\phi_{wt}\big)_{W \times T}$. Аналогично условные вероятности тем в документах будем обозначать $p(t|d) = \theta_{td}$ и представлять в виде матрицы $\Theta = \big(\theta_{td}\big)_{T \times D}$. Также договоримся о том, что BTM, в которых рассматриваются только отдельные слова, будем называть *униграммными*.

Пусть нам задана коллекция документов D и словарь W. Тогда правдоподобие коллекции как функцию матриц Φ и Θ можно записать как

$$p(D; \Phi, \Theta) = \prod_{d \in D} \prod_{w \in d} p(w, d)^{n_{dw}} = \prod_{d \in D} \prod_{w \in d} p(w|d)^{n_{dw}} \underbrace{p(d)^{n_{dw}}}_{const(\Phi, \Theta)} \to \max_{\Phi, \Theta}. \tag{3}$$

Используя гипотезу условной независимости и логарифмирование, запишем логарифм правдоподобия коллекции документов и условия неотрицательности и нормировки на распределения

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi, \Theta};$$

$$\sum_{w \in W} \phi_{wt} = 1, \quad \phi_{wt} \geq 0 \qquad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0 \tag{4}$$

Так выглядит основная задача BTM для униграммных моделей [1, 2].

На задачу построения BTM можно смотреть как на задачу неотрицательного матричного разложения матрицы частот слов в документах $P = \Big(p(w|d)\Big)_{W \times D}$ в произведение матриц $P \approx \Phi\Theta$. Заметим, что для всякой невырожденной матрицы S, которая сохраняет стохастичность найденного разложения, выполнено

$$P \approx \Phi\Theta = \Phi S S^{-1}\Theta = \widetilde{\Phi}\widetilde{\Theta}.$$

То есть существует бесконечно много различных решений (матриц $\widetilde{\Phi}$ и $\widetilde{\Theta}$) таких, что их произведение дает одно и то же приближение к матрице P. Для того, чтобы ограничить число степеней свободы в задаче применяется регуляризация.

2.2 Мультиграммные модели

BTM, в которых рассматриваются вхождения не только слов, но и наборов слов длины до n включительно будем называть *n-граммными* или *мультиграммными*

BTM. Модели с $n = 2$ будем называть *биграммными*.

Пусть теперь в коллекции документов D мы рассматриваем не отдельные слова, а, скажем, пары слов. Рассмотрим, как меняется правдоподобие коллекции. Поскольку пары слов в документе — события, которые не являются независимыми, то надо понять, как можно записать правдоподобие коллекции.

$$p(D) = \prod_{d \in D} p(w_{n_d}, w_{n_d-1}, \ldots, w_2, w_1|d)p(d) = C \prod_{d \in D} p(w_{n_d}, w_{n_d-1}, \ldots, w_2, w_1|d) \tag{5}$$

Здесь через n_d обозначено количество слов в документе d, а вероятности документов будем считать постоянными величинами, произведение которых равно C.

Для биграммных моделей для упрощения формул условных вероятностей вводится требование марковости. Учтем это требование и перепишем правдоподобие слов документа d.

$$p(w_{n_d}, w_{n_d-1}, \ldots, w_2, w_1|d) = p(w_{n_d}|w_{n_d-1}, d)p(w_{n_d-1}|w_{n_d-2}, d) \ldots p(w_2|w_1, d)p(w_1|d) \tag{6}$$

Теперь введем дополнительные переменные. Через v будем обозначать слово, которое предшествует слову w в тексте. Для каждого слова w определим величину x_{vw}, которая равна 1, если v и w образуют биграмму и нулю в противном случае

$$x_{vw} = \begin{cases} 1, & v \text{ и } w \text{ бразуют биграмму;} \\ 0, & v \text{ и } w \text{ независимы.} \end{cases} \tag{7}$$

Тогда для пар подряд идущих слов v, w, которые не образуют биграмму выполнено

$$p(w|v, d) = p(w|d) \tag{8}$$

то есть вероятность появления слова w после слова v не зависит от слова v. Отсюда получаем, что

$$p(w|v, d) = x_{vw} \cdot p(w|v, d) + (1 - x_{vw}) \cdot p(w|d). \tag{9}$$

Из это формулы видно, что переменной x_{vw} можно приписать вероятностный смысл и получить более общую модель. Так и сделаем, договорившись, что

$$x_{vw} \in [0, 1]$$

Перепишем правдоподобие коллекции документов

$$p(D) = C \prod_{d \in D} \prod_{w \in d} \prod_{v \in d} \Big(x_{vw} \cdot p(w|v, d) + (1 - x_{vw}) \cdot p(w|d) \Big)^{n_{vwd}}. \tag{10}$$

Через n_{vwd} обозначено число раз, когда пара слов v, w встретилась подряд в документе d (для первого слова w_1 в каждом документе будем считать, что $x_{w_1} = 0$). Здесь надо дополнительно договориться об обозначениях. В предыдущей формуле формально записано произведение $\prod\limits_{w \in d} \prod\limits_{v \in d}$, то есть произведение по всем парам слов. На самом деле, из-за наличия степени n_{vwd} неединичные множители в произведение могут давать только слова, стоящие в документе рядом. Учитывая этот факт, иногда произведение $\prod\limits_{w \in d} \prod\limits_{v \in d}$ будет записываться как $\prod\limits_{(v,w) \in d}$ или $\prod\limits_{vw \in d}$. При логарифмировании знаки произведений будут заменяться на соответствующие знаки суммы.

Формула выше дает возможность взглянуть на мультиграммные BTM под более широким углом, так как многие уже существующие модели описываются таким разложением правдоподобия коллекции. Рассмотрим основные работы в области мультиграммных BTM, чтобы понять, на каких предположениях предлагаемая в данной работе теория описывает каждую из уже существующих моделей.

2.3 Биграммная тематическая модель

Снова будем полагать, что появление слова w ассоциировано с некоторой скрытой темой t, но теперь это появление ассоциировано еще и с предыдущим словом v

$$p(w|v,d) = \sum_{t \in T} p(w,t|v,d) = \sum_{t \in T} p(w|t,v,d) p(t|v,d) \tag{11}$$

Модификация гипотезы условной независимости для биграммных моделей

$$\begin{aligned} p(w|t,v,d) &= p(w|t,v), \\ p(t|v,d) &= p(t|d), \end{aligned} \tag{12}$$

то есть распределение слов в теме при заданном предыдущем слове не зависит от документа, который мы сейчас рассматриваем и распределение тем в документе не зависит от того, какое слово v предшествует слову w, которое мы сейчас рассматриваем. Тогда условную вероятность слова w можно переписать так

$$p(w|v,d) = \sum_{t \in T} p(w|v,t) p(t|d). \tag{13}$$

Обозначим условные вероятности $p(w|v,t)$ через ϕ_{wvt} и будем хранить их в трехмерной матрице $\widetilde{\Phi} = \left(\phi_{wvt}\right)_{W \times W \times T}$. Матрицу $\widetilde{\Phi}$ можно рассматривать как набор из W

плоских матриц $\Phi_v = \big(\phi_{wvt}\big)_{W\times T}$, каждая из которых — матрица распределений слов в темах при фиксированном предыдущем слове v. Условные вероятности тем в документах по-прежнему будем обозначать $p(t|d) = \theta_{td}$ и представлять в виде матрицы $\Theta = \big(\theta_{td}\big)_{T\times D}$.

Первая вероятностная тематическая модель, которая включает в себя n-граммы, была предложена Ханной Валлах в [5] и называлась Bigram topic model (Bigram TM). Основное предположение Bigram TM состояло в том, что каждая пара подряд идущих слов в документе является биграммой и что вероятность появления следующего слова в документе зависит только от скрытой темы и от предыдущего слова, которое уже известно. То есть $x_{vw} = 1$ для всех слов и документов, поэтому условные вероятности $p(w|d)$ исключаются из рассмотрения. Таким образом, задача построения BTM сводится к задаче оптимизации логарифма правдоподобия коллекции D:

$$L(D;\widetilde{\Phi},\Theta) = \sum_{d\in D}\sum_{(v,w)\in d} n_{vwd}\ln\Big(\sum_{t\in T}\phi_{wvt}\theta_{td}\Big) \tag{14}$$

Итого, предположения

1. Марковость алгоритма порождения коллекции
2. Все пары слов образуют биграммы.
3. Гипотеза условной независимости $p(w|v,t,d) = p(w|v,t)$
4. Гипотеза условной независимости $p(t|v,d) = p(t|d)$

приводят к модели Bigram TM и к оптимизационной задаче

$$\begin{gathered}\sum_{d\in D}\sum_{(v,w)\in d} n_{vwd}\ln\sum_{t\in T}\phi_{wvt}\theta_{td} \to \max_{\Phi,\Theta}\\ \sum_{w\in W}\phi_{wvt} = 1, \quad \phi_{wvt}\geq 0\\ \sum_{t\in T}\theta_{td} = 1, \quad \theta_{td}\geq 0\end{gathered} \tag{15}$$

2.4 Словосочетания в латентном размещении Дирихле

В работе [13] предлагается модель LDA Collocation model (LDACOL) — модель словосочетаний в латентном размещении Дирихле.

Модель LDACOL первой вводит множество переменных x_{vw} («меток биграммности») Эти переменные считаются случайными величинами. LDACOL имеет возможность решать — образовать биграмму или просто слово. В этом плане она более реалистична, чем Bigram TM, которая всегда генерирует биграммы. Предполагается, что x_1 наблюдаема, и что только униграммы разрешены в начале документа. Если надо наложить на модель больше ограничений (например, требование отсутствия биграмм в начале/конце предложения/параграфа), можно считать соответствующие метки биграммности тоже наблюдаемыми. Также считается, что x_{vw} порождается из распределения Бернулли с параметром ψ_v, а если метка $x_{vw} = 1$ (то есть v и w образуют биграмму), то следующее слово порождается из распределения $p(w|v)$, которое не зависит ни от каких скрытых тем. Распределения $p(w|v)$ хранятся в матрице $\Sigma = (\sigma_{vw})_{W \times W}, \quad p(w|v) = \sigma_{vw}$. Алгоритм порождения коллекции в этой модели такой

1. Для каждого документа d, для каждого слова w в документе d:

   (a) Получить метку биграммности x_{vw} из имеющегося распределения Бернулли ψ_v

   (b) Получить тему t из дискретного распределения тем в документе θ_d

   (c) Получить случайное слово w из имеющегося дискретного распределения σ_v если метка биграммности $x_{vw} = 1$, иначе
   получить случайное слово из w из дискретного распределения ϕ_t.

Перепишем правдоподобие коллекции

$$p(D) = C \prod_{d \in D} \prod_{(v,w) \in d} \Big(x_{vw} \cdot p(w|v,d) + (1 - x_{vw}) \cdot p(w|d) \Big)^{n_{vwd}}. \tag{16}$$

В модели LDACOL принимается предположение о том, что условное распределение $p(w|v,d)$ для всех документов одинаково, то есть $p(w|v,d) = p(w|v)$, а вот вероятность появления слова w в документе d ассоциирована с выпадением одной из скрытых тем, то есть $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$. Учет меток биграммности x_{vw} приводит в такому виду правдоподобия

$$p(D) \propto \prod_{d \in D} \prod_{(v,w) \in d} \Big(x_{vw} p(w|v) + (1 - x_{vw}) \sum_{t \in T} p(w|t) p(t|d) \Big)^{n_{vwd}}. \tag{17}$$

Итого, предположения

1. Марковость алгоритма порождения коллекции
2. Выборочное использование меток биграммности
3. Гипотеза условной независимости $p(w|t,d) = p(w|t)$

приводят к оптимизационной задаче с ограничениями равенства и неравенства

$$\mathcal{L}(D; \Phi, \Theta, \Sigma) = const + \sum_{d \in D} \sum_{(v,w) \in d} n_{vwd} \cdot \ln \Big(x_{vw} \sigma_{vw} + (1 - x_{vw}) \sum_{t \in T} \phi_{wt} \theta_{td} \Big) \to \max_{\Phi, \Theta, \Sigma, x}$$

$$\sum_{w \in W} \sigma_{vw} = 1, \quad \sigma_{vw} \geq 0$$

$$\sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0$$

$$\sum_{v \in W} \phi_{vt} = 1, \quad \phi_{vt} \geq 0$$

$$x_{vw} \in [0, 1] \tag{18}$$

2.5 Мультиграмная тематическая модель

В статье [14] предлагается мультиграммная тематическая модель Topical N-gramm model (TNG), которая автоматически определяет слова и фразы, основываясь на контексте и относит смесь тем как к отдельным словам, так и к мультиграммным фразам. То есть это самая общая из всех существующих моделей.

В TNG предполагается, что метка биграммности x_{vw} генерируется из распределения $\psi_{t_v v}$, которое для каждой пары «предыдущее слово v, тема предыдущего слова t_v» свое собственное. Если метка биграммности равна единице, то генерируется тема t из распределения θ_d, слово w генерируется из распределения $p(w|v,t) = \varphi_{wvt}$ Алгоритм порождения коллекции следующий. Для каждого слова w в документе d

1. получить метку биграммности x_{vw} из распределения Бернулли $\psi_{t_v v}$;
2. получить тему t из дискретного распределения θ_d;

3. если метка биграммности $x_{vw} = 1$ то сгенерировать случайное слово w из вектора дискретного распределения φ_{tv}, иначе сгенерировать случайное слово w из дискретного распределения слов в теме ϕ_t.

В модели TNG обе условные вероятности $p(w|v,d)$ и $p(v|d)$ связаны со скрытыми темами $t \in T$. Для условной вероятности $p(w|v,d)$ получаем такие тождества

$$p(w|v,d) = \sum_{t \in T} p(w,t|v,d) = \sum_{t \in T} p(w|t,v,d)p(t|v,d). \tag{19}$$

Здесь гипотеза условной независимости выглядит как набор из двух тождеств

$$\begin{aligned} p(w|t,v,d) &= p(w|t,v) \\ p(t|v,d) &= p(t|d), \end{aligned} \tag{20}$$

поэтому

$$p(w|v,d) = \sum_{t \in T} p(w|t,v)p(t|d) = \sum_{t \in T} \varphi_{wtv}\theta_{td}. \tag{21}$$

Для вероятности $p(w|d)$ выполнены предположения, что появление слова w связано с множеством скрытых тем $t \in T$, гипотеза условной независимости — как в униграммных моделях.

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}. \tag{22}$$

Запишем правдоподобие коллекции D в модели TNG.

$$\begin{aligned} p(D) &\propto \prod_{d \in D} \prod_{(v,w) \in d} \Big(x_{vw} p(w|v,d) + (1 - x_{vw})p(w|d)\Big)^{n_{vwd}} \propto \\ &\propto C \prod_{d \in D} \prod_{(v,w) \in d} \Big(x_{vw} \sum_{t \in T} p(w|t,v)p(t|d) + (1 - x_{vw}) \sum_{t \in T} p(w|t)p(t|d)\Big)^{n_{vwd}} \propto \\ &\propto \prod_{d \in D} \prod_{(v,w) \in d} \Big(x_{vw} \sum_{t \in T} \varphi_{wtv}\theta_{td} + (1 - x_{vw}) \sum_{t \in T} \phi_{wt}\theta_{td}\Big)^{n_{vwd}} \end{aligned} \tag{23}$$

Используемые при построении модели предположения:

1. марковость алгоритма порождения коллекции,
2. выборочное использование меток биграммности,
3. гипотеза условной независимости для униграммных слов $p(w|t,d) = p(w|t)$

4. гипотеза условной независимости для биграмм $p(w|t,v,d) = p(w|t,v)$

5. гипотеза условной независимости $p(t|v,d) = p(t|d)$.

Эти предположения приводят к оптимизационной задаче с ограничениями равенства и неравенства

$$\mathcal{L}(D;\Phi,\Theta,\widetilde{\Phi}) = \sum_{d\in D}\sum_{(v,w)\in d} n_{vwd}\ln\Big(x_{vw}\sum_{t\in T}\varphi_{wtv}\theta_{td} + (1-x_{vw})\sum_{t\in T}\phi_{wt}\theta_{td}\Big) \to \max_{\Phi,\Theta,\widetilde{\Phi},x}$$

$$\sum_{w\in W}\varphi_{wtv} = 1, \quad \varphi_{wtv} \geq 0$$

$$\sum_{t\in T}\theta_{td} = 1, \quad \theta_{td} \geq 0$$

$$\sum_{w\in W}\phi_{wt} = 1, \quad \phi_{wt} \geq 0$$

$$x_{vw} \in [0,1]. \tag{24}$$

2.6 Пути улучшения существующих моделей

Каждая из рассмотренных выше моделей в оригинальных работах [5,13,14] дополнительно использовала байесовский подход к регуляризации с наложением априорного распределения на оптимизируемые параметры модели. Предлагается обобщить этот подход и использовать метод APTM, основанный на регуляризации по Тихонову. Это снимет ограничения на неотрицательность и нормированность используемых функций регуляризации.Далее будет показано, что использование APTM позволяет предельно упростить нахождение условий для оптимального значения параметров модели.

В рамках данной работы рассматривался вариант усовершенствования модели LDACOL. Условные распределения $p(w|v) = \sigma_{vw}$ предлагалось восстанавливать, построив некоторую языковую модель коллекции текстовых документов. Изучение в рамках данной работы способов построения языковых, также называемых лингвистическими, моделей [15] позволило установить, что главная проблема, которая решается при построении языковых моделей состоит в том, чтобы выбрать оптимальный способ заполнения нулевых значений в матрице $p(w|v)$ на ненулевые. Это связано с тем, что в противном случае слово, которое не встретилось на обучении языковой

модели, будет иметь нулевую вероятность, а значит нулевую вероятность будет иметь правдоподобие тестового документа с таким словом. Для построения тематических моделей такой подход неэффективен, потому что требует хранения плотной матрицы вероятностей перехода от слова к слову, то есть квадратичного от размера словаря использования памяти. Поэтому в рамках данной работы такой путь развития модели LDACOL был отвергнут.

Модель TNG, которая является обобщением BigramTM и LDACOL, тоже является вычислительно неэффективной, потому что требует хранения в памяти меток биграммности для каждой из встреченных в коллекции пары слов, стоящих рядом. Предлагается отказаться от хранения меток биграммности, сделав некоторую нижнюю оценку максимизируемой в модели TNG функции.

3 Постановка задачи

Цель данного исследования состоит в том, чтобы построить вычислительно эффективную вероятностную тематическую модель, которая повысит интерпретируемость получаемых тем и позволит с помощью регуляризации задавать свойства, которыми будет обладать оптимальное решение задачи тематического моделирования. Интерпретируемость будет повышена с помощью использования и слов, и биграмм при построении TM, для регуляризации будет использоваться модель APTM, а вычислительная эффективность будет достигнута за счет меньшего числа переменных в модели.

4 Предлагаемая модель

Для начала рассмотрим предлагаемую модель без регуляризации. Предлагается в качестве целевого функционала в задаче тематического моделирования использовать сумму правдоподобий униграммной и биграммной модели. Не обязательно ограничиваться биграммами, модель обобщается и на словосочетания произвольной длины.

Итак, для построения ТМ будет решаться следующая задача оптимизации

$$\sum_{d\in D}\sum_{w\in d} n_{wd}\ln\Big(\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) + \sum_{d\in D}\sum_{v,w\in d} n_{vwd}\ln\Big(\sum_{t\in T}\varphi_{wtv}\theta_{td}\Big) \to \max_{\Phi,\widetilde{\Phi},\Theta}$$

$$\sum_{w\in W}\varphi_{wtv} = 1, \quad \varphi_{wtv} \geq 0$$

$$\sum_{t\in T}\theta_{td} = 1, \quad \theta_{td} \geq 0$$

$$\sum_{w\in W}\varphi_{wt} = 1, \quad \varphi_{wt} \geq 0$$

Здесь первая часть целевой функции — логарифм униграммного правдоподобия униграммной модели коллекции текстовых документов, а вторая часть — логарифм биграммного правдоподобия коллекции текстовых документов. Задача решается с ограничениями на неотрицательность и стохастичность матриц $\Phi, \Theta, \widetilde{\Phi}$.

Теперь добавим в модель регуляризацию в виде слагаемого к оптимизируемой функции. Получим окончательный вид предлагаемой модели - Bigram Additively Regularized Topic Model (Bigram-ARTM).

$$\sum_{d\in D}\sum_{w\in d} n_{wd}\ln\Big(\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) + \sum_{d\in D}\sum_{v,w\in d} n_{vwd}\ln\Big(\sum_{t\in T}\varphi_{wtv}\theta_{td}\Big) + R(\Phi,\widetilde{\Phi},\Theta) \to \max_{\Phi,\widetilde{\Phi},\Theta}$$

$$\sum_{w\in W}\varphi_{wtv} = 1, \quad \varphi_{wtv} \geq 0$$

$$\sum_{t\in T}\theta_{td} = 1, \quad \theta_{td} \geq 0$$

$$\sum_{w\in W}\varphi_{wt} = 1, \quad \varphi_{wt} \geq 0$$

Обозначим через $\underset{i\in I}{norm}(x)$ оператор нормирования вектора x, x_i — i-ая компонента вектора, I — множество индексов, по которым происходит нормировка. Оператор действует на компоненты вектора x следующим образом

$$\underset{i\in I}{norm}(x_i) = \frac{\max\{x_i, 0\}}{\sum\limits_{j\in J}\max\{x_j, 0\}}.$$

Если все элементы вектора неположительны, то оператор возвращает нулевой вектор.

Теорема 1. В точке максимума целевой функции для непрерывно дифференцируе-

мой функции R выполнены соотношения

$$p_{twd} = \underset{t\in T}{norm}\,\varphi_{wt}\theta_{td} \qquad p_{twvd} = \underset{t\in T}{norm}\,\varphi_{wvt}\theta_{td}$$

$$\varphi_{wt} = \underset{w\in W}{norm}\Big(n_{wt} + \varphi_{wt}\frac{\partial R}{\varphi_{wt}}\Big) \qquad n_{wt} = \sum_d n_{wd}p_{twd}$$

$$\varphi_{wvt} = \underset{w\in W}{norm}\Big(n_{vwt} + \varphi_{wvt}\frac{\partial R}{\partial\varphi_{wvt}}\Big) \qquad n_{wvt} = \sum_d n_{vwd}p_{twvd}$$

$$\theta_{td} = \underset{t\in T}{norm}\Big(n_{td} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\Big) \qquad n_{td} = \sum_{vw} n_{vwd}p_{twvd} + \sum_w n_{wd}p_{twd}$$

Доказательство. Запишем лагранжиан оптимизационной задачи с ограничениями типа равенства. Ограничения типа неравенства пока рассматривать не будем, а в конце покажем, что в точках оптимума они автоматически выполняются. Итак, лагранжиан

$$L = \sum_{d\in D}\sum_{w\in d} n_{wd}\ln\Big(\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) + \sum_{d\in D}\sum_{v,w\in d} n_{vwd}\ln\Big(\sum_{t\in T}\varphi_{wtv}\theta_{td}\Big) + R(\Phi,\Phi_v,\Theta) +$$
$$+ \sum_{v\in W,t\in T}\lambda_{vt}\Big(\sum_w\varphi_{wvt} - 1\Big) + \sum_{t\in T}\mu_t\Big(\sum_w\varphi_{wt} - 1\Big) + \sum_{d\in D}\gamma_d\Big(\sum_t\theta_{td} - 1\Big).$$

Посчитаем частные производные лагранжиана по оптимизируемым переменным

$$\frac{\partial L}{\partial\varphi_{wt}} = \sum_{d\in D} n_{wd}\frac{\theta_{td}}{\sum_t\varphi_{wt}\theta_{td}} + \mu_t + \frac{\partial R}{\partial\varphi_{wt}} = 0;$$

$$\frac{\partial L}{\partial\varphi_{wvt}} = \sum_{d\in D} n_{vwd}\frac{\theta_{td}}{\sum_t\varphi_{wvt}\theta_{td}} + \lambda_{vt} + \frac{\partial R}{\partial\varphi_{wvt}} = 0;$$

$$\frac{\partial L}{\partial\theta_{td}} = \sum_{v,w\in W} n_{vwd}\frac{\varphi_{wvt}}{\sum_t\varphi_{wvt}\theta_{td}} + \sum_{w\in W} n_{wd}\frac{\varphi_{wt}}{\sum_t\varphi_{wt}\theta_{td}} + \gamma_d + \frac{\partial R}{\partial\theta_{td}} = 0.$$

Умножим первое уравнение на φ_{wt}, просуммируем по $w\in W$ и учтем, что $\sum\limits_{w\in W}\varphi_{wt} = 1$. Получится, что

$$\mu_t = -\Big(\sum_{w\in W}\sum_{d\in D}\frac{n_{wd}\varphi_{wt}\theta_{td}}{\sum_t\varphi_{wt}\theta_{td}} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\Big),$$

откуда получим, что выполняется соотношение

$$\varphi_{wt} = \frac{\sum\limits_{d\in D} n_{wd}\dfrac{\varphi_{wt}\theta_{td}}{\sum_t\varphi_{wt}\theta_{td}} + \varphi_{wt}\dfrac{\partial R}{\partial\varphi_{wt}}}{\sum\limits_{w\in W}\sum\limits_{d\in D} n_{wd}\dfrac{\varphi_{wt}\theta_{td}}{\sum_t\varphi_{wt}\theta_{td}} + \varphi_{wt}\dfrac{\partial R}{\partial\varphi_{wt}}}$$

Обозначим $p_{twd} = \underset{t \in T}{norm}\, \varphi_{wt}\theta_{td}$, и $n_{wt} = \sum\limits_{d} n_{wd}p_{twd}$. Получим, что

$$\varphi_{wt} = \underset{w \in W}{norm}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\varphi_{wt}}\right)$$

Аналогично умножим второе уравнение на φ_{wvt}, просуммируем по $w \in W$ и исключим двойственные переменные λ_{vt} из уравнения. Введем вспомогательные переменные $p_{twvd} = \underset{t \in T}{norm}\, \varphi_{wvt}\theta_{td}$ и $n_{vwt} = \sum\limits_{d} n_{vwd}p_{twvd}$. Получится

$$\varphi_{wvt} = \underset{w \in W}{norm}\left(n_{vwt} + \varphi_{wvt}\frac{\partial R}{\partial \varphi_{wvt}}\right)$$

С третьим уравнением проделывается аналогичная процедура с умножением на θ_{td}, суммированием по $t \in T$ и по исключению двойственных переменных γ_d из уравнения. Получается, что

$$\theta_{td} = \underset{t \in T}{norm}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right).$$

Теорема доказана.

4.1 Алгоритм поиска локальных оптимумов

Полученные условия не позволяют явно выписать оптимальные значения, а лишь дают систему уравнений, соотношения между неизвестными параметрами построенной модели в точках локальных оптимумов. Можно применить метод простых итераций для пресчета новых значений φ и θ через предыдущие. Поэтому предлагается EM-алгоритм для нахождения приближения к оптимальному значению параметров.

Вход: Начальные приближения матриц параметров $\Phi, \Theta, \widetilde{\Phi}$, число итераций N;

Выход: Точка приближенного локального оптимума;

1: **для всех** $i = 0, \ldots, N$

2: **E-шаг.**

Посчитать вспомогательные переменые

$$p_{twd} = \underset{t \in T}{norm}\left(\varphi_{wt}\theta_{td}\right); \qquad n_{wt} = \sum_{d} n_{wd}p_{twd};$$

$$p_{tvwd} = \underset{t \in T}{norm}\left(\varphi_{wvt}\theta_{td}\right); \qquad n_{vwt} = \sum_{d} n_{vwd}p_{tvwd}.$$

3: **M-шаг.**

С помощью посчитанных на E-шаге вспомогательных переменных вычислить новые приближения значений параметров модели по формулам

$$\varphi_{wt} = \underset{w \in W}{norm}\Big(n_{wt} + \varphi_{wt}\frac{\partial R}{\varphi_{wt}}\Big);$$

$$\varphi_{wvt} = \underset{w \in W}{norm}\Big(n_{vwt} + \varphi_{wvt}\frac{\partial R}{\partial \varphi_{wvt}}\Big);$$

$$\theta_{td} = \underset{t \in T}{norm}\Big(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big).$$

4.2 Связь с моделью TNG

Покажем, как связаны модели TNG и Bigram-ARTM. Для простоты будем рассматривать вариант без регуляризации, для варианта с регуляризацией будут верны те же выкладки.

Оценим снизу целевую функцию в задаче TNG. Для этого воспользуемся свойством вогнутости логарифма.

$$\sum_{d \in D}\sum_{v,w \in d} n_{vwd} \ln\Big(x_{vw}\sum_{t \in T}\varphi_{wtv}\theta_{td} + (1 - x_{vw})\sum_{t \in T}\varphi_{wt}\theta_{td}\Big) \geq$$

$$\geq \sum_{d \in D}\sum_{v,w \in d} n_{vwd}x_{vw} \ln\Big(\sum_{t \in T}\varphi_{wtv}\theta_{td}\Big) + \sum_{d \in D}\sum_{v,w \in d} n_{vwd}(1 - x_{vw}) \ln\Big(\sum_{t \in T}\varphi_{wt}\theta_{td}\Big) \to \max_{\Phi,\Theta,\tilde{\Phi},x}$$

И положим все метки биграммности равными: $x_{vw} = \frac{\lambda}{1+\lambda}$

$$\frac{\lambda}{1+\lambda}\sum_{d \in D}\sum_{v,w \in d} n_{vwd} \ln\Big(\sum_{t \in T}\varphi_{wtv}\theta_{td}\Big) + \frac{1}{1+\lambda}\sum_{d \in D}\sum_{w \in d} n_{wd} \ln\Big(\sum_{t \in T}\varphi_{wt}\theta_{td}\Big) \to \max_{\Phi,\Theta,\tilde{\Phi}}$$

Отсюда видно, что модель Bigram-ARTM является оценкой снизу для модели TNG.

4.3 Биграммная модель как мультимодальная

Через некоторое время после доказательства теоремы 1 удалось установить, что эта теорема является частным случаем более общей теоремы, а построенная модель — частный случай мультимодальной TM с регуляризацией [16]. Альтернативный взгляд

на задачу всегда полезен, поэтому ниже приводится описание Bigram-ARTM как мультимодальной.

Будем рассматривать документ не только как набор слов, но еще и как набор *метаданных* или *токенов*. Примеры токенов, описывающих документ — слова словаря, авторы данного документа, время написания, ссылки на другие документы. Токены одного типа называют *модальностью* со словарем W^m, $m \in M$, M — множество различных модальностей, $\bigcup\limits_{m\in M} W^m = W$ — общий словарь для всех токенов. Через w будем обозначать токен с модальностью $m(w)$ Примем гипотезу условной независимости — распределения токенов в темах одно и то же для всех документов.

$$p(w|d) = \sum_{t\in T} p(w|t,d)p(t|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td}, \quad w \in W^m, \quad d \in D$$

Параметры модели $\Phi^m = (\varphi_{wt})_{W^m\times T}, \Theta = (\theta_{td})_{T\times D}$, записанные в столбик матрицы Φ^m образуют матрицу Φ.

Логарифм правдоподобия мультимодальной модели с регуляризацией

$$\sum_{m\in M}\tau_m \sum_{d\in D}\sum_{w\in W^m} n_{wd}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

$$\sum_{w\in W^m}\varphi_{wt} = 1, \varphi_{wt} \geq 0; \qquad \sum_{t\in T}\theta_{td} = 1, \theta_{td} \geq 0$$

Теорема 2, [16] Для непрерывно дифференцируемой $R(\Phi,\Theta)$ в точке локального экстремума описанной выше задачи выполнены соотношения

$$p_{tdw} = \underset{t\in T}{norm}\,\varphi_{wt}\theta_{td}$$

$$\varphi_{wt} = \underset{w\in W^m}{norm}\Big(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\Big); \qquad n_{wt} = \sum_{d\in D}\tau_m n_{dw}p_{tdw}$$

$$\theta_{td} = \underset{t\in T}{norm}\Big(n_{td} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\Big); \qquad n_{td} = \sum_{m\in M}\tau_m\sum_{w\in W^m} n_{dw}p_{tdw}$$

Чтобы увидеть, что предлагаемая модель оказалась частным случаем мультимодальной модели с регуляризацией, представим словарь W мультимодальной модели в виде

$$W = W^0 \sqcup W^{v_1} \sqcup \cdots \sqcup W^{v_W},$$

W^0 — униграммный словарь коллекции W^{v_i} — множество пар слов, идущих подряд в коллекции, начинающихся на слово $v_i \in W^0$.

5 Эксперимент

5.1 Данные

В качестве данных для эксперимента использовалась коллекция из $|D| = 1009$ текстов тезисов конференций ИОИ и ММРО за 2007 —2013 годы. Размер словаря составил $|W| = 13678$ слов и 327694 пар слов.

Заметим, что у модели Bigram-ARTM есть полезное свойство, имеющееся и в модели TNG, но отсутствующее в модели Bigram-TM. Свойство состоит в том, что модель восстанавливает одновременно вероятности $p(w|t)$ и $p(w|v,t)$, поэтому на выходе получается готовый инструмент ранжирования вероятностей пар слов в теме, поскольку $p(w,v|t) = p(w|v,t) \cdot p(v|t)$.

В работе [17] было показано, что оптимальное число тем для используемой коллекции документов ИОИ-ММРО лежит в диапазоне от 60 до 80, поэтому эксперименты проводились для $T = 70$ тем.

5.2 Сравнение интерпретируемости униграммной TM и Bigram-ARTM

Сравнение интерпретируемости проводилось для значения $\lambda = 0.01$ и для модели без регулизации. Результаты представлены в таблицах ниже. В каждой из таблиц приведены наиболее вероятные в темах биграммы, вероятность посчитана через параметры модели $p(w,v|t) = p(w|v,t) \cdot p(v|t) = \varphi_{vwt} \cdot \varphi_{vt}$.

Тема 1	Тема 1
СКЕЛЕТ	МНОГОУГОЛЬНЫЙ ФИГУРА
ТОЧКА	БАЗОВЫЙ СКЕЛЕТ
ФИГУРА	РЕБРО СКЕЛЕТ
РЕБРО	ПУСТОЙ КРУГ
ВЕРШИНА	КРУГ ЦЕНТР
ГРАНИЦА	СКЕЛЕТ ФИГУРА
КРУГ	ТЕРМИНАЛЬНЫЙ ВЕРШИНА
РИС	МАКСИМАЛЬНЫЙ ПУСТОЙ
ЦЕНТР	СКЕЛЕТНЫЙ ГРАФ
МНОГОУГОЛЬНЫЙ	ТЕРМИНАЛЬНЫЙ РЕБРО
БАЗОВЫЙ	МНОЖЕСТВО ТОЧКА
МНОЖЕСТВО	СКЕЛЕТ МНОГОУГОЛЬНЫЙ
ФОРМА	ТОЧКА СКЕЛЕТ
СКЕЛЕТНЫЙ	ФРАГМЕНТ ГРАНИЦА
МОДЕЛЬ	БАЗОВЫЙ КРУГ
РАДИУС	ВЕТВЬ СКЕЛЕТ
ЦЕПОЧКА	ЦЕНТР ТОЧКА
МАКСИМАЛЬНЫЙ	ФИГУРА СКЕЛЕТ
СРАВНЕНИЕ	ТОЧНОСТЬ АППРОКСИМАЦИЯ
ПРОЕКЦИЯ	СЕРЕДИННЫЙ ОСЬ

Тема 2	Тема 2
ЗАДАЧА	ЗАДАЧА РАСПОЗНАВАНИЕ
МЕТРИКА	АЛГЕБРАИЧЕСКИЙ ЗАМЫКАНИЕ
ОБЪЕКТ	МОДЕЛЬ АВО
АЛГЕБРАИЧЕСКИЙ	ВЫБОР МЕТРИКА
МНОЖЕСТВО	АЛГЕБРАИЧЕСКИЙ ПОДХОД
МАТРИЦА	ЛИНЕЙНЫЙ ЗАМЫКАНИЕ
МОДЕЛЬ	КРИТЕРИЙ КОРРЕКТНОСТЬ
АЛГОРИТМ	РЕШЕНИЕ ЗАДАЧА
РАСПОЗНАВАНИЕ	СИСТЕМА ТОЧКА
ЗАМЫКАНИЕ	КОРРЕКТНОСТЬ АЛГЕБРАИЧЕСКИЙ
ОЦЕНКА	РАСПОЗНАВАТЬ ОПЕРАТОР
АВО	МАТРИЦА ОЦЕНКА
КОРРЕКТНЫЙ	ПОПАРНЫЙ РАССТОЯНИЕ
УСЛОВИЕ	УСЛОВИЕ РЕГУЛЯРНОСТЬ
КОРРЕКТНОСТЬ	ВЫЧИСЛЕНИЕ ОЦЕНКА
РАССТОЯНИЕ	МАТРИЦА ПОПАРНЫЙ
ОПЕРАТОР	ЗАМЫКАНИЕ МОДЕЛЬ
СИСТЕМА	КОНТРОЛЬНЫЙ ОБЪЕКТ
ПРИЗНАК	ЗАМЫКАНИЕ АВО
ЛИНЕЙНЫЙ	АЛГОРИТМ ВЫЧИСЛЕНИЕ

Тема 3	Тема 3
ОЦЕНКА	ВЕРОЯТНОСТЬ ПЕРЕОБУЧЕНИЕ
АЛГОРИТМ	ОЦЕНКА ВЕРОЯТНОСТЬ
ВЫБОРКА	ЭМПИРИЧЕСКИЙ РИСКА
ВЕРОЯТНОСТЬ	МИНИМИЗАЦИЯ ЭМПИРИЧЕСКИЙ
МЕТОД	МЕТОД ОБУЧЕНИЕ
ПЕРЕОБУЧЕНИЕ	ОБУЧАТЬ ВЫБОРКА
ОШИБКА	ТОЧНЫЙ ОЦЕНКА
МОНОТОННЫЙ	ВЕРХНИЙ ОЦЕНКА
СЕМЕЙСТВО	ГЕНЕРАЛЬНЫЙ ВЫБОРКА
ОБЪЕКТ	ЧАСТОТА ОШИБКА
МНОЖЕСТВО	СЕМЕЙСТВО АЛГОРИТМ
ЭМПИРИЧЕСКИЙ	ЧИСЛО ОШИБКА
ОБУЧЕНИЕ	ВЕКТОР ОШИБКА
РИСКА	ОБОБЩАТЬ СПОСОБНОСТЬ
ЧИСЛО	ЭМПИРИЧЕСКИЙ РИСК
ТОЧНЫЙ	КОМБИНАТОРНЫЙ ОЦЕНКА
ОБУЧАТЬ	ОЦЕНКА CCV
КОМБИНАТОРНЫЙ	МНОЖЕСТВО АЛГОРИТМ
СЛУЧАЙ	ОШИБКА АЛГОРИТМ
КЛАССИФИКАЦИЯ	МЕТОД МИНИМИЗАЦИЯ

Видно, что полученные темы описывают тематику одних и тех же научных школ. Также видно, что темы, представленные в списка топа биграмм, являются более интерпретируемыми и полнее отражают содержание темы.

5.3 Сглаживающий регуляризатор для выделения фоновых тем

Сглаживающий регуляризатор уменьшает расстояние Кульбака-Лейблера от заданных столбцов матриц $\Phi, \overline{\Phi}$ и заданных строк матрицы Θ до равномерного распределения

$$R(\Phi, \overline{\Phi}, \Theta) = \sum_{t\in\{T-1,T\}} \sum_{w\in W} \alpha \cdot \ln(\varphi_{wt}) + \sum_{(v,w)\in W} \sum_{t\in\{T-1,T\}} \alpha \cdot \ln(\varphi_{vwt}) + \sum_{d\in D} \sum_{t\in\{T-1,T\}} \alpha \ln(\theta_{td}).$$

Используется на 2 последний темах (из 70 тем), для того, чтобы сделать эти две темы фоновыми, то есть содержащими слова общей лексики данной коллекции. Результаты работы представлены в таблицах ниже.

Сглаживающий регулизатор, $\alpha = 3.0$

Тема 69	Тема 69
ВЛАДИМИР	РАБОТА ПОДДЕРЖАТЬ
АЛЕКСАНДР	ГРАНТ РФФИ
ЮРИЙ	ПОДДЕРЖАТЬ ГРАНТ
СЕРГЕЙ	ВЛАДИМИР ВЛАДИМИР
МИХАИЛ	КРАСОТКИН ОЛЬГА
НИКОЛАЙ	ОЛЬГА ВЯЧЕСЛАВ
ВЯЧЕСЛАВ	НАГОРНЫЙ ЮРИЙ
ПОДДЕРЖАТЬ	ОЛЕГ СЕРГЕЙ
АЛЕКСЕЙ	ВАДИМ ВЯЧЕСЛАВ
ГРАНТ	СЕРГЕЙ ВЛАДИМИР
РФФИ	ВИШНЯК БОРИС
КОНСТАНТИН	СЕРЕДИНА ОЛЕГ
ОЛЕГ	МИХАИЛ СЕРГЕЙ
ИВАН	РАСТЯЖИМОСТЬ СОПРОТИВЛЕНИЕ
ИГОРЬ	ФИЗИК РАДИОЭЛЕКТРОНИКА
ДМИТРИЙ	РАЗИН НИКОЛАЙ
БОРИС	СУЛИМОВ ВАЛЕНТИН
ОЛЬГА	ОБЩИЙ ПРЕДОК
ВАЛЕРИЙ	ИВАН КОНСТАНТИН

Сглаживающий регулизатор, $\alpha = 30.0$

Тема 70	Тема 70
РАБОТА	ПОДДЕРЖКА РФФИ
МЕТОД	ГРАНТ РФФИ
ОСНОВА	РФФИ ПРОЕКТ
ПОЗВОЛЯТЬ	РАБОТА ПОДДЕРЖАТЬ
АНАЛИЗ	ПОДДЕРЖАТЬ ГРАНТ
ЗАДАЧА	ВЫПОЛНИТЬ ПОДДЕРЖКА
ИССЛЕДОВАНИЕ	ТОЧКА ЗРЕНИЕ
ПОДХОД	МАТЕМАТИЧЕСКИЙ ОЖИДАНИЕ
РЕЗУЛЬТАТ	РАБОТА ВЫПОЛНИТЬ
ИСПОЛЬЗОВАНИЕ	РАБОТА ПОСВЯТИТЬ
ВЫПОЛНИТЬ	СКАЛЯРНЫЙ ПРОИЗВЕДЕНИЕ
СЛЕДОВАТЬ	PATTERN RECOGNITION
ПОКАЗАТЬ	ОБОБЩАТЬ СПОСОБНОСТЬ
ВИД	ДРУГ ДРУГ
РЕШЕНИЕ	DATUM MINE
НЕОБХОДИМЫЙ	ДИНАМИЧЕСКИЙ ПРОГРАММИРОВАНИЕ
РАЗЛИЧНЫЙ	КОСТНЫЙ ТКАНЬ
ОСНОВАТЬ	ЛИНЕЙНЫЙ КОМБИНАЦИЯ
ОБРАЗ	РЕШЕНИЕ ПОСТАВИТЬ

Видно, что в теме 69, когда $\alpha = 3.0$ в качестве слов общей лексики выделяются в основном имена и фамилии, а в теме 70, когда $\alpha = 30.0$, выделяются слова общей лексики, характерные для данной коллекции.

6 Заключение

В работе предложена вычислительно эффективная мультиграммная тематическая модель Bigram-ARTM, использующая аддитивную регуляризацию. Построена общая теоретическая модель, описывающая эти существующие мультиграммные тематические модели. Показана связь модели Bigram-ARTM с моделью Topical N-grams (TNG) - наиболее общей из существующих мультиграммных моделей. Показано, что предлагаемая модель является частным случаем мультимодальной модели с регуляризацией. Работа Bigram-ARTM проиллюстрирована на коллекции текстов тезисов конференций ИОИ и ММРО. Показано, что использование биграмм улучшает интерпретируемость выделяемых тем, и что использование сглаживающего регуляризатора позволяет выделять в отдельные темы слова общей лексики, характерные для данной коллекции.

Список литературы

[1] Blei David M, Ng Andrew Y, Jordan Michael I. Latent dirichlet allocation // the Journal of machine Learning research. — 2003. — Vol. 3. — P. 993–1022.

[2] Hofmann Thomas. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval / ACM. — 1999. — P. 50–57.

[3] Lee Daniel D, Seung H Sebastian. Algorithms for non-negative matrix factorization // Advances in neural information processing systems. — 2001. — P. 556–562.

[4] Algorithms and applications for approximate nonnegative matrix factorization / Michael W Berry, Murray Browne, Amy N Langville et al. // Computational statistics & data analysis. — 2007. — Vol. 52, no. 1. — P. 155–173.

[5] Wallach Hanna M. Topic modeling: beyond bag-of-words // Proceedings of the 23rd international conference on Machine learning / ACM. — 2006. — P. 977–984.

[6] Wallach Hanna M, Mimno David M, McCallum Andrew. Rethinking LDA: Why priors matter // Advances in neural information processing systems. — 2009. — P. 1973–1981.

[7] Newman David, Bonilla Edwin V, Buntine Wray. Improving topic coherence with regularized topic models // Advances in neural information processing systems. — 2011. — P. 496–504.

[8] Lindsey Robert V, Headden III William P, Stipicevic Michael J. A phrase-discovering topic model using hierarchical pitman-yor processes // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning / Association for Computational Linguistics. — 2012. — P. 214–222.

[9] Jameel Shoaib, Lam Wai. An N-Gram Topic Model for Time-Stamped Documents. // ECIR. — 2013. — P. 292–304.

[10] Vorontsov Konstantin, Potapenko Anna. Additive regularization of topic models // Machine Learning. — 2014. — P. 1–21.

[11] Тихонов Арсенин. Методы решения некорректных задач. — 1979.

[12] Vorontsov Konstantin, Potapenko Anna. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // Analysis of Images, Social Networks and Texts. — Springer, 2014. — P. 29–46.

[13] Griffiths Thomas L, Steyvers Mark, Tenenbaum Joshua B. Topics in semantic representation. // Psychological review. — 2007. — Vol. 114, no. 2. — P. 211.

[14] Wang Xuerui, McCallum Andrew, Wei Xing. Topical n-grams: Phrase and topic discovery, with an application to information retrieval // Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on / IEEE. — 2007. — P. 697–702.

[15] Jurafsky Dan, Martin James H. Speech & language processing. — Pearson Education India, 2000.

[16] Konstantin Vorontsov Oleksandr Frei Murat Apishev Peter Romov, Dudarenko Marina. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections // Proceedings of the 4th international conference on analisys of images, social networks, and texts. — 2015.

[17] Vorontsov K. Potapenko A. Plavin A. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization. — 2015.