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

На правах рукописи

Потапенко Анна Александровна

# Семантические векторные представления текста на основе вероятностного тематического моделирования

05.13.17 – Теоретические основы информатики

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
д. ф.-м. н., проф. РАН
Воронцов Константин Вячеславович

Москва – 2018

# Оглавление

# Введение

**Актуальность темы исследования.** В задачах анализа текста (Natural Language Processing, NLP) часто возникает необходимость представления слов или сегментов текста векторами низкой размерности, отражающими их семантику. Если два близких по смыслу слова удается представить близкими векторами, то такие представления затем могут эффективно использоваться для широкого класса задач NLP, в частности, для задач информационного поиска, классификации, категоризации и суммаризции текстов, анализа тональности, определения границ именованных сущностей, разрешения омонимии, генерации ответов в диалоговых системах.

Подходы векторного представления слов активно развиваются в последние годы [1–4]. Постоянно расширяется спектр их приложений, и улучшается качество предсказания семантической близости слов. Однако признаковые описания слов в большинстве случаев представляют собой «черный ящик»: координаты вектора не удается интерпретировать как определенные аспекты смысла. Это затрудняет применение данных моделей в системах разведочного информационного поиска и других приложениях, где важна не только оценка близости, но и ее объяснение для пользователя.

В большинстве методов строятся плотные вектора низкой размерности, таким образом, что каждое слово представляется набором фиксированного числа признаков. Это противоречит гипотезе об экономном хранении, согласно которой человеческий мозг представляет более специфичные концепты большим числом характеристик, а более общие – меньшим [5, 6]. Проводя параллели с когнитивными науками, векторные представления должны быть сильно разреженными, а их компоненты должны соответствовать отдельным семантическим признакам кодируемого понятия.

В данной работе исследуется применимость вероятностного тематического моделирования для получения таких представлений. Тематическая модель поз-

воляет представить слова и документы вероятностными распределениями на множестве тем. При этом ставятся вопросы об интерпретируемости и различности тем, разреженности полученных распределений, устойчивости модели к шуму в данных и случайности начальных приближений. Эти вопросы являются открытыми в области тематического моделирования и представляют отдельный интерес.

**Степень разработанности темы исследования.** Дистрибутивная гипотеза, утверждающая что смысл слова можно определить по его контекстам, была предложена в 1950-х годах [7, 8]. Модели векторного представления слов, основанные на частотных распределениях слов в контекстах, развиваются на протяжении последних десятилетий и хорошо изучены. Одними из первых работ можно считать модели 1990-х годов латентного семантического анализа (Latent Semantic Analysis, LSA) [9] и семантической памяти (Hyperspace Analogue to Language, HAL) [10]. Эти модели позволяют представлять слова векторами в некотором низкоразмерном пространстве, так что семантически близкие слова имеют близкие вектора [11]. Для оценивания моделей существуют составленные вручную наборы пар слов с экспертными оценками близости.

Недавно большую популярность получили модели *обучаемых* векторных представлений слов, в частности, семейство моделей word2vec [1], предложенное Томасом Миколовым в 2013 году. Эта архитектура возникла как результат упрощения глубоких нейросетевых моделей языка. Она содержит один скрытый слой, не содержит нелинейных преобразований и может интерпретироваться как матричное разложение PMI-частот слов в контекстах [12]. Недавно предложенная модель GloVe [2] также решает задачу матричного разложения, но с другим оптимизационным критерием. Таким образом, модели обучаемых векторных представлений слов (word embeddings) можно считать, скорее, новым витком развития хорошо изученных подходов, нежели революционно новыми технологиями в данной области.

Обе группы методов обладают рядом недостатков, среди которых можно

назвать отсутствие интерпретируемости компонент построенных векторов.

Вероятностное тематическое моделирование развивалось параллельно, начиная с модели вероятностного латентного семантического анализа (Probabilistic Latent Semantic Analysis, PLSA), которая была предложена Томасом Хофманном в 1999 году [13]. Эта модель позволяет осуществлять мягкую би-кластеризацию слов и документов по темам. Каждая тема при этом описывается вероятностным распределением на множестве слов. Как правило, темы являются хорошо интерпретируемыми, т.е. эксперт можно понять, о чем данная тема, посмотрев на список наиболее вероятных слов.

Наиболее известной тематической моделью является латентное размещение Дирихле (Latent Dirichlet Allocation, LDA), в которой дополнительно предполагается, что параметры модели имеют априорное распределение Дирихле [14]. Эта модель позиционируется авторами как способ получать разреженные тематические распределения, однако на практике достигаемой разреженности часто оказывается недостаточно. На больших корпусах текстов модели PLSA и LDA показывают сопоставимое качество [15–17]. Позднее были построены сотни расширений LDA, и предложены алгоритмы их обучения в рамках байесовского подхода [18, 19]. Важной проблемой этой линии исследований остается сложность вывода алгоритмов обучения для новых моделей, а также сложность комбинирования моделей и дополнительных требований, таких как иерархии тем, учет мета-данных, отказ от гипотезы мешка слов.

Альтернативный подход аддитивной регуляризации тематических моделей (АРТМ) предлагается в работе [20] и развивается в данном диссертационном исследовании. АРТМ позволяет строить тематические модели, оптимизирующие заданный набор критериев. В частности, ставится вопрос о возможности повышения различности и разреженности тем без существенного ухудшения основного критерия правдоподобия.

Применимость подхода вероятностного тематического моделирования к задаче определения семантической близости слов является мало изученной. Как

правило, в статьях исследуется модель LDA, которая показывает на этой задаче низкое качество. В данном исследовании устанавливаются взаимосвязи между тематическими моделями и моделями дистрибутивной семантики. Разрабатываемый подход аддитивной регуляризации расширяется для решения задач семантической близости слов и для обработки мультимодальных данных.

**Цели и задачи диссертационной работы.** Цель диссертационного исследования – разработка методов построения интерпретируемых разреженных векторных представлений текста, применимых в задачах определения семантической близости.

Для достижения данной цели в диссертации решаются следующие задачи.

1. Обобщение известных алгоритмов тематического моделирования. Построение разреженных тематических векторных представлений.

2. Повышение различности и интерпретируемости тем с помощью регуляризации в рамках подхода APTM. Разработка методики оценивания различности и интерпретируемости.

3. Построение интерпретируемых разреженных тематических представлений слов и сегментов текста на основе моделирования со-встречаемости слов в локальных контекстах.

4. Построение единого векторного пространства для токенов различных *модальностей* (авторы, даты и другие мета-данные документов).

**Научная новизна.** Объединяются преимущества вероятностного тематического моделирования и моделей векторного представления слов на основе их совместной встречаемости. Это позволяет построить векторное пространство с интерпретируемыми размерностями, с помощью которого успешно решается задача определения семантической близости слов или сегментов текста. Разрабатывается подход аддитивной регуляризации тематических моделей, позволя-

ющий встраивать новые требования, мотивированные лингвистическими предположениями или специфичными свойствами конечных приложений.

**Теоретическая и практическая значимость.** Предлагается аддитивно регуляризованная тематическая модель, позволяющая достичь высокой разреженности, различности и интерпретируемости предметных тем. Данные свойства тематических моделей важны в задачах разведочного поиска, навигации по коллекциям научных статей, категоризации и суммаризации документов.

Предлагается формализация дистрибутивной гипотезы в рамках подхода АРТМ. В обучении моделей используется информация о совместной встречаемости слов. Это позволяет уйти от гипотезы о представлении документа в виде «мешка слов», являющейся одним из самых критикуемых допущений в тематическом моделировании. Предлагается алгоритм построения единого векторного пространства для слов, сегментов текста и мета-данных документа, в котором сохраняется свойство интерпретируемости компонент.

Примером применения интерпретируемых семантических векторных представлений слов является задача автоматического пополнения ключевых слов в заданных категориях при построении системы показов рекламы. Расширение на данные других модальностей применимо в рекомендательных системах, анализе социальных сетей, анализе транзакционных данных и других приложениях.

**Методология и методы исследования.** В работе использованы методы теории вероятностей, оптимизации, теории машинного обучения и компьютерной лингвистики. Экспериментальное исследование проводится на языках C++ и Python с использованием библиотек NLTK, Gensim, BigARTM и удовлетворяет принципам воспроизводимости результатов.

**Положения, выносимые на защиту:**

- Предложен обобщенный EM-алгоритм, позволяющий комбинировать известные тематические модели, обеспечивая контроль перплексии, робастности и разреженности.

- В рамках подхода аддитивной регуляризации предложена тематическая модель фоновых и предметных тем, обладающих свойствами различности, интерпретируемости и высокой разреженности.

- Предложен алгоритм построения тематических векторных представлений, сохраняющих информацию о семантической близости слов и обладающих интерпретируемыми компонентами.

- С помощью подхода аддитивной регуляризации тематических моделей алгоритм построения векторных представлений слов обобщен на случай мультимодальных данных и сегментированного текста.

**Степень достоверности и апробация результатов.** Достоверность результатов обеспечивается математическими доказательствами теорем и серией подробно описанных вычислительных экспериментов на реальных текстовых коллекциях. Основные результаты диссертации докладывались на следующих конференциях и семинарах:

1. BlackboxNLP: Analyzing and interpreting neural networks for NLP (co-located with EMNLP), октябрь 2018, Брюссель (постер).

2. 7th International Conference - Analysis of Images, Social networks and Texts (AIST), Москва, июль 2018.

3. Доклад в группе Томаса Хофманна, ETH Zurich, ноябрь 2017.

4. Artificial Intelligence and Natural Language (AINL), Санкт-Петергбург, сентябрь 2017.

5. 2nd Workshop on Representation Learning for NLP (co-located with ACL), август 2017, Ванкувер (постер).

6. Доклад в группе Криса Биманна по языковым технологиям, Технический Университет Дармштадта, июль 2016.

7. Доклад на семинаре по анализу текстов в Google, Цюрих, июнь 2016.
8. Yandex School Conference “Machine Learning: Prospects and Applications”, октябрь 2015, Берлин (постер).
9. Доклад на семинаре в Microsoft Research Cambridge, апрель 2015.
10. The Third International Symposium On Learning And Data Sciences (SLDS), Лондон, апрель 2015.
11. Школа Russian Summer School on Information Retrieval, август 2014, Нижний Новгород (постер).
12. The 35-th European Conference on Information Retrieval (ECIR), Москва, март 2013 (постер).
13. Международная конференция по компьютерной лингвистике “Диалог”, Москва, июнь 2014.
14. XXI Международная научная конференция студентов, аспирантов и молодых ученых “Ломоносов-2014”, Москва, 2014.
15. 16-ая Всероссийская конференция “Математические методы распознавания образов” (ММРО), Казань, 2013.

**Публикации.** Материалы диссертации опубликованы в 12 печатных работах, из них 6 статей индексируются в базе Scopus [21–26], еще одна [27] опубликована в журнале, входящем в перечень ВАК. Работа [28] опубликована в рецензируемом научном журнале, работа [29] представлена на воркшопе международной конференции EMNLP, работы [30–32] являются тезисами докладов. Еще одна статья [33] принята в печать (Scopus).

**Личный вклад автора.** Подход аддитивной регуляризации тематических моделей разрабатывался в соавторстве с Воронцовым К.В. [21, 23–25]. Основные положения, выносимые на защиту, являются личным вкладом автора

в опубликованные работы. Результаты по комбинированию тематического моделирования с моделями дистрибутивной семантики, представленные в пятой главе, получены автором лично, за исключением некоторых экспериментов, проведенных совместно с Поповым А.С. [26].

**Структура и объем диссертации.** Диссертация состоит из введения, двух обзорных глав, трех глав с результатами проведенного исследования, заключения и библиографии. Общий объем диссертации 147 страниц, из них 131 страница текста, включая 15 рисунков и 12 таблиц. Библиография включает 143 наименования на 16 страницах.

**Краткое содержание по главам.** В главе 1 приводятся основные принципы дистрибутивной семантики и типы семантической близости слов. Подробно рассматриваются математические модели, позволяющие произвести переход от корпусных частот к низкоразмерным семантическим векторным представлениям текста. При систематичном анализе становится ясно, что методы, пришедшие из различных областей (языковое моделирование, тематическое моделирование, матричные разложения, глубокие нейронные сети) обладают очень схожей структурой. Это понимание позволяет построить гибридные подходы, описанные в главе 5.

Глава 2 содержит обзор классических тематических моделей и алгоритмов их обучения. Особенно подробно рассматриваются байесовские методы, широко распространенные в литературе по тематическому моделированию. В частности, описывается три способа обучения тематической модели LDA. Далее в работе обсуждаются сложности байесовского подхода и развивается альтернативный подход – аддитивная регуляризация тематических моделей.

В главе 3 рассматривается ряд эвристик, применимых к базовой тематической модели PLSA. Исследуются различные модификации EM-алгоритма, используемого для ее обучения. В результате удается построить набор моделей, комбинирующих свойства разреженности тематических распределений, робастности к шуму в данных и экономного сэмплирования.

В главе 4 эти результаты обобщаются в рамках подхода аддитивной регуляризации. Рассматривается проблема неединственности стохастического матричного разложения, и вводятся дополнительные критерии оптимизации. Разрабатывается модель предметных и фоновых тем, позволяющая разделить специфичные термины от фоновой лексики и тем самым повысить интерпретируемость, различность и разреженность тем.

В главе 5 предлагается алгоритм построения семантических представлений текста на основе разработанного аппарата аддитивной регуляризации тематических моделей. В результате удается построить тематические представления слов, сопоставимые по качеству на задачах определения семантической близости слов со стандартными нейросетевыми моделями семейства word2vec. При этом в экспериментах демонстрируются преимущества предлагаемого подхода: интерпретируемость компонент векторов, высокая разреженность, возможность дополнительной регуляризации. Помимо слов в единое векторное пространство погружаются сущности дополнительных модальностей: метки времени, категории и другие метаданные, связанные с документами. На задаче предсказания семантической близости документов предложенный метод превосходит модель doc2vec — стандартное расширение модели wod2vec для документов.

# Глава 1

# Дистрибутивная семантика

*Дистрибутивная семантика (distributional semantics)* изучает способы определения семантической близости слов на основе их распределения в большом корпусе текстов. В основе лежит *дистрибутивная гипотеза*, утверждающая, что слова со схожим распределением контекстов имеют схожие смыслы: “You shall know the word by the company it keeps.” [7, 8]. Под контекстом могут пониматься соседи слова в окне фиксированной ширины или более сложные синтаксические конструкции [34].

В данной главе рассматриваются различные типы семантической близости слов. Приводится общая схема обработки текста для получения оценок близости, и подробно рассматривается ее ключевой компонент – математические модели построения низкоразмерных векторов слов. Все модели излагаются в едином формализме без разделения на частотные (count-based) и предсказывающие (predictive), принятого во многих обзорах. В результате удается выделить общие принципы и придти к гибридным моделям (глава 5).

## 1.1. Типы семантической близости слов

В компьютерной лингвистике два слова называются *семантически близкими (semantically similar)*, если они имеют общий *гипероним* (родительскую категорию, дословно с греческого - “сверх-имя”). Например, машина и велосипед близки, потому что оба являются транспортным средством [35]. Такой тип отношения между словами иногда также называют *таксономической близостью* [11]. Семантическая близость является частным случаем *семантической связанности (semantic relatedness)* слов [36]. Семантически связанными называют слова, находящиеся в отношении меронимии (отношение часть-целое: колесо и машина), гипонимии (родо-видовое отношение: машина и транспорт), синони-

мии (кружка и чашка), антонимии (горячий и холодный). Также сюда могут включаться слова, которые связаны синтаксическими конструкциями или просто часто встречаются совместно (пчела и мед).

В приложениях важно уметь отличать *семантическую близость* слов от других типов семантической связанности. При этом точное определение семантической близости может варьироваться в зависимости от постановки прикладной задачи. Например, для автоматизации колл-центра в банке важно исключить антонимы (открыть-закрыть вклад) из понятия близких слов. В то же время, для системы автоматического пополнения ключевых слов по категории «действия со вкладом», данные слова могут считаться близкими. При этом в обоих случаях семантически связанные слова «открыть» и «вклад» близкими считаться не должны.

Для определения типа близости слов полезно заметить, что существует два принципиально различных типа совместной встречаемости слов в корпусе [37]. Если два слова часто встречаются в тексте рядом, будем называть их *синтагматически ассоциированными (syntagmatic associates)*. Пример: «открыть» и «вклад». Если два слова взаимозаменяемы в одних и тех же контекстах, будем называть их *парадигматически параллельными (paradigmatic parallels)*. Пример: «открыть» и «закрыть» (оба слова встречаются в контексте слова «вклад»). Также говорят, что синтагматически ассоциированные слова имеют высокую *совстречаемость первого порядка* (или просто совстречаемость), а парадигматически паралаллельные слова – высокую *совстречаемость второго порядка* (близость векторов, составленных из совстречаемостей первого порядка со всеми словами словаря). Как правило, нас будет интересовать совстречаемость второго порядка, т.к. именно она позволяет выделять семантически близкие слова.

При этом стоит отметить, что разные авторы не придерживаются единой терминологии. Например, в известной выборке пар слов с экспертными оценками близости WordSim353 [38] семантическая близость слов противопоставляет-


Рис. 1.1. Схема терминов о семантической связанности слов.

ся семантической связанности, а не является ее частным случаем. На рис. (1.1) представлена диаграмма, соответствующая такому подходу. При этом разделение типов отношений на семантически близкие и связанные также варьируется. Так, в выборке WordSim353 антонимы считаются семантически близкими, что противоречит доминирующему в литературе подходу.

В когнитивных науках семантическую связанность слов принято называть *атрибутивной близостью (attributional similarity)*. Помимо нее, изучается так называемся *близость отношений (relational similarity)*. Она подробно рассматривается в статье [39] 1998 года. В этом понятии участвуют уже не отдельные слова, а отношения слов, например, высокую близость будут иметь пары «кот:мяукать» и «собака:лаять». Такой тип близости в 2013 году был сильно популяризирован статьями Миколова [1, 40], где предлагалось решать *задачу аналогий*. В этой задаче нужно угадать четвертое слово по трем данным, например: (мужчина:женщина, король:?; Россия:Москва, Франция:?). Была разработана программа word2vec, которая успешно предсказывала четвертое слово на подготовленном наборе аналогий. Также, метод хорошо решал *задачу близости* слов. В ней оценивалась корреляция между экспертными оценками атрибутивной близости слов и предсказаниями модели.

Интересной представляется дискуссия о том, возможно ли сведение близости отношений к атрибутивной близости слов. Например, можно наивно предположить, что пара Россия:Москва и Франция:Париж имеет высокую близость отношений, т.к. Париж и Франция, Париж и Москва имеют высокую атрибу-

тивную близость, в то время как Париж и Россия – низкую. Согласно экспериментам [41, 42], близость отношений не сводится к атрибутивной близости слов. Это также соответствует пониманию из когнитивных наук о том, что задача аналогий является на порядки более сложной, чем задача близости, даже для людей. Тем не менее, в статье [43] теоретически показано, что в семействе моделей word2vec [1] такое сведение происходит.

## 1.2. Этапы обработки: от корпуса к смыслам

*Модели векторного представления слов (Vector Space Models of Semantics, VSM)* используют частоты в корпусе, чтобы представить каждое слово некоторым вектором, отражающим его смысл [44]. Ожидается, что слова, представленные близкими векторами, будут близки по смыслу. Такие модели изучаются на протяжении последних нескольких десятилетий и подробно описаны в обзоре [11]. Можно выделить несколько ключевых этапов при их построении.

**1. Лингвистическая предобработка.** На первичном этапе анализа текста, как правило, производится токенизация, нормализация, аннотирование [45]. *Токенизация* включает в себя разбиение текста на токены, корректную обработку пунктуации. *Нормализация* заключается в приведении текста к нижнему регистру, а также лемматизации или стемминге. При *лемматизации* каждое слово приводится к нормальной форме, в то время как при *стемминге* слово усекается до его основы. Из определений ясно, что лемматизация является более сложным процессом, поэтому часто приводит к лучшему качеству, но и большим временным затратам. *Аннотирование* – это необязательный этап, которой может включать в себя присвоение каждому слову аннотаций, таких как часть речи, грамматическая роль в предложении, тип именованной сущности, и т.д.

**2. Построение частотной матрицы.** На данном этапе строится матрица, строки которой соответствуют словам в словаре, а столбцы — контекстам. Элементами матрицы являются счетчики $n_{uv}$, которые показывают, сколько раз определенное слово $u$ встретилось в определенном контексте $v$. Понятие контекста может быть определено несколькими способами. В простейшем случае *контекстами* будем считать все слова, расположенные в тексте не далее, чем на $h$ позиций от заданного, т.е. в окне фиксированного радиуса $h$. Таким образом, рассматриваемая частотная матрица будет *квадратной симметричной матрицей* счетчиков совместной встречаемости слов. Это наиболее распространенный случай, рассматриваемый в литературе.

В более сложном случае в определении контекста может участвовать синтаксическая структура предложения, например, контекстами можно считать:

- глаголы в конструкциях типа субъект-глагол и глагол-объект [46];
- одно существительное влево и одно существительное вправо для главного существительного в предложении [47];
- все прилагательные, зависящие от данного существительного [48].

Больше деталей о синтаксических контекстах можно найти в работе [49]. Однако работа [50] показывает, что в случае достаточно большого корпуса методы без учета синтаксиса способны достичь сопоставимого качества.

В задачах, требующих векторного описания документов, на данном этапе может строиться матрица частот слов в документах.

**3. Частотное взвешивание.** Важной проблемой частотной матрицы, построенной на предыдущем этапе, является несбалансированность между редкими и частыми словами. Например, строка, соответствующая союзу «и» будет содержать счетчики на порядки большие, чем строка, соответствующая редкому термину «симметрия». Поэтому простые счетчики совместной встречаемости некоторым образом штрафуют с учетом отдельной встречаемости слов.

Для матриц слова-документы обычно используют TF-IDF *(Term Frequency - Inverted Document Frequency)*, где в качестве штрафа выступает логарифм документной частоты слова (числа документов, в которых оно встречается хотя бы раз). Для матриц слова-слова часто подсчитывают *поточечную взаимную информацию (Pointwise Mutual Information)*:

$$\mathrm{PMI}(u, v) = \log \frac{p(u, v)}{p(u)p(v)},$$

где $p(u, v)$ – эмпирическая вероятность встретить два слова в окне фиксированной ширины, а $p(u)$ и $p(v)$ – эмпирические вероятности встретить $u$ и $v$ в корпусе.

PMI успешно штрафует слишком частотные слова, однако имеет ряд недостатков. Во-первых, этот подход выводит в топ слишком редкие слова, во-вторых, значения не определены для слов, которые ни разу не встретились вместе. В работе [51] предлагается решить обе проблемы введением *положительной поточечной взаимной информации (positive Pointwise Mutual Information)*:

$$\mathrm{pPMI}(u, v) = \max(0, \mathrm{PMI}(u, v))$$

Эта простая эвристика зануления отрицательных значений хорошо работает на практике.

**4. Понижение размерности.** В матрице, составленной на предыдущем этапе, каждое слово представлено длинным разреженным вектором некоторых счетчиков. Аналогично представлены контексты или документы. Такое представление содержит шум, кроме того, работа с длинными разреженными векторами (например, их сравнение) может быть неэффективна. Поэтому важным этапом является переход к плотным векторам в пространстве меньшей размерности.

Одним из наиболее простых и классических методов является сингулярное разложение (SVD), при этом выбираются строки и столбцы, соответствующие первым $k$ сингулярным значениям. В результате строится аппроксимация

исходной матрицы ранга $k$, которая является наилучшей в смысле квадратичной нормы. Применение такого разложения к TF-IDF матрице слова-документы приводит к методу латентного семантического анализа (LSA). Он широко используется в анализе текстов для описания документов в низкоразмерном признаковом пространстве. Существует также огромное число альтернативных методов, применяемых на данном этапе.

**5. Использование построенных векторов.** В результате предыдущих шагов мы смогли представить некоторые единицы языка (слова, контексты, документы) векторами низкой размерности. При этом предполагается, что эти вектора сохраняют семантику. Например, близкие по смыслу слова имеют близкие представления в построенном векторном пространстве.

Существует множество способов оценить близость слов через близость строк матрицы слова-контексты. В работе [51] сравниваются некоторые популярные подходы: косинусная близость, евклидово расстояние, метрика Манхэттена, расстояние Бхаттачарья, расстояние Хеллингера, дивергенция Кульбака-Лейблера. По результатам четырех различных задач косинусная близость показала наилучшие результаты.

Помимо использования векторов напрямую для оценки близости определенных слов, построенные представления могут использоваться как признаковые описания и подаваться на вход алгоритмам машинного обучения. Например, для классификации текстов, кластеризации слов или выражений, построения онтологий предметных областей, информационного и разведочного поиска.

Далее мы подробно рассмотрим возможные математические модели, возникающие на этапах 3 и 4.

## 1.3. Математические модели векторных представлений

В литературе предложено большое число моделей, которые позволяют получить векторные представления слов. Некоторые их них, например, языковые модели, решают задачу генерации текста и получают векторные представления как побочный продукт. Другие, например, тематические модели, вообще редко рассматриваются в контексте дистрибутивной семантики. Цель данного раздела – систематическое изложение большого числа методов и демонстрация их тесной взаимосвязи. Будем обращать внимание на следующие ключевые компоненты: тип подсчитываемых статистик по корпусу, оптимизационная задача и ограничения на параметры, численный метод оптимизации.

Введем некоторые обозначения. Пусть $W$ – размер словаря, $T$ – размерность скрытого слоя или, другими словами, число компонент вектора. Мы также будем использовать $W$ и $T$, чтобы обозначать сами множества слов и компонент соответственно. Большинство моделей в данном разделе будут параметризированы двумя матрицами: $\Phi^{W\times T}$ и $\Theta^{T\times W}$. Через $\phi_w$ будем обозначать вектор-строку для слова $w$, аналогично, через $\theta_w$ – вектор-столбец для слова $w$. При этом мы предполагаем симметричный случай, при котором словарь контекстов совпадает со словарем слов. Таким образом, имеем задачу построения матричного разложения вида:

$$F^{W\times W} \approx \Phi^{W\times T} \cdot \Theta^{T\times W}, \tag{1.1}$$

где матрица $F$ содержит статистики совместной встречаемости слов, найденные по корпусу, а матрицы $\Phi$ и $\Theta$ — настраиваемые параметры модели (векторные представления слов и контекстов). Краткое резюме рассматриваемых в этой главе моделей приведено в таблице 1.1.

**Модель неотрицательных разреженных представлений.** Модель Non-negative Sparse Embeddings (NNSE, 2012) [6] основана на технике Non-Negative

Таблица 1.1. Низкоранговое матричное разложение для получения представлений слов.

| | | |
|---|---|---|
| NNSE | данные | $F_{uv} = \max(0, \log \frac{n_{uv} n}{n_u n_v})$ или результат SVD |
| | критерий | $\sum_u \left(\|f_u - \phi_u \Theta\|^2 + \|\phi_u\|_1\right) \to \min_{\Phi,\Theta}$ |
| | условия | $\phi_{ut} \geqslant 0, \forall u \in W, t \in T \quad \theta_t \theta_t^T \leqslant 1, \forall t \in T$ |
| | метод | Онлайновый алгоритм из [52] |
| WNTM | данные | $F_{uv} = \frac{n_{uv}}{n_v} = \hat{p}(u\|v)$ |
| | критерий | $\sum_{v \in W} n_v \operatorname{KL}\left(\hat{p}(u\|v) \big\| \langle \phi_u \theta_v \rangle\right) \to \min_{\Phi,\Theta}$ |
| | условия | $\phi_{ut} > 0, \quad \sum_u \phi_{ut} = 1; \quad \theta_{tv} > 0, \quad \sum_t \theta_{tv} = 1$ |
| | метод | Сэмплирование Гиббса |
| SGNS | данные | $F_{uv} = \log \frac{n_{uv} n}{n_u n_v} - \log k$ |
| | критерий | $\sum_u \sum_v n_{uv} \log \sigma\left(\langle \phi_u \theta_v \rangle\right) + k \, \mathbb{E}_{\bar{v}} \log \sigma\left(-\langle \phi_u \theta_v \rangle\right) \to \max_{\Phi,\Theta}$ |
| | условия | Без ограничений |
| | метод | SGD (по корпусу) |
| GloVe | данные | $F_{uv} = \log n_{uv}$ |
| | критерий | $\sum_v \sum_u f(n_{uv}) \left(\langle \phi_u \theta_v \rangle + b_u + \tilde{b}_v - \log n_{uv}\right)^2 \to \min_{\Phi,\Theta,b,\tilde{b}}$ |
| | условия | Без ограничений |
| | метод | AdaGrad (по элементам $F$) |

Sparse Coding [53] и позволяет строить неотрицательные разреженные представления слов с помощью решения следующей оптимизационной задачи:

$$\sum_{u \in W} \left(\|f_u - \phi_u \Theta\|^2 + \|\phi_u\|_1\right) \to \min_{\Phi,\Theta}; \tag{1.2}$$

$$\phi_{ut} \geqslant 0, \ \forall u \in W, \, t \in T \quad \theta_t \, \theta_t^T \leqslant 1, \ \forall t \in T, \tag{1.3}$$

где $f_u$ — строка раскладываемой матрицы $F$, $\phi_u$ — строка матрицы $\Phi$.

Эта модель реализует подход *обучения справочника (dictionary learning)*. Строки матрицы $\Theta$ задают новый базис и интерпретируются как «записи справочника». Предполагается, что они имеют ограниченную $L_2$-норму. Каждое слово представляется в виде смеси таких записей. Веса смеси для слова $u$ определяются строкой $\phi_u$, причем вводится ограничение неотрицательности весов, и минимизируется $L_1$-норма для достижения разреженности. Строки матрицы $\Phi$, полученные в результате обучения модели, предлагается использовать как векторные представления слов в прикладных задачах.

В данном методе непринципиально, как получена исходная матрица $F$.

Это могут быть pPMI-счетчики совместной встречаемости слов, частоты слов в документах или результат применения к такого рода данным усеченного SVD-разложения. В случае SVD, матрица $F$ будет несимметричной, каждая строка будет представлять слово в виде $k$-мерного вектора. Записи справочника будут также $k$-мерными. Всего будет $T$ записей, но каждое слово будет представляться смесью небольшого числа записей в силу разреженности матрицы $\Phi$.

Модель обучается онлайновым алгоритмом из [52].

**Тематическая модель сети слов.** Word Network Topic Model, WNTM [54] была предложена как тематическая модель для коротких текстов, преодолевающая проблему чрезмерной разреженности исходных данных. В отличие от традиционных тематических моделей, которые будут рассмотрены в следующей главе, исходные данные собираются не в виде частотной матрицы слова-документы, а в виде частотной матрицы слова-контексты, где в роли контекстов выступают слова из скользящего окна фиксированной ширины. Таким образом, эта модель использует для обучения те же данные, что и другие модели данной главы, и тоже производит некоторое низкоразмерное матричное разложение. Тем не менее, эта модель никогда не рассматривалась как способ построения векторных представлений слов. Подробнее мы вернемся к этому вопросу в главе 5. Здесь же приведем формальную постановку оптимизационной задачи:

$$\sum_{v\in W}\sum_{u\in W} n_{uv} \ln \langle \phi_u, \theta_v \rangle \to \max_{\Phi,\Theta}, \tag{1.4}$$

$$\phi_{ut} \geqslant 0; \quad \sum_{u\in W} \phi_{ut} = 1; \qquad \theta_{tv} \geqslant 0; \quad \sum_{t\in T} \theta_{tv} = 1. \tag{1.5}$$

Столбцы матриц $\theta_v$ и $\phi_t$ в данной модели являются дискретными вероятностными распределениями, на которые накладываются ограничения неотрицательности и нормировки. Дополнительно предполагается, что они имеют априорное распределение Дирихле. Это усложняет задачу (1.4), которая записана для аналога модели WNTM без учета априорных распределений. Обучение модели WNTM производится с помощью сэмплирования Гиббса.

Задача (1.4) эквивалентна разложению матрицы $F = (F_{uv})^{W \times W}$, составленной из эмпирических распределений $\hat{p}(u|v)$, по взвешенной сумме дивергенций Кульбака-Лейблера:

$$F_{uv} = \frac{n_{uv}}{n_v} = \hat{p}(u|v); \quad \sum_{v \in W} n_v \, \mathrm{KL}\left(\hat{p}(u|v) \middle|\middle| \langle \phi_u, \theta_v \rangle\right) \to \max_{\Phi, \Theta}. \tag{1.6}$$

**Нейросетевые языковые модели.** Рассмотрим задачу языкового моделирования, а именно предсказания слова $u$ по предшествующим $n$ словам $v_{1:n}$. Эта задача традиционно решалась марковскими моделями со сглаживанием. В 2003 году была предложена вероятностная нейросетевая модель (Neural Probabilistic Language Model, NPLM) [55], ставшая классической в этой области. В ряде последующих работ ее архитектура упрощалась, а в 2013 было предложено семейство моделей word2vec [1, 56], которые в терминах нейронных сетей содержат только один скрытый слой и не содержат нелинейных преобразований. Этот подход оказался очень удачным для обучения векторных представлений слов. Языковое моделирование в последующие годы ушло в сторону рекуррентных нейронных сетей и их различных модификаций. Приведем краткий обзор языковых нейросетевых моделей, исторически важных для обучения векторных представлений.

**Вероятностная нейросетевая модель языка.** Модель NPLM [55] в процессе предсказания слова $u$ по предшествующим словам $v_{1:n}$ обучает матрицу векторных представлений $\Theta$ размерности $T \times W$. Предсказания осуществляются по формуле:

$$p(u|v_{1:n}) = \mathrm{softmax}(b + Wx + U\mathrm{th}(d + Hx)), \tag{1.7}$$

где $x$ – это вектор размерности $nT \times 1$, составленный из векторных представлений контекстов $\theta_{v_i}$, $i = 1 \ldots n$. Все остальные вектора и матрицы $b$, $W$, $U$, $d$, $H$ – это параметры нейронной сети. Преобразование softmax переводит произвольный вещественный вектор в нормированный неотрицательный вектор той

же размерности (в нашем случае, $W \times 1$):

$$\text{softmax}(z) = \frac{\exp z_k}{\sum_k \exp z_k}. \tag{1.8}$$

Недостатком модели является огромное число параметров и долгое обучение.

**Лог-билинейная языковая модель.** Модель изначально предложена как языковая модель и лишь позднее использована как способ обучения векторных представлений слов. Она гораздо проще модели NPLM [55] и потому лучше применима на практике:

$$p(u|v_{1:n}) = \frac{\exp\left(\phi_u \sum_{i=1}^n C_i \theta_{v_i} + b_u\right)}{\sum_{w \in W} \exp\left(\phi_w \sum_{i=1}^n C_i \theta_{v_i} + b_w\right)}. \tag{1.9}$$

Матрицы $\Phi$ и $\Theta$ содержат векторные представления слов и контекстов. Дополнительно для каждого слова $u$ учитывается скалярный сдвиг $b_u$. Матрицы $C_i$ содержат веса, специфичные для позиции $i$, т.е. коэффициенты векторов контекстов зависят от расстояния до предсказываемого слова. Формула (1.9) имеет прозрачную интерпретацию. Предсказываемое слово $u$ имеет большую вероятность, если его вектор близок к агрегированному вектору контекстов в смысле скалярного произведения. Для получения вероятностей применяется softmax:

$$p(u|v_{1:n}) = \text{softmax}\left(\phi_u \theta_{v_{1:n}} + b_u\right), \quad \theta_{v_{1:n}} = \sum_{i=1}^n C_i \theta_{v_i}. \tag{1.10}$$

Модель называется лог-билинейной (Log-Bilinear Language Model, LBL), т.к. линейна по векторам слов и контекстов после взятия логарифма. В случае предсказаний по одному слову, модель принимает вид:

$$p(u|v) = \text{softmax}(\phi_u C \theta_v + b_u). \tag{1.11}$$

Модель Skip-Gram [1], которая будет рассмотрена далее, упрощает эту формулу еще сильнее, избавляясь от матрицы весов $C$ и вектора сдвига $b_u$.

**Иерархическая лог-билинейная языковая модель.** Рассмотренная лог-билинейная модель обучается по-прежнему долго. Одна из причин – необходимость подсчитывать softmax для каждого предсказываемого слова. Иерархическая лог-билинейная модель [57] обходит эту проблему с помощью иерархического софтмакса.

Все слова словаря организуются в некоторое сбалансированное бинарное дерево, в листьях которого находятся слова. Оно может быть построено случайным образом или объединять в под-деревья слова, близкие по смыслу. И листовые, и не листовые вершины получают некоторые векторные описания $\phi_{node}$ и сдвиги $b_{node}$, которые настраиваются в ходе обучения модели.

Каждое слово кодируется как путь от корня к соответствующей вершине, т.е. бинарным кодом, где каждая цифра равна 1 в случае решения пойти в левое поддерево и 0 в случае решения пойти в правое поддерево. Тогда вероятность слова можно представить как произведение вероятностей всех бинарных решений вдоль пути $\mathrm{Path}_u$ от вершины $u$ до корня дерева:

$$p(u|v_{1:n}) = \prod_{\mathrm{node} \in \mathrm{Path}(u)} \sigma(\phi_{\mathrm{node}} \theta_{v_{1:n}} + b_{\mathrm{node}}), \tag{1.12}$$

Здесь каждая вероятность моделируется сигмоидой, а значит быстро вычислима. Нетрудно показать, что это корректная вероятностная модель, обеспечивающая нормировку и неотрицательность распределений. Представления $\phi_{node}$ для листовых вершин можно использовать в качестве векторных представлений слов.

**Модели CBOW и Skip-Gram.** Следующим шагом упрощения языковых моделей, позволившим обработать данные больших объемов и получить высокое качество векторных представлений слов, стало семейство моделей word2vec [56]. Принято выделять две архитектуры и считать их в некотором смысле противоположными. В одной происходит предсказание слова по его окрестности (CBOW), в другой – предсказание окрестности по слову (Skip-Gram). Распи-

шем эти модели более подробно, чтобы увидеть, насколько они, на самом деле, близки.

Будем обозначать через $H_i$ множество индексов из окрестности позиции $i$, не включая $i$. Под окрестностью будем понимать окно фиксированной ширины $h$, так что $H_i$ содержит $2h$ индексов: $H_i = \{i-h, \dots, i-1, i+1, \dots, i+h\}$. Через $w_i$ будем обозначать слово на позиции $i$ в корпусе, где нумерация сквозная от 1 до суммарной длины текстов $N$.

При введенных обозначениях модель Skip-Gram имеет следующий вид:

$$p(w_{i-h}, \dots, w_{i+h}|w_i) = \prod_{j \in H_i} p(w_j|w_i) = \prod_{j \in H_i} \frac{\exp\langle \phi_{w_j}, \theta_{w_i} \rangle}{\sum_w \exp\langle \phi_w, \theta_{w_i} \rangle} = \frac{1}{Z_i} \prod_{j \in H_i} \exp\langle \phi_{w_j}, \theta_{w_i} \rangle. \quad (1.13)$$

Для каждой словопозиции в корпусе моделируются слова в скользящем окне, при этом предполагается их независимость.

Модель CBOW, напротив, моделирует центральное слово для каждого скользящего окна контекстов и имеет вид:

$$p(w_i|w_{i-h}, \dots, w_{i+h}) = \frac{\exp\left\langle \phi_{w_i}, \sum_{j \in H_i} \theta_{w_j} \right\rangle}{\sum_w \exp\left\langle \phi_w, \sum_{j \in H_i} \theta_{w_j} \right\rangle} = \frac{1}{Z'_i} \prod_{j \in H_i} \exp\langle \phi_{w_i}, \theta_{w_j} \rangle. \quad (1.14)$$

Модели CBOW и Skip-Gram, как и другие, параметризированы двумя матрицами векторных представлений. Их могут называть *входными* и *выходными* векторами, подчеркивая, на каком слое нейронной сети они используются. Также их могут называть векторами слов и векторами контекстов, подчеркивая, что каждое слово в тексте можно рассматривать и как слово, и как контекст для соседних слов. К сожалению, понятие *контекст* сильно перегружено. В литературе контекстом называют также всю окрестность данной словопозиции (скользящее окно). Таким образом, создается неверное впечатление, что модели хранят вектора, относящиеся сразу к *группам слов*.

Для оптимизации обеих моделей используется подход максимизации правдоподобия, где правдоподобие является произведением выражений (1.13) или

(1.14) по позициям в корпусе $i$. Таким образом, на уровне формул модели отличаются только нормировочными константами (т.к. моделируют вероятности в разных пространствах). На уровне алгоритма оптимизации, стохастический градиентный спуск организован по-разному: в случае модели Skip-Gram перебираются отдельные пары $(w_i, c_j)$, а в случае модели CBOW – более сложные объекты $(w_i, c_{i-h}, \dots c_{i+h})$.

Согласно Миколову[1], модель Skip-gram лучше применима для редких слов и маленьких корпусов, однако в [34] показано обратное.

**Модель SGNS.** Один из подходов, используемых на практике для обучения модели Skip-Gram, это *негативное сэмплировние (negative sampling)*. Оно позволяет избежать вычисления нормировочных констант в (1.13) и таким образом эффективно обучаться на больших коллекциях. Стоит заметить, что с этим подходом существенно меняется оптимизируемый функционал и постановка задачи в целом, поэтому можно говорить о SGNS (Skip-Gram Negative Sampling) как об отдельной модели.

Решается задача бинарной классификации: по данной паре (слово $u$, контекст $v$) необходимо определить, встречаются ли они в корпусе совместно. Напомним, что под контекстом понимается слово из словаря, а под совместной встречаемостью – окно фиксированной ширины. Модель параметризована вещественными матрицами $\Phi$ и $\Theta$. Обучение заключается в оптимизации следующего функционала:

$$\sum_{u \in W} \sum_{v \in W} n_{uv} \big[ \log \sigma \langle \phi_u, \theta_v \rangle + k \, \mathbb{E}_{\bar{v}} \log \sigma \left( - \langle \phi_u, \theta_{\bar{v}} \rangle \right) \big] \to \max_{\Phi, \Theta}. \tag{1.15}$$

Этот функционал можно интерпретировать как логистическую функцию потерь для бинарной классификации. Первое слагаемое отвечает за положительные примеры со-встречаемости слов, которые мы наблюдаем в корпусе. Второе слагаемое на практике означает сэмплирование случайных контекстов, обеспе-

[1] https://code.google.com/archive/p/word2vec/

чивая отрицательные примеры:

$$\mathbb{E}_{\bar{v}} \log \sigma\left(-\langle \phi_u, \theta_{\bar{v}} \rangle\right) \approx \frac{1}{k} \sum_{s=1}^{k} \log \sigma\left(-\langle \phi_u, \theta_{\bar{v}_s} \rangle\right), \tag{1.16}$$

где $\bar{v}_s \sim p(v)^{3/4}$ — это слова, сэмплируемые из распределения слов в корпусе, возведенного в степень $\tau = \frac{3}{4}$, в результате чего распределение слов становится ближе к равномерному. Оставив в стороне дискуссию о чисто эвристическом выборе коэффициента $\tau$, обратим внимание на другую деталь исходной реализации модели [1]. Параметр $k$, отвечающий в (1.15) за баланс между положительными и отрицательными примерами, выбран равным параметру $k$ из (1.16), отвечающим за точность приближения математического ожидания. Таким образом, на каждую положительную пару $(u, v)$ приходится $k$ отрицательных пар $(u, \bar{v}_s)$. Связанность двух различных по смыслу параметров может приводить к неожиданным артефактам, в частности, в [58] обсуждается странная геометрия полученного векторного пространства. Слова и контексты проецируются в узкие конусы, направленные в *противоположные стороны*, что может быть связано с преобладанием отрицательных примеров в обучении. В модели GloVe, описание которой будет представлено ниже, такого не происходит.

С точки зрения численного метода, оптимизация в модели SGNS осуществляется одной из модификаций стохастического градиентного спуска. При этом обучение происходит онлайновым проходом по корпусу текстов. Таким образом, не требуется ни хранение матрицы совместной встречаемости слов, ни ее предварительный подсчет по корпусу.

**Модель SGNS как матричное разложение.** В работе [59] показано, что оптимизация функционала (1.15) соответствует разложению матрицы $F$ смещенных PMI-оценок пар слов (shifted PMI, sPMI):

$$F_{uv} = \log \frac{p(u, v)}{p(u)p(v)} - \log k = \log \frac{n_{uv} n}{n_u n_v} - \log k, \tag{1.17}$$

где $k$ – это гиперпараметр (например, 10), а вероятности приближены частотными оценками по корпусу. Несмотря на их простой интуитивный смысл, на практике они могут подсчитываться по-разному. Приведем процедуру из [59], которой мы будем придерживаться. Для каждой словопозиции в исходном корпусе выпишем все пары, в которых она участвует. При ширине окна $h$ получится $2h$ пар (для словопозиций на краях корпуса – меньше). Обозначим за $n$ общее число пар, за $n_{uv}$ число пар $(u, v)$, за $n_u$ число пар, на первом месте у которых стоит $u$, и наконец, за $n_v$ число пар на втором месте у которых стоит $v$. Заметим, что так как каждая реальная совстречаемость двух слов в окне будет выписана дважды, как $(u, v)$ и как $(v, u)$, то полученные счетчики симметричны: $n_{uv} = n_{vu}$; а определение $n_u$ совпадает с определением $n_v$ и равно любой из сумм: $n_u = \sum_v n_{uv} = \sum_v n_{vu}$.

Вернемся к интерпретации SGNS как матричного разложения. Строго говоря, в работе [59] показано лишь следующее: функционал (1.15) достигает своего оптимума, когда скалярное произведение $\langle \phi_u, \theta_v \rangle$ равно значению $F_{uv}$. Однако из-за низкой размерности векторов это значение не может быть достигнуто в точности для всех пар. При этом не очевидно, что происходит в окрестности точки оптимума. Другими словами, остается открытым вопрос о том, какая функция расстояния между $\langle \phi_u, \theta_v \rangle$ и $\mathrm{sPMI}(u, v)$ минимизируется в модели SGNS.

Одно из возможных объяснений дано в [60]. Рассмотрим подробнее, из какого распределения сгенерированы пары в функционале (1.15). Можно представить следующий процесс. Сначала с некоторой априорной вероятностью выбирается класс: положительный или отрицательный. В данном случае, вероятность положительного класса $\alpha = \frac{1}{1+k}$. Затем используются следующие частотные вероятности в классах:

$$p(u, v|+) = \frac{n_{uv}}{n}; \qquad p(u, v|-) = \frac{n_u}{n}\frac{n_v}{n}. \tag{1.18}$$

Применим формулу Байеса аналогично тому, как это происходит в опти-

мальных байесовских классификаторах, и получим формулу для апостериорной вероятности классов:

$$p(+|u,v) = \frac{p(+)p(u,v|+)}{p(u,v)} = \frac{\frac{1}{k+1}\frac{n_{uv}}{n}}{\frac{1}{k+1}\frac{n_{uv}}{n} + \frac{k}{k+1}\frac{n_u n_v}{nn}} = \frac{1}{1 + k\frac{n_u n_v}{n_{uv} n}}. \tag{1.19}$$

Теперь заметим, что если выразить получившуюся вероятность с помощью сигмоиды $\sigma(x) = \frac{1}{1+e^{-x}}$, то получим в точности формулу sPMI:

$$p(+|u,v) = \sigma\Big(\ln\frac{n_{uv}n}{n_u n_v} - \ln k\Big). \tag{1.20}$$

Таким образом, задачу (1.15) можно понимать как максимизацию логарифма правдоподобия по выборке пар с бинарными ответами, распределенными согласно (1.20). Или, эквивалентно, как минимизацию дивергенции Кульбака-Лейблера между распределениями:

$$\mathrm{KL}(\sigma(\mathrm{sPMI}(u,v)) \,||\, \sigma(\langle\phi_u, \theta_v\rangle)) \to \min_{\Phi,\Theta}. \tag{1.21}$$

До сих пор мы опускали степень $3/4$ в распределении контекстов. Если ее вернуть, то придем к модифицированной формуле PMI: $\ln\frac{p(u,v)}{p(u)p(v)^{3/4}}$. В [12] показано, что такая модификация предпочтительна во многих задачах.

**Модель GloVe.** Стенфордовская модель *глобальных векторов* GloVe [2] сразу была предложена как низкоранговое матричное разложение. По корпусу текстов строится матрица $F = (F_{uv})^{W\times W}$ логарифмов частот совместной встречаемости слов:

$$F_{uv} = \log n_{uv}.$$

В разложении используется взвешенная квадратичная функция потерь:

$$\sum_{v\in W}\sum_{u\in W} f(n_{uv})\big(\langle\phi_u,\theta_v\rangle + b_u + \tilde{b}_v - \log n_{uv}\big)^2 \to \min_{\Phi,\Theta,b,b'}. \tag{1.22}$$

Дополнительных ограничений нет, матрицы $\Phi$ и $\Theta$ содержат любые вещественные значения. Кроме того, появляются дополнительные параметры модели: вектора сдвига $b_u$ и $\tilde{b}_v$. Интересно, что аналогичные параметры есть в матричных

разложениях для рекомендательных систем, где они называются базовыми предикторами. Функция весов $f(n_{uv})$ монотонно невозрастающая, принимает значение 0 для нулевых счетчиков и штрафует слишком большие счетчики таким образом, чтобы модель не перенастраивалась на них.

Параметры модели настраиваются стохастическим градиентным спуском (методом AdaGrad) по элементам входной матрицы. Таким образом, одним объектом является агрегированная по корпусу совстречаемость двух слов, в то время как в предыдущей SGNS модели одним объектом являлось конкретное вхождение двух слов в текст.

## 1.4. Замечания о терминологии

Из-за быстрого развития области, многие английские термины несут в себе непереводимую отсылку ко времени появления подхода или к научной школе. Например, модели векторного представления слов *(Vector Space Models)* плавно сменились другими моделями векторного представления слов *(Word Embeddings)*. При этом в обоих случаях решается задача представления слова некоторым низкоразмерным вектором, который отражает его семантику. Первый термин приходит из компьютерной лингвистики и ассоциируется с классическими методами, например, преобразованием SVD, примененным к матрице поточечной взаимной информации. Второй термин возник в результате развития глубоких нейронных сетей и необходимости представлять объекты произвольной природы (слова, картинки, сигналы) в виде векторов на входном слое. Подходы этой группы получили широкое распространение в последние годы [1–4]. Термин *ембединг (embedding)* дословно означает погружение объекта в линейное векторное пространство.

Интересно обратить внимание и на другие терминологические особенности. Например, два созвучных термина *distributional vector representations* и *distributed vector representations* обозначают практически противоположные под-

ходы [61]. Первый подход восходит к дистрибутивной гипотезе, предложенной в 1950-ых годах [7, 8] и сводится к построению разреженных высокоразмерных векторов. Второй подход был предложен Хинтоном в 1986 году [62] и стал популярен в языковом моделировании благодаря статье Бенжо 2003 года [55]. Опишем оба подхода более детально.

Дистрибутиваня гипотеза (distributional hypothesis) происходит от слова *distribution* и полагает, что смысл слова полностью определяется частотным *распределением* слов в объединении всего его контекстов. Представим каждое слово вектором из нулей с единственной единицей, соответствующей индексу слова в словаре. В таком случае, каждое слово является уникальной независимой сущностью. Информация о его смысле накапливается из его контекстов. Однако эта информация никаким образом не переиспользуется для семантически близких слов. Это является большой проблемой, особенно при обучении векторных представлений для редких слов.

Для преодоления этого ограничения были предложены так называемые *распределенные (distributed)* представления. Это плотные векторы низкой размерности, у которых каждая компонента отвечает за некоторый (возможно, неинтерпретируемый) признак. Слова и признаки находятся в отношении много ко многим, т.е. знание о смысле слова *распределено* между всеми компонентами вектора. Как правило, такие векторы обучаются как параметры некоторой сложной модели, например, нейронной сети, решающей задачу языкового моделирования. При таком подходе информация о со-встречаемости накапливается для слов совместно, и смысл более редких слов может уточняться с помощью знания о смысле их частотных синонимов.

Несмотря на различность подходов, они тесно связаны. Так, в статье [59] показано, что векторные представления word2vec можно интерпретировать как результат разложения матрицы частотных корпусных оценок. Здесь логично упомянуть еще об одной паре терминов. В литературе часто противопоставляются *частотные (count-based)* и *предсказательные (predictive)* модели. В ста-

тье [63] с говорящим названием “Don’t count! Predict.” показано, что методы, основанные на *обучении* векторов (например, CBOW) существенно превосходят более старые подходы, основанные на разложении матриц частот или других простых статистик. Однако после детального анализа и тщательного подбора гиперпараметров для обоих классов моделей это было опровергнуто [12].

В данной работе мы рассматриваем все упомянутые классы методов и не проводим столь жесткой классификации, т.к. для некоторых моделей она была бы слишком субъективна.

# Глава 2

# Вероятностное тематическое моделирование

Рассматриваются две наиболее популярные тематические модели, а также способы их обучения. Модель вероятностного латентного семантического анализа (Probabilistic Latent Semantic Analysis, PLSA) [13] является одной из первых и классических работ в этой области. Модель латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) [14] является расширением модели PLSA; для ее обучения используются методы байесовского подхода, при этом их детали в литературе по тематическому моделированию часто опускаются.

В данной главе приводится описание EM-алгоритма в общем виде, его применение для максимизации правдоподобия в модели PLSA и максимизации апостериорной вероятности в модели LDA. Для модели LDA также рассматриваются два альтернативных способа обучения: вариационный вывод и сэмплирование Гиббса. Обсуждается взаимосвязь формул вариационного вывода в модели LDA с формулами E-шага обучения PLSA.

Обзорный материал данной главы используется в экспериментах в главе 3, в результате которых выводятся гибридные схемы обучения, позволяющие совмещать полезные свойства известных алгоритмов. В главе 4 рассматриваются ограничения байесовского подхода и предлагается альтернативный метод аддитивной регуляризации тематических моделей.

## 2.1. Задача тематического моделирования

Тематическое моделирование (topic modeling) — одно из современных приложений машинного обучения к анализу текстов, активно развивающееся с конца 90-х годов. *Вероятностная тематическая модель* (ВТМ) коллекции текстовых документов определяет каждую тему как дискретное распределение на множестве терминов, каждый документ — как дискретное распределение на множе-

стве тем. Предполагается, что коллекция документов — это последовательность терминов, выбранных случайно и независимо из смеси таких распределений, и ставится задача восстановления компонент смеси по выборке.

Вероятностная «мягкая» кластеризация документов и терминов по кластерам-темам обходит проблемы синонимии и омонимии слов, возникающие при обычной «жёсткой» кластеризации. Синонимы, появляющиеся в схожих контекстах, с большой вероятностью попадают в одну тему. Омонимы, употребляемые в разных контекстах, распределяются между несколькими темами пропорционально частоте их употребления.

BTM применяются для выявления трендов в научных публикациях и новостных потоках [64, 65], классификации и категоризации документов [66] и изображений [67, 68], семантического информационного поиска [69], в том числе многоязычного [70], тегирования веб-страниц [71], и других приложениях. BTM могут учитывать тематическую иерархию [72], динамику изменения тем во времени и связи слов в предложениях [73], связи между документами через авторство или ссылки, внутреннюю структуру документов, различные особенности языка.

Многочисленные разновидности BTM описаны в обзоре [18]. Большинство моделей являются модификациями модели *латентного размещения Дирихле* LDA [74]. Открытой проблемой является сочетание разнородных требований в рамках одной модели. В частности, для обработки больших коллекций научных публикаций нужна модель, одновременно иерархическая, динамическая, мультиязычная, $n$-граммная, разреженная, робастная, инкрементная, с частичным обучением, и это далеко не полный список требований. Байесовские модели оказываются слишком сложны для совмещения в них более 2–3 требований.

## 2.2. Вероятностный латентный семантический анализ

Введем некоторые обозначения. Пусть дана коллекция $D$ документов, где каждый документ — это последовательность слов. Будем использовать сквозную индексацию словопозиций: $i = 1, \ldots, N$, где $N$ — длина всей коллекции (суммарная длина документов). Про каждую позицию $i$ известно, что она входит в определенный документ $d_i$ и содержит определенное слово $w_i$. Введем обозначение для этой пары $x_i = (w_i, d_i)$ и объединим все пары в совокупность *наблюдаемых переменных* $X$.

Предполагается, что есть некоторый набор тем $\{1, \ldots, T\}$, и с каждой словопозицией в документе связана ровно одна тема $t_i$ из этого набора. обозначим через $Z$ темы всех словопозиций в коллекции. Это *скрытые переменные*, значения которых необходимо восстановить. Число тем $T$ считается гиперпараметром и фиксируется заранее.

Будем моделировать совместную вероятность скрытых и наблюдаемых переменных $p(X, Z)$, при этом сделаем два допущения:

1. *Гипотеза мешка слов:* порядок слов в документах не важен. Все словопозиции независимы.

2. *Гипотеза условной независимости:* вероятность слова при условии темы не зависит от документа: $p(w_i|t_i, d_i) = p(w_i|t_i)$.

С учетом второй гипотезы для каждой словопозиции можно записать:

$$p(x_i, t_i) \equiv p(w_i, d_i, t_i) = p(w_i|t_i, d_i)p(t_i|d_i) = p(w_i|t_i)p(t_i|d_i). \qquad (2.1)$$

Параметрами модели являются две матрицы вероятностных распределений. Матрица $\Phi$ содержит дискретные распределения на множестве слов для каждой темы: $\phi_{wt} = p(w|t)$. Матрица $\Theta$ содержит вероятностные распределения на множестве тем для каждого документа: $\theta_{td} = p(t|d)$.

Модель вероятностного латентного семантического анализа (Probabilistic

Latent Semantic Analysis, PLSA) [75] имеет следующий вид:

$$p(X, Z|\Phi, \Theta) = \prod_{i=1}^{N} p(x_i, t_i|\Phi, \Theta) = \prod_{i=1}^{N} p(d_i)\phi_{w_i t_i}\theta_{t_i d_i}. \tag{2.2}$$

Заметим, что сомножители $p(d_i)$ не содержат настраиваемых параметров, поэтому не будут играть роли при обучении модели.

### 2.2.1. Метод максимума правдоподобия

Для обучения модели PLSA, т.е. настройки параметров $\Phi$ и $\Theta$, воспользуемся методом максимума правдоподобия. В случае, когда наблюдаемыми переменными являются не только слова в документах $X$, но и их темы $Z$, можно сразу же записать:

$$\log \mathcal{L}(\Phi, \Theta) = \log p(X, Z|\Phi, \Theta) = \sum_{i=1}^{N} (\log \phi_{w_i t_i} + \log \theta_{t_i d_i}) + \text{const} \to \max_{\Phi,\Theta}. \tag{2.3}$$

При этом необходимо учесть дополнительные ограничения на $\Phi$ и $\Theta$, т.к. их столбцы образуют дискретные распределения:

$$\forall w, t \quad \phi_{wt} \geqslant 0, \quad \sum_{w=1}^{W} \phi_{wt} = 1; \tag{2.4}$$

$$\forall t, d \quad \theta_{td} \geqslant 0, \quad \sum_{t=1}^{T} \theta_{td} = 1. \tag{2.5}$$

Ограничения неотрицательности выполнятся автоматически. Для учета ограничений нормировки воспользуемся методом множителей Лагранжа:

$$\sum_{i=1}^{N} (\log \phi_{w_i t_i} + \log \theta_{t_i d_i}) - \sum_{t=1}^{T} \lambda_t \left( \sum_{w=1}^{W} \phi_{wt} - 1 \right) - \sum_{d=1}^{D} \mu_d \left( \sum_{t=1}^{T} \theta_{td} - 1 \right) \to \max_{\Phi,\Theta}.$$

Возьмем производную по элементу $\phi_{wt}$ матрицы $\Phi$ и приравняем ее нулю:

$$\frac{1}{\phi_{wt}} \sum_{i=1}^{N} [w_i = w][t_i = t] - \lambda_t = 0. \tag{2.6}$$

Здесь квадратные скобки обозначают индикатор: 1, если выражение внутри скобок истинно, и 0 иначе. Таким образом, $\sum_{i=1}^{N} [w_i = w][t_i = t]$ – это число

раз, когда в коллекции встретилось слово $w$, отнесенное к теме $t$. Обозначим эту величину за $n_{wt}$:

$$n_{wt} = \lambda_t \phi_{wt} \ \forall w = 1, \ldots, W \quad \Rightarrow \quad \sum_{w=1}^{W} n_{wt} = \sum_{w=1}^{W} \lambda_t \phi_{wt}, \quad \Rightarrow \quad \sum_{w=1}^{W} n_{wt} = \lambda_t.$$

Тогда искомая оценка:

$$\phi_{wt} = \frac{n_{wt}}{\sum_{w=1}^{W} n_{wt}}.$$

Получился хорошо интерпретируемый результат: частотная оценка вероятности слова $w$ в теме $t$ — отношение числа раз, когда тема $t$ связывалась со словом $w$, к общему числу появлений темы $t$ в коллекции.

Совершенно аналогичный результат можно получить для параметров $\theta_{td}$:

$$\theta_{td} = \frac{n_{td}}{\sum_{t=1}^{T} n_{td}},$$

где $n_{td} = \sum_{i=1}^{N} [t_i = t][d_i = d]$.

Однако в реальности переменные $Z$ — это скрытые переменные, которые не известны. Таким образом, при обучении PLSA ставится следующая задача:

$$p(X|\Phi, \Theta) = \sum_{Z} p(X, Z|\Phi, \Theta) \to \max_{\Phi,\Theta}. \tag{2.7}$$

Такую задачу называют максимизацией *неполного* правдоподобия, т.к. из функции правдоподобия выведены скрытые переменные $Z$. Чтобы их вывести, производится суммирование по всем возможным значениям набора переменных $Z$. Так как таких значений экспоненциальное число, то напрямую применить метод максимизации правдоподобия не удается, и вместо него используется EM-алгоритм. Подробное описание можно найти в книге [76].

### 2.2.2. EM-алгоритм для максимизации неполного правдоподобия

**EM-алгоритм в общем виде.** Запишем задачу максимизации неполного правдоподобия для вероятностной модели, в которой есть некоторые наблюда-

емые переменные $X$, скрытые переменные $Z$ и параметры $\Omega$:

$$\log p(X|\Omega) \to \max_{\Omega}.$$

Пусть $q(Z)$ — произвольное распределение. Справедлива следующая цепочка равенств:

$$\log p(X|\Omega) = \int q(Z) \log p(X|\Omega) dZ =$$
$$\int q(Z) \log \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)} dZ = \int q(Z) \log \frac{p(X,Z|\Omega)}{q(Z)} \frac{q(Z)}{p(Z|X,\Omega)} dZ =$$
$$\underbrace{\int q(Z) \log p(X,Z|\Omega) dZ - \int q(Z) \log q(Z) dZ}_{L(q,\Omega)} + \underbrace{\int q(Z) \log \frac{q(Z)}{p(Z|X,\Omega)} dZ}_{\mathrm{KL}(q(Z)||p(Z|X,\Omega))}. \quad (2.8)$$

Дивергенция Кульбака-Лейблера $\mathrm{KL}(q(Z)||p(Z|X,\Omega))$ оценивает расстояние между двумя распределениями. Дивергенция Кульбака-Лейблера неотрицательна, несимметрична и равна нулю тогда и только тогда, когда распределения совпадают.

В силу неотрицательности $\mathrm{KL}(q(Z)||p(Z|X,\Omega))$ слагаемое $L(q,\Omega)$ является нижней оценкой на величину $\log p(X|\Omega)$. От максимизации $\log p(X|\Omega)$ по $\Omega$ предлагается перейти к максимизации нижней границы $L(q,\Omega)$ по $q$ и $\Omega$. Такая постановка в общем случае может давать приближенный ответ, однако оказывается существенно более простой. Основная идея EM-алгоритма заключается в том, чтобы итеративно повторять два шага:

1. $L(q,\Omega) \to \max\limits_{q}$;
2. $L(q,\Omega) \to \max\limits_{\Omega}$.

На первом шаге максимизация $L(q,\Omega)$ по $q$ эквивалентна минимизации $\mathrm{KL}(q(Z)||p(Z|X,\Omega))$, т.к. их сумма $\log p(X|\Omega)$ от $q$ не зависит. Из свойств дивергенции Кульбака-Лейблера следует, что минимум равен 0 и достигается при $q(Z) = p(Z|X,\Omega)$. Поэтому если удается выписать аналитически распределение $p(Z|X,\Omega)$, то именно его и нужно взять в качестве $q$, при этом нижняя оценка $L(q,\Omega)$ будет являться точной нижней оценкой.

Рассмотрим теперь второй шаг:

$$\int q(Z)\log p(X,Z|\Omega)dZ - \int q(Z)\log q(Z)dZ \to \max_{\Omega} \Leftrightarrow$$

$$\Leftrightarrow \int q(Z)\log p(X,Z|\Omega)dZ \to \max_{\Omega}, \quad (2.9)$$

т.к. второе слагаемое не зависит от $\Omega$. Первое слагаемое соответствует математическому ожиданию:

$$\int q(Z)\log p(X,Z|\Omega)dZ = \mathbb{E}_{q(Z)}\log p(X,Z|\Omega).$$

Таким образом, EM-алгоритм заключается в чередовании двух типов шагов: E-шаг (Expectation) соответствует подготовке к вычислению математического ожидания, M-шаг (Maximization) – максимизации математического ожидания логарифма правдоподобия по параметрам.

- **E-шаг:** $\mathrm{KL}(q(Z)||p(Z|X,\Omega)) \to \min\limits_{q(Z)} \Leftrightarrow q(Z) = p(Z|X,\Omega)$;
- **M-шаг:** $\mathbb{E}_{q(Z)}\log p(X,Z|\Omega) \to \max\limits_{\Omega}$.

Заметим, что из описанной процедуры следует следующее утверждение о сходимости EM-алгоритма.

**Утверждение 1.** *Последовательность значений параметров, получаемых в ходе итераций EM-алгоритма, дает неубывающую последовательность значений величины $L(q,\Omega)$, являющейся нижней оценкой логарифма правдоподобия модели $\log p(X|\Omega)$.*

**Применение EM-алгоритма для обучения PLSA.** Решим задачу (2.7), действуя согласно общей схеме. На E-шаге необходимо оценить распределение скрытых переменных при условии наблюдаемых переменных и параметров: $p(Z|X,\Phi,\Theta)$. Т.к. словопозиции независимы, то сразу перейдем к отдельным вероятностям:

$$p(Z|X,\Phi,\Theta) = \prod_{i=1}^{N} p(t_i|x_i,\Phi,\Theta).$$

Чтобы найти эти вероятности, воспользуемся формулой Байеса и для краткости записи опустим $\Phi$ и $\Theta$ после черты:

$$p(t_i|x_i) \equiv p(t_i|w_i, d_i) = \frac{p(w_i|t_i, d_i)p(t_i|d_i)}{\sum_{t=1}^{T} p(w_i|t, d_i)p(t|d_i)} = \frac{\phi_{w_i t_i}\theta_{t_i d_i}}{\sum_{t=1}^{T} \phi_{w_i t}\theta_{t d_i}}. \tag{2.10}$$

Теперь запишем выражение, которое нужно максимизировать на M-шаге:

$$\mathbb{E}_{p(Z|X,\Phi,\Theta)} \log p(X, Z|\Phi, \Theta) = \sum_{i=1}^{N} \mathbb{E}_{p(t_i|x_i,\Phi,\Theta)} \left(\log \phi_{x_i t_i} + \log \theta_{t_i d_i}\right) + \text{const} \to \max_{\Phi,\Theta}.$$

Мы учли, что $q(Z) = p(Z|X, \Phi, \Theta)$, и пронесли математическое ожидание внутрь суммы в силу независимости словопозиций. Теперь распишем по определению:

$$\sum_{i=1}^{N}\sum_{t=1}^{T} p(t_i = t|x_i, \Phi, \Theta) \left(\log \phi_{w_i t} + \log \theta_{t d_i}\right) + \text{const} \to \max_{\Phi,\Theta}. \tag{2.11}$$

Эта задача очень похожа на задачу (2.3), сформулированную в предположении известных тем $Z$. Если записать функцию Лагранжа для учета ограничений нормировки и взять производную по одному элементу $\phi_{wt}$, то получим:

$$\frac{1}{\phi_{wt}} \sum_{i=1}^{N} [w_i = w] p(t_i = t) - \lambda_t = 0.$$

Здесь $\sum_{i=1}^{N} [w_i = w] p(t_i = t)$ аналогично выражению в (2.6) интерпретируется как число раз, когда слово $w$ было отнесено к теме $t$. Однако если раньше мы это число знали точно, то теперь вместо индикаторов тем появляются вероятности, посчитанные на E-шаге. Таким образом, это наилучшая оценка интересующей нас величины. Обозначим ее как и прежде за $n_{wt}$. Аналогично получим оценки $n_{td}$ для документов. Тогда итоговые формулы M-шага будут иметь вид:

$$\phi_{wt} = \frac{n_{wt}}{\sum_{w=1}^{W} n_{wt}}; \quad \theta_{td} = \frac{n_{td}}{\sum_{t=1}^{T} n_{td}} \tag{2.12}$$

Итак, итерационно повторяя формулы (2.10) и (2.12), мы оценим параметры $\Phi$ и $\Theta$, т.е. обучим модель PLSA с помощью EM-алгоритма.

## 2.3. Латентное размещение Дирихле

**Априорные распределения в общем случае.** В байесовском подходе введение априорных распределений $p(\Omega|\alpha)$ — это способ учесть предположения о возможных значениях параметров $\Omega$ до каких-либо наблюдений. Здесь $\alpha$ — это новый гиперпараметр модели. Наблюдаемые переменные уточняют наши представления о значении параметров $\Omega$:

$$\underbrace{p(\Omega|X,\alpha)}_{\text{posterior}} = \frac{p(\Omega,X|\alpha)}{p(X|\alpha)} \propto p(\Omega,X|\alpha) = \underbrace{p(X|\Omega,\alpha)}_{\text{likelihood}}\underbrace{p(\Omega|\alpha)}_{\text{prior}}. \tag{2.13}$$

Таким образом, значение параметров $\Omega$ имеет большую *апостериорную* вероятность, если оно одновременно хорошо вписывается в априорные предположения (prior) и хорошо описывает результаты реальных наблюдений (likelihood, правдоподобие).

**Максимизация апостериорной вероятности.** Одним из возможных способов оценивания параметров модели, имеющих заданные априорные распределения, является *максимизация апостериорной вероятности*:

$$\log p(\Omega|X,\alpha) = \log p(X|\Omega,\alpha) + \log p(\Omega|\alpha) \to \max_{\Omega}. \tag{2.14}$$

Вернемся к общей схеме EM-алгоритма и учтем априорное распределение $p(\Omega)$. Тогда разложение будет выглядеть следующим образом:

$$\log p(X|\Omega,\alpha) + \log p(\Omega|\alpha) = L(q,\Omega) + \mathrm{KL}(q(Z)||p(Z|X,\Omega,\alpha)) + \log p(\Omega|\alpha),$$

где $L(q,\Omega)$ введено в (2.8).

Как и прежде, учтем неотрицательность дивергенции Кульбака-Лейбера и перейдем к максимизации *нижней оценки* логарифма апостериорной вероятности:

$$L(q,\Omega) + \log p(\Omega|\alpha) \to \max_{q,\Omega}. \tag{2.15}$$

Введем обозначение $R(\Omega) = \log p(\Omega|\alpha)$ и выпишем итерации EM-алгоритма для задачи (2.15):

- **E-шаг** остается без изменения, т.к. $\log p(\Omega)$ не зависит от $q$:

$$\mathrm{KL}(q(Z)||p(Z|X,\Omega,\alpha)) \to \min_{q(Z)} \Leftrightarrow q(Z) = p(Z|X,\Omega,\alpha). \tag{2.16}$$

- **M-шаг** содержит оценки параметров $\Omega$, найденные из условия:

$$\mathbb{E}_{q(Z)} \log p(X,Z|\Omega,\alpha) + R(\Omega) \to \max_{\Omega}. \tag{2.17}$$

**Утверждение 2.** *Последовательность значений параметров, получаемых в результате итераций EM-алгоритма* (2.16), (2.17), *дает неубывающую последовательность значений нижней оценки логарифма апостериорной вероятности* (2.14).

Заметим, что при выводе EM-алгоритма мы нигде не пользовались тем фактом, что дополнительное слагаемое $R(\Omega) = \log p(\Omega)$ имеет вероятностную интерпретацию логарифма априорного распределения. Следовательно, рассуждения останутся справедливы для задачи максимизации *регуляризованного правдоподобия*:

$$\log p(X|\Omega,\alpha) + R(\Omega) \to \max_{\Omega}, \tag{2.18}$$

где $R(\Omega)$ — произвольная дифференцируемая функция.

**Утверждение 3.** *Последовательность значений параметров, получаемых в результате итераций EM-алгоритма* (2.16), (2.17) *с произвольной дифференцируемой функцией* $R(\Omega)$, *дает неубывающую последовательность значений нижней оценки логарифма регуляризованного правдоподобия* (2.18).

Последнее утверждение обосновывает сходимость EM-алгоритма для аддитивно регуляризованных тематических моделей, предлагаемых в главе 4.

**Априорные распределения в модели LDA.** Модель латентного размещения Дирихле (Latent Durichlet Allocation, LDA) [14] отличается от PLSA введением *априорных распределений Дирихле* на параметры $\Phi$ и $\Theta$.

Рассмотрим вектор $\theta = (\theta_1, \ldots, \theta_K)$ такой, что $\theta_k \geqslant 0$, $\forall k = 1, \ldots, K$ и $\sum_{k=1}^{K} \theta_k = 1$. То есть это вектор, задающий вероятности $K$ возможных исходов некоторой дискретной случайной величины. Распределение Дирихле — это непрерывное вероятностное распределение на симплексе:

$$\mathrm{Dir}(\theta|\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \quad \alpha_k > 0, \, \forall k = 1, \ldots, K,$$

где $\alpha = (\alpha_1, \ldots, \alpha_K)$ — параметры. Важное свойство распределения Дирихле заключается в том, что если $\alpha_k < 1$, $\forall k = 1, \ldots, K$, то наиболее вероятными будут *разреженные* вектора $\theta$, в которых лишь несколько значений существенно отличны от нуля. При этом заметим, что остальные значения не будут нулевым, а будут положительными, хотя и близкими к нулю величинами. Также нам понадобятся еще несколько свойств:

1. Математическое ожидание: $\mathbb{E}\theta_k = \frac{\alpha_k}{\sum_{i=1}^{K} \alpha_i}$.
2. Мода (точка максимума вероятности): $\hat{\theta}_k^{MP} = \frac{\alpha_k - 1}{\sum_{i=1}^{K} \alpha_i - K}$.
3. Математическое ожидание логарифма: $\mathbb{E} \ln \theta_k = \psi(\alpha_k) - \psi\left(\sum_{i=1}^{K} \alpha_i\right)$,

   где $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ – дигамма-функция.

   При $x > 1$ справедливо приближение: $\exp(\psi(x)) \approx x - \frac{1}{2}$.

В модели LDA предполагается, что вероятности слов для каждой темы $\phi_t$ имеют априорное распределение Дирихле с вектор-параметром $\beta$. Аналогично, вероятности тем для каждого документа $\theta_d$ имеют априорное распределение Дирихле с вектор-параметром $\alpha$:

$$\phi_t \sim \mathrm{Dir}(\phi_t|\beta), \, \forall t = 1, \ldots, T; \quad \theta_d \sim \mathrm{Dir}(\theta_d|\alpha), \, \forall d = 1, \ldots, D.$$

Одна из основных мотиваций такой модели — разреживающее свойство распределения Дирихле: в реальности каждый документ содержит лишь неболь-

шое число тем, а каждая тема описывается лишь небольшим числом слов. Запишем совместную вероятность, задающую модель LDA:

$$p(X, Z, \Phi, \Theta|\alpha, \beta) = \prod_{i=1}^{N} p(d_i)\phi_{w_i t_i}\theta_{t_i d_i} \prod_{d=1}^{D} \mathrm{Dir}(\theta_d|\alpha) \prod_{t=1}^{T} \mathrm{Dir}(\phi_t|\beta). \tag{2.19}$$

Если сравнить это выражение с аналогичным для PLSA (2.2), то заметим, что наборы переменных $\Phi$ и $\Theta$ переместились налево от черты, т.е. теперь оценивается их совместная вероятность с наблюдаемыми переменными $X$ и скрытыми $Z$. В частности, это означает, что теперь мы можем зафиксировать некоторые $\alpha$ и $\beta$ и сгенерировать $X$, $Z$, $\Phi$ и $\Theta$ согласно (2.19). Кроме того, теперь мы можем считать $\Phi$ и $\Theta$ скрытыми переменными (как и темы $Z$), а $\alpha$ и $\beta$ — параметрами модели.

За счет усложнения модели появляется несколько различных сценариев оценивания матриц параметров $\Phi$ и $\Theta$. В PLSA практически единственным разумным сценарием была максимизация неполного правдоподобия. Далее мы рассмотрим три алгоритма обучения модели LDA:

- максимизация апостериорной вероятности (MAP: maximum a posteriori probability);
- вариационный байесовский вывод (VB: Variational Bayes);
- сэмплирование Гиббса (CGS: Collapsed Gibbs Sampling).

Они строятся по-разному, но приводят к близким оценкам. В литературе также используются их различные модификации [77].

### 2.3.1. Метод максимума апостериорной вероятности

Этот метод является наиболее близким аналогом максимизации неполного правдоподобия в модели PLSA. Как следует из названия, необходимо найти значения параметров, в которых достигается максимум апостериорной вероятности $p(\Phi, \Theta|X, \alpha, \beta)$. Переписывая задачу (2.14) в обозначениях модели LDA,

получаем:

$$\log p(X|\Phi,\Theta,\alpha,\beta) + \underbrace{\log p(\Theta|\alpha) + \log p(\Phi|\beta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta}. \tag{2.20}$$

Данная задача может быть решена с помощью EM-алгоритма (2.16), (2.17). Таким образом, в алгоритме LDA-MAP сохраняется E-шаг из алгоритма PLSA-EM, а на M-шаге максимизируется выражение:

$$\sum_{i=1}^{N}\sum_{t=1}^{T} p(t_i = t|x_i,\Phi,\Theta)\,(\log\phi_{w_i t} + \log\theta_{td_i}) + \\ + \sum_{d=1}^{D}\sum_{t=1}^{T}(\alpha_t - 1)\log\theta_{td} + \sum_{t=1}^{T}\sum_{w=1}^{W}(\beta_w - 1)\log\phi_{wt} \to \max_{\Phi,\Theta}. \tag{2.21}$$

при условиях неотрицательности и нормировки (2.4). Если записать функцию Лагранжа и взять производную по одному элементу $\phi_{wt}$, то получим:

$$\frac{1}{\phi_{wt}}\left(\sum_{i=1}^{N}[w_i = w]p(t_i = t) + \beta_w - 1\right) - \lambda_t = 0.$$

Отсюда, аналогично уже разобранным случаям:

$$n_{wt}+\beta_w-1 = \lambda_t\phi_{wt} \Rightarrow \quad \sum_{w=1}^{W}(n_{wt}+\beta_w-1) = \sum_{w=1}^{W}\lambda_t\phi_{wt} \Rightarrow \quad \sum_{w=1}^{W}(n_{wt}+\beta_w-1) = \lambda_t.$$

Искомая оценка:

$$\phi_{wt} = \frac{n_{wt} + \beta_w - 1}{\sum_{w=1}^{W}(n_{wt} + \beta_w - 1)}. \tag{2.22}$$

Заметим, что в этот раз условия неотрицательности автоматически не выполняются, и возможна ситуация, когда $n_{wt}+\beta_w-1 < 0$. Таким образом, формула (2.22) корректна для значений $\beta_w > 1$. Можно показать, что при корректном учете ограничений-неравенств $\phi_{wt} \geqslant 0$, $\theta_{td} \geqslant 0$ отрицательные значения заменяются на нули, и итоговые формулы M-шага имеют вид:

$$\phi_{wt} = \frac{(n_{wt} + \beta_w - 1)_+}{\sum_{w=1}^{W}(n_{wt} + \beta_w - 1)_+}; \quad \theta_{td} = \frac{(n_{td} + \alpha_t - 1)_+}{\sum_{t=1}^{T}(n_{td} + \alpha_t - 1)_+}. \tag{2.23}$$

В результате, при маленьких значениях счетчиков $n_{wt}$ или $n_{td}$ соответствующие вероятности обнулятся, таким образом, сработает разреживающее

свойство априорного распределения Дирихле. Тем не менее, на практике этих обнулений может оказаться недостаточно.

### 2.3.2. Вариационный байесовский вывод

Приведем другой подход к оцениванию матриц $\Phi$ и $\Theta$ в модели LDA, заданной совместным вероятностным распределением $p(X, Z, \Phi, \Theta|\alpha, \beta)$. Поставим задачу максимизации неполного правдоподобия, т.е. максимизации вероятности наблюдаемых данных при условии параметров модели: $p(X|\alpha, \beta) \to \max\limits_{\alpha,\beta}$. Такая постановка неудобна тем, что интересующие нас матрицы $\Phi$ и $\Theta$ пропали, а максимизация осуществляется по гиперпараметрам модели.

Тем не менее, распишем EM-алгоритм для этой задачи:

- E-шаг: $\mathrm{KL}(q(Z, \Phi, \Theta)||p(Z, \Phi, \Theta|X, \alpha, \beta)) \to \min\limits_{q}$;
- M-шаг: $\mathbb{E}_{q(Z,\Phi,\Theta)}\, p(X, Z, \Phi, \Theta|\alpha, \beta) \to \max\limits_{\alpha,\beta}$.

Чтобы получить нулевую дивергенцию Кульбака-Лейблера на E-шаге, необходимо вычислить совместное распределение трех групп скрытых переменных: $p(Z, \Phi, \Theta|X, \alpha, \beta)$. К сожалению, сделать это аналитически не удается. Введем дополнительные упрощающие предположения, и в этих предположениях приближенно оценим искомое распределение. А именно, предположим, что группы переменных взаимно независимы, т.е. будем искать распределение $q$ в виде:

$$q(Z, \Phi, \Theta) \approx \prod_{i=1}^{N} q(t_i) \prod_{t=1}^{T} q(\phi_t) \prod_{d=1}^{D} q(\theta_d). \tag{2.24}$$

Это можно интерпретировать как минимизацию KL-дивергенции по $q$ на множестве допустимых распределений $q$, в данном случае, на множестве распределений вида (2.24). При такой постановке KL-дивергенция может оказаться ненулевой, а нижняя оценка $L(q, \alpha, \beta)$ — неточной оценкой $\log p(X|\alpha, \beta)$. Итоговые оценки параметров будут также неточными. Данный прием называется *приближенным байесовским выводом*.

Итак, теперь задача формулируется так:

$$\mathrm{KL}\left(\prod_{i=1}^{N} q(t_i)\prod_{t=1}^{T} q(\phi_t)\prod_{d=1}^{D} q(\theta_d)\,||\,p(Z,\Phi,\Theta|X,\alpha,\beta)\right)\to\min_{q(t_{di}),q(\phi_t),q(\theta_d)}. \quad (2.25)$$

Она решается итерационным процессом, который на каждом шаге оценивает очередной фактор $q_j$ по всем остальным факторам $q_{\setminus j}$:

$$\log q_j \propto \mathbb{E}_{q_{\setminus j}} \log p(X,Z,\Phi,\Theta|\alpha,\beta). \quad (2.26)$$

Доказательство этого утверждения можно найти, например, в [78].

Распишем формулы байесовского вывода для оценки распределений $Z$, $\Phi$ и $\Theta$. Начнем с $q(\theta_d)$ для некоторого документа $d$. В последующих выкладках нас будет интересовать только зависимость от $\theta_d$, все остальные члены будем опускать. Они повлияют только на нормировочную константу, которую мы найдем отдельно.

$$\begin{aligned}\log q(\theta_d) &\propto \mathbb{E}_{\setminus q(\theta_d)} \log p(X,Z,\Phi,\Theta|\alpha,\beta) \propto \\ &\propto \mathbb{E}_{\setminus q(\theta_d)} \sum_{i=1}^{N}(\log\phi_{w_i t_i} + \log\theta_{t_i d_i}) + \sum_{d=1}^{D}\log \mathrm{Dir}(\theta_d|\alpha) + \sum_{t=1}^{T}\log\mathrm{Dir}(\phi_t|\beta) \propto \\ &\propto \mathbb{E}_{\setminus q(\theta_d)} \sum_{i=1}^{N}[d_i = d]\log\theta_{t_i d} + \log\mathrm{Dir}(\theta_d|\alpha).\end{aligned}$$

Раскроем математическое ожидание и распишем плотность распределения Дирихле, опуская его нормировочную константу:

$$\begin{aligned}\log q(\theta_d) &\propto \sum_{i=1}^{N}\mathbb{E}_{q(t_i)}[d_i = d]\log\theta_{t_i d} + \log\mathrm{Dir}(\theta_d|\alpha) \propto \\ &\propto \sum_{i=1}^{N}[d_i = d]\sum_{t=1}^{T} q(t_i = t)\log\theta_{td} + \sum_{t=1}^{T}(\alpha_t - 1)\log\theta_{td} \propto \\ &\propto \sum_{t=1}^{T}\left(\sum_{i=1}^{N}[d_i = d]q(t_i = t) + \alpha_t - 1\right)\log\theta_{td}.\end{aligned}$$

Проанализируем это выражение как функцию от $\theta_{td}$. Можно заметить, что с точностью до константы оно совпадает с распределением Дирихле:

$$\theta_d \sim \mathrm{Dir}(\theta_d|\gamma); \quad \gamma_t = \sum_{i=1}^{N}[d_i = d]q(t_i = t) + \alpha_t. \tag{2.27}$$

То, что мы получили зависимость от $\theta_{td}$, соответствующую известному распределению, позволяет не вычислять нормировочную константу. Сделать это напрямую часто бывает невозможно. Если апостериорное распределение лежит в том же семействе, что и априорное распределение, то *говорят, что априорное распределение является сопряженным* к функции правдоподобия. Вспоминая представление апостериорного распределения через функцию правдоподобия и априорное распределение $p(\Theta|X, \alpha, \beta) \propto p(X|\alpha, \beta)p(\Theta|\alpha)$, можно перефразировать определение так: априорное распределение образует сопряженную пару с функцией правдоподобия, если при их перемножении получается распределение из того же семейства. В нашем случае правдоподобие является мультиномиальным распределением. Распределение Дирихле является сопряженным к мультиномиальному. Отчасти именно этим мотивирован его выбор в модели LDA.

Остановимся теперь подробнее на результате (2.27). Это оценка апостериорного распределения параметров $\theta_d$. В отличие от метода максимума апостериорной вероятности получено распределение целиком, а не точечная оценка. С одной стороны, появилось больше информации об оцениваемых параметрах. С другой стороны, в большинстве случаев нас по-прежнему интересует одно значение $\theta_d$. Чтобы его получить, можно подсчитать какую-либо статистику распределения (2.27), например, математическое ожидание. Для распределения Дирихле получим:

$$\mathbb{E}\,\theta_{td} = \frac{\gamma_t}{\sum_{t=1}^{T}\gamma_t}.$$

Следуя введенным ранее обозначениям,

$$\mathbb{E}\,\theta_{td} = \frac{n_{td} + \alpha_t}{\sum_{t=1}^{T}(n_{td} + \alpha_t)}.$$

Вместо математического ожидания можно взять моду, тогда

$$\hat{\theta}_{td}^{MP} = \frac{\gamma_t - 1}{\sum_{t=1}^{T}(\gamma_t - 1)} = \frac{n_{td} + \alpha_t - 1}{\sum_{t=1}^{T}(n_{td} + \alpha_t - 1)}.$$

Такая оценка в точности соответствует оценкам (2.23), полученными другим методом, но также из соображений максимизации апостериорного распределения на $\theta_d$.

Однако взятие точечных оценок остается за рамками байесовского вывода, который заключается в итеративном пересчете всех $q_j$ с помощью текущих значений остальных $q_{\setminus j}$ согласно (2.26). Нетрудно показать, что $q(\phi_t)$ вычисляются аналогичным образом:

$$\phi_t \sim \mathrm{Dir}(\phi_t|\lambda); \quad \lambda_w = \sum_{i=1}^{N}[w_i = w]q(t_i = t) + \beta_w. \tag{2.28}$$

Остается вывести формулы для $\log q(t_i = t|w_i = w, d_i = d)$:

$$\mathbb{E}_{\setminus q(t_i)} \log p(X, Z, \Phi, \Theta|\alpha, \beta) \propto \mathbb{E}_{q(\phi_t)} \log \phi_{wt} + \mathbb{E}_{q(\theta_d)} \log \theta_{td} \propto$$
$$\propto \psi(n_{wt} + \beta_w) - \psi\left(\sum_{w=1}^{W}(n_{wt} + \beta_w)\right) + \psi(n_{td} + \alpha_t) - \psi\left(\sum_{t=1}^{T}(n_{td} + \alpha_t)\right). \tag{2.29}$$

Чтобы проинтерпретировать полученный результат, перейдем от логарифма к вероятности и воспользуемся известным приближением для экспонент дигамма-функций:

$$q(t_{di} = t|w_i = w, d_i = d) \propto \frac{n_{wt} + \beta_w - 0.5}{\sum_{w=1}^{W}(n_{wt} + \beta_w) - 0.5} \frac{n_{td} + \alpha_t - 0.5}{\sum_{t=1}^{T}(n_{td} + \alpha_t) - 0.5} \approx \phi_{wt}\,\theta_{td}$$

Таким образом, приближенно выражение (2.29) соответствует формулам пересчета распределения тем в PLSA-MLE и LDA-MAP.

Итак, байесовский вывод дает приближенные оценки распределения скрытых переменных на E-шаге EM-алгоритма. M-шаг заключается в нахождении оценок максимума правдоподобия для гиперпараметров $\alpha$ и $\beta$. На практике он часто опускается, а $\alpha$ и $\beta$ фиксируются. Рекомендации по оптимизации гиперпараметров можно найти в [79].

### 2.3.3. Сэмплирование Гиббса

Collapsed Gibbs Sampling — еще один часто используемый алгоритм обучения модели LDA. В нем приближенно оценивается распределение $p(Z|X,\alpha,\beta)$, затем вычисляются матрицы $\Phi$ и $\Theta$ из принципа максимума правдоподобия.

Первый шаг (collapsing) заключается в том, чтобы проинтегрировать совместное распределение по $\Phi$ и $\Theta$:

$$p(Z|X,\alpha,\beta) = \int p(X,Z,\Phi,\Theta|\alpha,\beta)d\Phi d\Theta. \tag{2.30}$$

Это удается сделать аналитически за счет того, что интеграл представляется в виде произведения двух интегралов, по $\Phi$ и по $\Theta$:

$$p(X,Z|\alpha,\beta) = \underbrace{\int \prod_{i=1}^{N} \theta_{t_i d_i} \prod_{d=1}^{D} \mathrm{Dir}(\theta_d|\alpha) d\Theta}_{I_1} \underbrace{\int \prod_{i=1}^{N} \phi_{w_i t_i} \prod_{t=1}^{T} \mathrm{Dir}(\phi_t|\beta) d\Phi}_{I_2}. \tag{2.31}$$

Распишем интеграл по $\Theta$; по $\Phi$ все будет аналогично.

$$I_1 = \prod_{d=1}^{D} \int \prod_{i=1}^{N} [d_i = d] \prod_{t=1}^{T} \theta_{td}^{[t_i=t]} \mathrm{Dir}(\theta_d|\alpha) d\theta_d =$$

$$= \left( \frac{\Gamma\left(\sum_{t=1}^{T} \alpha_t\right)}{\prod_{t=1}^{T} \Gamma(\alpha_t)} \right)^{D} \prod_{d=1}^{D} \int \prod_{t=1}^{T} \theta_{td}^{\sum_{i=1}^{N} [d_i=d][t_i=t]+\alpha_t-1} d\theta_d.$$

Под знаком каждого интеграла стоит распределение Дирихле без учета нормировочной константы, параметры которого:

$$\tilde{\alpha}_{td} = \sum_{i=1}^{N} [d_i = d][t_i = t] + \alpha_t, \quad t = 1, \ldots, T.$$

Следовательно, окончательно получаем:

$$I_1 = \left( \frac{\Gamma(\sum_{t=1}^{T} \alpha_t)}{\prod_{t=1}^{T} \Gamma(\alpha_t)} \right)^{D} \prod_{d=1}^{D} \frac{\prod_{t=1}^{T} \Gamma(\tilde{\alpha}_{td})}{\Gamma\left(\sum_{t=1}^{T} \tilde{\alpha}_{td}\right)}. \tag{2.32}$$

Таким образом, получено распределение $p(X, Z|\alpha, \beta)$. Чтобы получить из него распределение тем, воспользуемся формулой Байеса:

$$p(Z|X, \alpha, \beta) = \frac{p(X, Z|\alpha, \beta)}{\sum\limits_{Z} p(X, Z|\alpha, \beta)}.$$

В знаменателе стоит сумма по всем наборам $Z$, которую невозможно найти аналитически. Поэтому возникает задача приближенного оценивания параметров $\Phi$ и $\Theta$ по выборке, сгенерированной из распределения $p(Z|X, \alpha, \beta)$. Сэмплирование Гиббса принадлежит семейству методов Монте-Карло на марковских цепях (Markov Chain Monte Carlo, MCMC) и служит для эффективной генерации выборки из многомерного распределения, известного с точностью до нормировочной константы. Согласно нему, нужно запустить итерационный процесс, который на каждом шаге генерирует точку $t_i$ из одномерного распределения $p(t_i|X, Z_{\setminus i}, \alpha, \beta)$, где $Z_{\setminus i}$ – текущие темы всех словопозиций кроме $i$-ой. Тогда через некоторое время процесс сойдется к генерации точек из искомого распределения $p(Z|X, \alpha, \beta)$. Доказательство этого факта можно найти, например, в [78].

Чтобы выписать распределение $p(t_j|X, Z_{\setminus j}, \alpha, \beta)$ для фиксированной позиции $j$ в документе $d_j$, нужно взять уже известное нам совместное распределение $p(X, Z|\alpha, \beta) = I_1 I_2$, оставить только члены, зависящие от $t_j$, и нормировать получившуюся зависимость. Поскольку это одномерное дискретное распределение, то в данном случае нормировка не является проблемой. Снова сократим выкладки вдвое и будем расписывать только сомножитель $I_1$. Выделим из всех сумм отдельно слагаемые, содержащие $t_j$:

$$I_1 = \text{const}\ \frac{\prod\limits_{t=1}^{T} \Gamma\left(\sum\limits_{i=1, i\neq j}^{N} [d_i = d_j][t_i = t] + [t_j = t] + \alpha_t\right)}{\Gamma\left(\sum\limits_{t=1}^{T}\sum\limits_{i=1, i\neq j}^{N} [d_i = d_j][t_i = t] + \alpha_t + 1\right)}.$$

В знаменателе мы воспользовались тождеством $\sum_{t=1}^{T} [t_j = t] = 1$. Введем

обозначения для громоздких счетчиков:

$$n_{td}^{\setminus j} = \sum_{i=1, i \neq j}^{N} [d_i = d][t_i = t].$$

Это число слов в документе $d$, связанных с темой $t$, без учета слова на $j$-ой позиции. В числителе условие $[t_j = t]$ при всех темах $t \neq t_j$ равно 0, и его можно не учитывать. При $t = t_j$ оно равно 1. Воспользуемся свойством гамма-функции $\Gamma(x+1) = x\Gamma(x)$:

$$I_1 = const \frac{\left(n_{t_j d_j}^{\setminus j} + \alpha_{t_j}\right) \prod_{t=1}^{T} \Gamma\left(n_{d_j t}^{\setminus j} + \alpha_t\right)}{\left(\sum_{t=1}^{T} n_{t d_j}^{\setminus j} + \alpha_t\right) \Gamma\left(\sum_{t=1}^{T} n_{t d_j}^{\setminus j} + \alpha_t\right)}.$$

Здесь все оставшиеся гамма-функции не зависят от $t_j$, поэтому их можно опустить в счет нормировочной константы. Проделывая то же самое для $I_2$ в итоге получаем:

$$p(t_j | X, Z_{\setminus j}, \alpha, \beta) \propto \frac{\left(n_{d_j t_j}^{\setminus j} + \alpha_{t_j}\right)}{\sum_{t=1}^{T} \left(n_{t d_j}^{\setminus j} + \alpha_t\right)} \frac{\left(n_{w_j t_j}^{\setminus j} + \beta_{w_j}\right)}{\sum_{w=1}^{W} \left(n_{w t_j}^{\setminus j} + \beta_w\right)}. \tag{2.33}$$

Итак, получено распределение тем. Предположим, мы сделали $S$ сэмплов. Тогда оценить распределения $\phi_{wt}$ и $\theta_{td}$ можно по формулам:

$$\theta_{td} = \frac{n_{td} + \alpha_t}{\sum_{t=1}^{T}(n_{td} + \alpha_t)}, \quad \phi_{wt} = \frac{n_{wt} + \beta_w}{\sum_{w=1}^{W}(n_{wt} + \beta_w)} \tag{2.34}$$

где

$$n_{wt} = \sum_{i=1}^{N} [w_i = w] \sum_{s=1}^{S} \frac{1}{S} [t_i^s = t], \quad n_{td} = \sum_{i=1}^{N} [d_i = d] \sum_{s=1}^{S} \frac{1}{S} [t_i^s = t]. \tag{2.35}$$

Эти оценки можно трактовать как промежуточный вариант между случаем, когда известны темы каждой словопозиции, и случаем, когда известно аналитическое распределение тем. Здесь известны $S$ сэмплов, которые усредняются

для оценивания числа сопоставлений слова $w$ теме $t$ или темы $t$ документу $d$. Для больших коллекций достаточно полагать $S = 1$.

В случае $S = 1$ заметим, что дроби в (2.33) почти совпадают с оценками $\theta_{td}$ и $\phi_{wt}$ в (2.34), поэтому на практике $\Phi$ и $\Theta$ не хранятся, а пересчитываются налету из счетчиков. Единственная разница заключается в исключении одной позиции в корпусе при подсчете счетчиков. Это важное требование из теории MCMC, необходимое для сходимости. Однако в случае тематического моделирования оно оказывается не столь существенным, и распределения сходятся, даже если его нарушить. Эксперименты приведены в следующем разделе.

# Глава 3

# Схемы обучения тематических моделей

В данной главе предлагается обобщённое семейство EM-подобных методов и рассматриваются эвристики регуляризации, сэмплирования, частого обновления параметров, робастности относительно шума и фона. Все они могут включаться независимо друг от друга в любых сочетаниях, порождая как известные модели (PLSA, LDA, SWB), так и новые. Изучаются различные режимы обучения моделей и формулируются рекомендации по выбору эвристик.

## 3.1. Обобщенное семейство EM-подобных алгоритмов

В данном разделе излагается ряд модификаций EM-алгоритма, который интерпретируется как численный метод решения системы уравнений. Рассматриваются эвристики частого обновления параметров, сэмплирования и регуляризации Дирихле.

**Обозначения и предположения.** Пусть $D$ — множество (коллекция) текстовых документов, $W$ — множество (словарь) всех употребляемых в них терминов, $T$ — множество тем. Каждый документ $d \in D$ представляет собой последовательность $n_d$ терминов $(w_1, \ldots, w_{n_d})$ из словаря $W$. Коллекция документов рассматривается как случайная и независимая выборка троек $(w_i, d_i, t_i)$, $i = 1, \ldots, n$ из дискретного распределения $p(w, d, t)$ на конечном множестве $W \times D \times T$.

*Гипотеза независимости или «мешка слов»* позволяет перейти к компактному представлению документа как подмножества $d \subset W$, в котором каждому элементу $w \in d$ поставлено в соответствие число $n_{dw}$ вхождений термина $w$ в документ $d$. *Гипотеза условной независимости* позволяет сформулировать

вероятностную модель порождения коллекции $D$ по известным $p(t \mid d)$ и $p(w \mid t)$:

$$p(w \mid d) = \sum_{t \in T} p(w \mid t)\, p(t \mid d). \tag{3.1}$$

Построение тематической модели — это обратная задача: по известной коллекции $D$ требуется восстановить породившие её $p(t \mid d)$ и $p(w \mid t)$. Обычно число тем $|T|$ много меньше $|D|$ и $|W|$, и задача сводится к поиску приближённого представления заданной матрицы частот

$$F = \big(\hat{p}_{wd}\big)_{W \times D}, \quad \hat{p}_{wd} = \hat{p}(w \mid d) = \tfrac{n_{dw}}{n_d},$$

в виде произведения $F \approx \Phi\Theta$ двух неизвестных матриц меньшего размера — *матрицы терминов тем* $\Phi$ и *матрицы тем документов* $\Theta$:

$$\begin{aligned} \Phi &= (\phi_{wt})_{W \times T}, & \phi_{wt} &= p(w \mid t), & \phi_t &= (\phi_{wt})_{w \in W}; \\ \Theta &= (\theta_{td})_{T \times D}, & \theta_{td} &= p(t \mid d), & \theta_d &= (\theta_{td})_{t \in T}. \end{aligned}$$

Матрицы $F, \Phi, \Theta$ являются *стохастическими*, то есть имеют неотрицательные нормированные столбцы, представляющие дискретные распределения.

**Оценка качества моделей.** Наиболее распространённым внутренним критерием является *перплексия* (perplexity), используемая для оценивания моделей языка в компьютерной лингвистике. Это мера несоответствия или «удивлённости» модели $p(w \mid d)$ токенам $w$, наблюдаемым в документах $d$. Перплексия определяется через лог-правдоподобие (чем меньше, тем лучше):

$$\mathcal{P}(D') = \exp\left(-\frac{1}{n} \sum_{d \in D'} \sum_{w \in d''} n_{dw} \ln p(w \mid d)\right). \tag{3.2}$$

Обычно коллекцию разделяют на обучающую $D$ и контрольную $D'$ случайным образом в пропорции $9:1$ [74]. Параметры $\phi_{wt}$ оцениваются по обучающей выборке. Каждый документ $d$ контрольной коллекции $D'$ случайным образом делится на две половины, $d'$ и $d''$. Параметры $\theta_{td}$ оцениваются по $d'$. Перплексия вычисляется по $d''$.

На графиках данной главы приводится зависимость перплексии от номера итерации обучения (одна итерация — это один проход по коллекции). Число итераций 40; число тем $|T| = 100$. Вопрос выбора числа тем дополнительно исследуется в четвертой главе и в работе [25].

**Рациональный EM-алгоритм для модели PLSA.** Во введенных обозначениях запишем задачу максимизации логарифма правдоподобия при ограничениях нормировки и неотрицательности для модели PLSA (2.2):

$$L(\Phi,\Theta) = \ln \prod_{d\in D}\prod_{w\in d} p(w\,|\,d)^{n_{dw}} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}; \tag{3.3}$$

$$\sum_{w\in W} \phi_{wt} = 1, \quad \phi_{wt} \geqslant 0; \qquad \sum_{t\in T} \theta_{td} = 1, \quad \theta_{td} \geqslant 0. \tag{3.4}$$

**Теорема 1.** *Точка* $(\Phi,\Theta)$ *локального экстремума задачи* (3.3), (3.4) *удовлетворяет системе уравнений со вспомогательными переменными* $p_{tdw}$*:*

$$p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}}; \tag{3.5}$$

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}, \qquad n_t = \sum_{w\in W} n_{wt}; \tag{3.6}$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}, \qquad n_d = \sum_{t\in T} n_{td}. \tag{3.7}$$

Доказательство следует из условий Каруша–Куна–Таккера. В следующей главе будет сформулировано и доказано более общее утверждение.

Система уравнений (3.5)–(3.7) может быть решена различными численными методами. В частности, метод простых итераций приводит к EM-алгоритму, который чаще всего используется на практике. В Алгоритме 1 EM-итерации организованы так, чтобы E-шаг вычислялся внутри M-шага. Это позволяет избежать хранения трёхмерного массива $p_{tdw}$, содержащего оценки вероятностей тем для всех слововхождений коллекции.

Заметим, что согласно шагам 5–8, если $\theta_{td} = 0$ (тема $t$ не представлена в документе $d$) или если $\phi_{wt} = 0$ (термин $w$ не относится к теме $t$), то нулевое

**Алгоритм 1** ЕМ-алгоритм для тематической модели PLSA.

**Вход:** коллекция документов $D$, число тем $|T|$, начальные приближения $\Theta$ и $\Phi$;

**Выход:** распределения $\Theta$ и $\Phi$;

**повторять**
  обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D,\ w \in W,\ t \in T$;
  **для всех** $d \in D,\ w \in d$
    $Z := \sum_{t \in T} \phi_{wt}\theta_{td}$;
    **для всех** $t \in T$ таких, что $\phi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на $\delta = n_{dw}\phi_{wt}\theta_{td}/Z$;
  $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D,\ t \in T$;
**пока** $\Theta$ и $\Phi$ не стабилизируются.

значение будет сохраняться на протяжении всех итераций. И, наоборот, если все значения $\theta_{td}$, $\phi_{wt}$ были положительны в начальном приближении, то они так и останутся положительными. Таким образом, PLSA не позволяет находить оптимальную структуру разреженности распределений и требует задавать её через начальное приближение.

**Частое обновление параметров.** В ЕМ-алгоритме нет необходимости очень точно решать задачу максимизации правдоподобия на каждом М-шаге. Достаточно сместиться в направлении максимума и затем выполнить Е-шаг. Модификация ЕМ-алгоритма, при которой Е-шаг выполняется чаще, называется *обобщённым ЕМ-алгоритмом* (generalized EM-algorithm, GEM). Для него справедливы те же доказательства сходимости, что и для основного варианта ЕМ-алгоритма [80].

Обобщённый ЕМ-алгоритм в случае PLSA сводится к более частому обновлению параметров $\theta_{td}$ и $\phi_{wt}$ по значениям счётчиков $\hat{n}_{wt}$ и $\hat{n}_{dt}$. В Алгоритме 1 это происходит после каждого просмотра всей коллекции. На больших коллекциях более частые обновления должны повышать скорость сходимости. Обновления

**Алгоритм 2** Обобщённый EM-алгоритм для тематической модели PLSA.

**Вход:** коллекция документов $D$, число тем $|T|$, начальные приближения $\Theta$ и $\Phi$;

**Выход:** распределения $\Theta$ и $\Phi$;

1: обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$, $n_{dwt}$ для всех $d \in D,\ w \in W,\ t \in T$;
2: **повторять**
3: **для всех** $d \in D,\ w \in d$
4: $Z := \sum_{t \in T} \phi_{wt}\theta_{td}$;
5: **для всех** $t \in T$ таких, что $n_{dwt} > 0$ или $\phi_{wt}\theta_{td} > 0$
6: $\delta := n_{dw}\phi_{wt}\theta_{td}/Z$;
7: увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$ на $(\delta - n_{dwt})$;
8: $n_{dwt} := \delta$;
9: **если** не первая итерация и пора обновить параметры $\Phi$, $\Theta$ **то**
10: $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W,\ t \in T$ таких, что $\hat{n}_{wt}$ изменился;
11: $\theta_{td} := \hat{n}_{dt}/\hat{n}_d$ для всех $d \in D,\ t \in T$ таких, что $\hat{n}_{dt}$ изменился;
12: **пока** $\Theta$ и $\Phi$ не стабилизируются.

можно делать после каждой пары $(d, w)$ или после заданного числа пар $(d, w)$ или после каждого документа. Далее будет показано, что частые обновления ускоряют сходимость.

В Алгоритме 2 выбор условия обновления на шаге 9 оставлен на усмотрение разработчика, при этом на первой итерации частые обновления не делаются, чтобы в счётчиках накопилась информация по всей коллекции. В противном случае оценки параметров $\theta_{td}$ и $\phi_{wt}$ по начальному фрагменту выборки могут оказаться хуже начального приближения. Начиная со второй итерации, для каждой пары $(d, w)$ из счётчиков $\hat{n}_{wt}$ и $\hat{n}_{dt}$ вычитается $n_{dwt}$ — то самое значение $\delta$, которое было к ним прибавлено при обработке пары $(d, w)$ на предыдущей итерации. Таким образом, счётчики $\hat{n}_{wt}$ и $\hat{n}_{dt}$ всегда содержат актуальное значение, сформированное при последнем просмотре всей коллекции.

**Сэмплирование.** В Алгоритме 2 для каждой пары $(d, w)$ хранится весь массив значений $n_{dwt}$, $t \in T$. Даже при небольшом числе тем такой расход памяти на хранение каждой пары $(d, w)$ может оказаться неприемлемым. В то же время, согласно гипотезе разреженности, вхождение термина $w$ в документ $d$ связано, скорее всего, с небольшим числом тем. Эксперименты показывают, что тривиальное отбрасывание близких к нулю значений $n_{dwt}$ на каждом шаге может приводить к накоплению большой систематической ошибки и смещению модели.

В таком случае лучше использовать сэмплирование — для каждой пары $(d, w)$ генерировать $s$ случайных тем $t_{dwi}$, $i = 1, \ldots, s$, из распределения $p(t \mid d, w) = p_{tdw} = \phi_{wt}\theta_{td}$. Тогда число ненулевых значений $n_{dwt}$ будет невелико, и в то же время оценки будут несмещёнными. Сэмплирование можно рассматривать как замену условного распределения $p(t|d, w)$ его эмпирической оценкой по сгенерированной случайной выборке длины $s$:

$$\hat{p}(t \mid d, w) = \frac{1}{s} \sum_{i=1}^{s} \big[t_{dwi} = t\big]. \tag{3.8}$$

Сэмплирование в Алгоритме 2 реализуется путём трёх модификаций:

1) перед шагом 5 сэмплируется $s$ тем $t = t_{dwi}$, $i = 1, \ldots, s$ из $p(t \mid d, w)$;

2) на шаге 5 цикл по всем темам заменяется циклом по $t = t_{dwi}$, $i = 1, \ldots, s$;

3) на шаге 6 вычисляется $\delta := n_{dw}/s$.

Таким образом, в обычном PLSA $n_{dw}$ вхождений термина $w$ в документ $d$ распределяются между $|T|$ темами пропорционально вероятностям $p(t \mid d, w)$, тогда как при сэмплировании задействуется не более $s$ тем. Далее в экспериментах мы будем сравнивать эти опции и обозначать P (proportional) и S (sampling).

Сэмплирование Гиббса для модели LDA [81] во многом аналогично сэмплированию в модифицированном Алгоритме 2. PLSA-GEM с сэмплированием (модифицированный Алгоритм 2) и LDA-GS (Алгоритм 3) имеют несколько отличий, но только одно из них оказывается существенным с точки зрения качества модели.

**Алгоритм 3** Сэмплирование Гиббса LDA-GS.

**Вход:** коллекция $D$, число тем $|T|$, начальные приближения $\Theta$ и $\Phi$, гиперпараметры $\alpha$, $\beta$;

**Выход:** распределения $\Theta$ и $\Phi$;

обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\forall d \in D$, $\forall w \in W$, $\forall t \in T$;
**повторять**
  **для всех** $d \in D$, $w \in d$, $i = 1, \ldots, n_{dw}$
    **если** не первый проход коллекции **то**
      $t := t_{dwi}$; уменьшить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на 1;
    вычислить $\phi_{wt}$, $\theta_{td}$ согласно (3.9);
    сэмплировать $t_{dwi}$ из $p(t \mid d, w) \propto \phi_{wt}\theta_{td}$;
    $t := t_{dwi}$; увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на 1;
**пока** $\Theta$ и $\Phi$ не стабилизируются.
обновить $\phi_{wt}$, $\theta_{td}$, $\forall d \in D$, $\forall w \in W$, $\forall t \in T$;

1. В LDA-GS жёстко фиксируется параметр $s = n_{dw}$. Однако *гипотеза разреженности* предполагает, что появление термина $w$ в документе $d$ вряд ли может быть связано с большим числом тем. В наших экспериментах $s = 5$ тем оказалось достаточно, но одной темы явно мало, см. рис. 3.2. Эта эвристика, названная *экономным сэмплированием* [27], сокращает затраты времени и памяти в тех случаях, когда средняя по коллекции величина $n_{dw}$ превышает $s$.

В эксперименте проверялась также гипотеза, что число тем, связанных с парой $(d, w)$, не должно превышать числа употреблений данного слова $n_{dw}$. Для этого производилось сэмплирование $\min\{s, n_{dw}\}$ тем, однако результаты для этой эвристики немного хуже, чем при сэмплировании ровно $s$ тем.

2. В LDA-GS параметры $\phi_{wt}$ и $\theta_{td}$ обновляются предельно часто — после обработки каждого вхождения термина $w$ в документ $d$. В LDA-SEM обновления могут производиться с любой частотой. Эксперименты показывают, что частота обновления влияет на скорость сходимости, но почти не влияет на значение кон-

трольной перплексии в конце итераций, рис. 3.3. По результатам эксперимента можно рекомендовать обновления после каждого термина или после каждого вхождения термина, как в LDA-GS.

3. В LDA-GS перед сэмплированием счётчики уменьшаются на единицу. Тем самым при оценивании распределений не учитывается $i$-е вхождение термина $w$ в документ $d$, для которого сэмплируется тема $t_{dwi}$. Из теории следует, что эта особенность исключительно важна [82]. Однако в экспериментах с коллекциями достаточно больших размеров оказывается, что она не влияет на качество модели — кривые «термин 1 раз» и «термин 1 раз (GS)» на рис. 3.3 практически совпадают. Можно одновременно уменьшать счётчики для старой темы и увеличивать для новой, как в Алгоритме 2.

4. Единственным существенным различием, влияющим на качество модели, является применение байесовской регуляризации в LDA, которая подробнее рассматривается в следующем параграфе. Различие заключается только в формулах частотных оценок условных вероятностей: в PLSA используются несмещённые оценки максимального правдоподобия (3.6)–(3.7), в LDA — байесовские сглаженные оценки (3.9).

Таким образом, LDA-GS отличается от PLSA-EM тремя эвристиками: частым обновлением параметров, сэмплированием и регуляризацией. Эти эвристики не связаны друг с другом и могут применяться в любых сочетаниях.

**Регуляризация.** Тематическая модель LDA [74] основана на разложении (3.1) при дополнительном предположении, что векторы документов $\theta_d \in \mathbb{R}^{|T|}$ и векторы тем $\phi_t \in \mathbb{R}^{|W|}$ порождаются распределениями Дирихле с гиперпараметрами $\alpha = (\alpha_t) \in \mathbb{R}^{|T|}$ и $\beta = (\beta_w) \in \mathbb{R}^{|W|}$ соответственно. Известно несколько способов оценивания параметров $\Theta$ и $\Phi$ в модели LDA, отличающиеся, главным образом, формулой сглаживания частотных оценок вероятностей. Сравнение шести наиболее известных способов в [77] показало, что оптимизация гиперпараметров практически нивелирует различия между ними. В данной работе используются

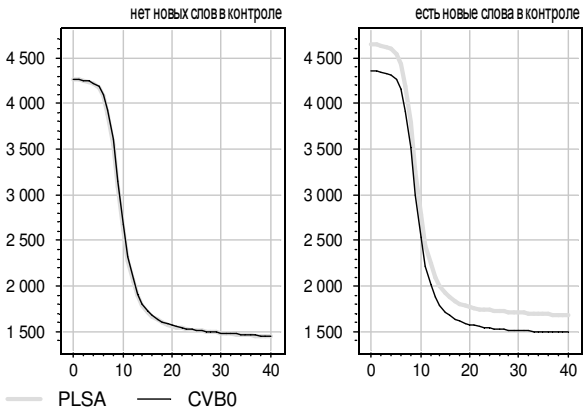



Рис. 3.1. Регуляризация даёт преимущество только когда в контроле есть новые термины (метод CVB0 — это PLSA-GEM с регуляризацией но без сэмплирования).

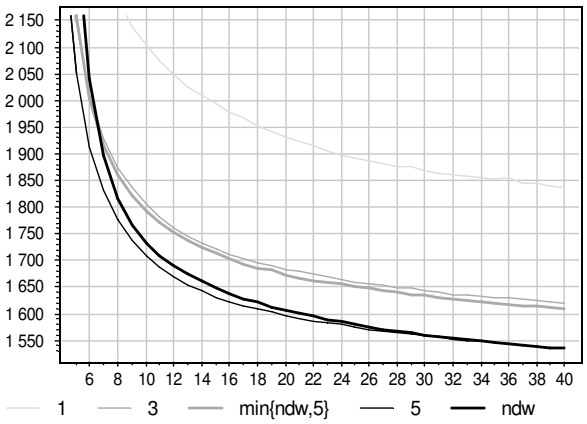



Рис. 3.2. При экономном сэмплировании пяти тем для каждой пары $(d, w)$ перплексия не хуже, чем при сэмплировании $n_{dw}$ тем. Но 1-3 тем не достаточно.

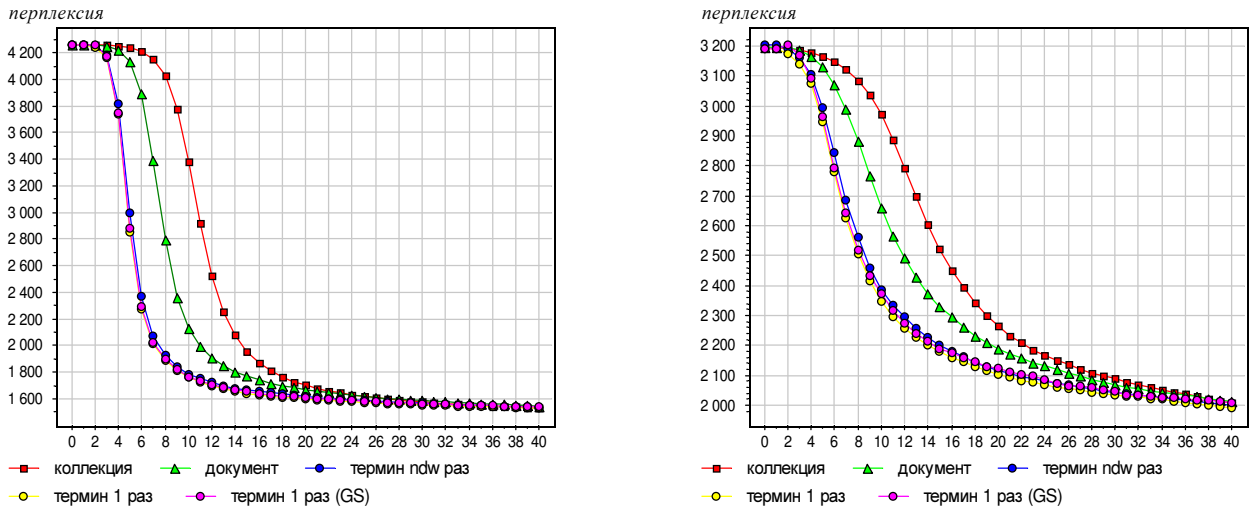
следующие оценки [81, 82]:

$$\phi_{wt} = \frac{\hat{n}_{wt} + \beta_w}{\hat{n}_t + \beta_0}, \quad \beta_0 = \sum_{w \in W} \beta_w; \qquad \theta_{td} = \frac{\hat{n}_{dt} + \alpha_t}{n_d + \alpha_0}, \quad \alpha_0 = \sum_{t \in T} \alpha_t. \tag{3.9}$$

Известно, что LDA обеспечивает существенно меньшие значения контрольной перплексии, чем PLSA [74]. По аналогии с задачами классификации и регрессии отсюда был сделан стандартный вывод, что модель PLSA имеет слишком много параметров $\theta_{td}$, $\phi_{wt}$, на которые не накладывается никаких ограничений, потому возникает переобучение, а в модели LDA эти оценки более устойчивы благодаря байесовской регуляризации, поэтому эффективная сложность модели меньше, и переобучение меньше.

Однако возможна и иная интерпретация этих экспериментов. Оптимальные значения гиперпараметров $\alpha$ и $\beta$ в LDA обычно близки к нулю и могут повлиять только на частотные оценки тем, редких в документе, и терминов, редких в теме. Полезность таких оценок для выявления тематики представляется сомнительной. Контрольная перплексия лучше у LDA только потому, что новым терминам, которых не было в обучающей коллекции, назначаются «чуть более адекватные» априорные оценки вероятностей $\phi_{wt} = \beta_w / \beta_0$.

Эта гипотеза была подтверждена в нашем эксперименте. Если коллекцию




RuDis NIPS

Рис. 3.3. Частое обновление параметров ускоряет сходимость (приведен результат для обобщенного ЕМ-алгоритма с сэмплированием и регуляризацией). Варианты обновлений: после каждого прохода коллекции, после каждого документа, после каждого термина $(d, w)$ по всем $n_{dw}$ его вхождениям, после каждого вхождения термина, GS — с предварительным уменьшением счётчиков как в алгоритме сэмплирования Гиббса. Коллекции: RuDis, NIPS. Число тем $|T| = 100$. Параметры регуляризации: $\alpha_t = 0.5$, $\beta_w = 0.01$.

разбить на обучающую и контрольную так, чтобы в контрольных документах новых терминов не было, то регуляризация не даёт никакого выигрыша, и перплексии PLSA и LDA практически совпадают, см. рис. 3.1. Этот результат согласуется с распространённым мнением, что для больших коллекций нет существенных различий в качестве моделей PLSA и LDA [15–17].

**Данные для обучения моделей.** Численные эксперименты данной главы проведены на двух коллекциях, доступных на странице «Коллекции документов для тематического моделирования»[1]. В ходе предварительной обработки отбрасывались стоп-слова, для русского языка проводилась лемматизация.

*Коллекция RuDis* содержит $|D| = 2000$ авторефератов диссертаций на русском языке; суммарная длина $n \approx 8.7 \cdot 10^6$, объём словаря $|W| \approx 3 \cdot 10^4$. Кон-

[1] На вики-ресурсе `www.MachineLearning.ru`

трольная коллекция $D'$ состоит из 200 авторефератов.

*Коллекция NIPS* содержит $|D| = 1566$ текстов статей научной конференции Neural Information Processing Systems на английском языке; суммарная длина $n \approx 2.3 \cdot 10^6$, объём словаря $|W| \approx 1.3 \cdot 10^4$. Контрольная коллекция $D'$ состоит из 174 документов.

## 3.2. Робастные и разреженные тематические модели

Предположение, что редкие и новые термины бесполезны для тематической модели, приводит к робастным моделям. Нетривиальным и неожиданным результатом сравнения робастных и неробастных версий моделей PLSA и LDA оказывается то, что робастные модели не нуждаются в регуляризации.

**Робастная тематическая модель.** Формализуем предположение о том, что лишь некоторые слова в текстах относятся к каким-либо темам, с помощью вероятностной смеси трёх компонент — тематической, шумовой и фоновой:

$$p(w \mid d) = \frac{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}{1 + \gamma + \varepsilon}; \qquad Z_{dw} = \sum_{t \in T} \phi_{wt}\theta_{td}. \tag{3.10}$$

*Шумовая компонента* $\pi_{dw} \equiv p_{\text{ш}}(w \mid d)$ — это слова, специфичные для конкретного документа $d$, либо редкие термины, относящиеся к темам, слабо представленным в данной коллекции. Отнесение шумовых слов к темам загрязняет распределения $\phi_{wt} = p(w \mid t)$, увеличивает перплексию и искажает тематическую модель.

*Фоновая компонента* $\pi_w \equiv p_{\text{ф}}(w)$ — это общеупотребительные слова, в частности, стоп-слова, не отброшенные на стадии предварительной обработки. Фоновые слова имеют значимые вероятности во многих темах и только мешают различать темы.

*Тематическая компонента* $Z_{dw}$ совпадает с моделью PLSA. Если она плохо объясняет избыточную частоту слова в документе, то слово относится к шу-

му или фону. Параметры $\gamma$ и $\varepsilon$, ограничивающие долю таких слов, связаны с априорными вероятностями тематической, шумовой и фоновой компонент, равными $\frac{1}{1+\gamma+\varepsilon}$, $\frac{\gamma}{1+\gamma+\varepsilon}$, $\frac{\varepsilon}{1+\gamma+\varepsilon}$ соответственно.

Похожая модель SWB (special words with background) на основе LDA и сэмплирования Гиббса предлагалась в [83]. В данной работе робастность рассматривается как ещё одна эвристика, дополняющая обобщённую модель PLSA/LDA, и экспериментально исследуются её сочетания с другими эвристиками.

Задача максимизации правдоподобия (3.3) для модели (3.10) решена в [27]. По аналогии со стандартным EM-алгоритмом, на E-шаге для каждой пары $(d, w)$ вычисляются по формуле Байеса условные вероятности тем $p_{tdw} = p(t \mid d, w)$,

$$p_{tdw} = \frac{\phi_{wt}\theta_{td}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}, \quad t \in T, \tag{3.11}$$

а также условные вероятности того, что слово $w$ является шумом $H_{dw}$ и фоном $H'_{dw}$:

$$H_{dw} = \frac{\gamma\pi_{dw}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}; \qquad H'_{dw} = \frac{\varepsilon\pi_w}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}. \tag{3.12}$$

На M-шаге переменные $\theta_{td}$ и $\phi_{wt}$ вычисляются по прежним формулам (3.6) и (3.7) с единственным отличием, что $p_{tdw}$ вычисляются по новой формуле (3.11). Переменные $\pi_{dw}$ и $\pi_w$ вычисляются как частотные оценки условных вероятностей шума и фона:

$$\pi_{dw} = \frac{\nu_{dw}}{\nu_d}, \qquad \nu_{dw} = n_{dw}H_{dw}, \qquad \nu_d = \sum_{w \in d} \nu_{dw},$$
$$\pi_w = \frac{\nu'_w}{\nu'}, \qquad \nu'_w = \sum_{d \in D} n_{dw}H'_{dw}, \qquad \nu' = \sum_{w \in W} \nu'_w,$$

где $\nu_d$ и $\nu'$ — оценки числа шумовых слов в документе $d$ и фоновых слов во всей коллекции. Эти формулы для $\pi_{dw}$ и $\pi_w$ называются *мультипликативным M-шагом*. Они порождают ту же проблему разреженности, что и переменные $\phi_{wt}$ и $\theta_{td}$: если в начальном приближении значение $\pi_{dw}$ или $\pi_w$ не равно нулю, то оно так и останется ненулевым.

Таблица 3.1. Контрольная перплексия $\mathscr{P}$ и оценки апостериорной вероятности шума $\hat{p}_{\text{ш}}$ и фона $\hat{p}_{\text{ф}}$ при различных значениях $\gamma$ и $\varepsilon$ (после 40 итераций, $|T| = 100$).

RuDis, $\varepsilon = 0.01$:

| $\gamma$ | $\mathscr{P}$ | $\hat{p}_{\text{ш}}$ |
|---|---|---|
| 0 | 1540 | 0.000 |
| 0.001 | 1434 | 0.026 |
| 0.01 | 1277 | 0.090 |
| 0.05 | 1076 | 0.196 |
| 0.1 | 974 | 0.266 |
| 0.3 | 805 | 0.413 |
| 0.5 | 750 | 0.498 |

RuDis, $\gamma = 0.3$:

| $\varepsilon$ | $\mathscr{P}$ | $\hat{p}_{\text{ф}}$ |
|---|---|---|
| 0 | 797 | 0.000 |
| 0.01 | 794 | 0.006 |
| 0.05 | 798 | 0.027 |
| 0.1 | 809 | 0.049 |
| 0.2 | 823 | 0.086 |
| 0.3 | 841 | 0.116 |
| 0.5 | 870 | 0.165 |

NIPS, $\varepsilon = 0.01$:

| $\gamma$ | $\mathscr{P}$ | $\hat{p}_{\text{ш}}$ |
|---|---|---|
| 0 | 2001 | 0.000 |
| 0.001 | 1763 | 0.044 |
| 0.01 | 1381 | 0.152 |
| 0.05 | 991 | 0.296 |
| 0.1 | 818 | 0.377 |
| 0.3 | 604 | 0.527 |
| 0.5 | 525 | 0.598 |

NIPS, $\gamma = 0.3$:

| $\varepsilon$ | $\mathscr{P}$ | $\hat{p}_{\text{ф}}$ |
|---|---|---|
| 0 | 598 | 0.000 |
| 0.01 | 596 | 0.005 |
| 0.05 | 605 | 0.023 |
| 0.1 | 613 | 0.043 |
| 0.2 | 630 | 0.079 |
| 0.3 | 640 | 0.109 |
| 0.5 | 668 | 0.157 |

Формула *аддитивного М-шага*, полученная в [27] из условий Куна–Таккера задачи (3.3), приводит к автоматическому выбору структуры разреженности матрицы $(\pi_{dw})_{D\times W}$:

$$\pi_{dw} = \left( \frac{n_{dw}}{\nu_d} - \frac{Z_{dw} + \varepsilon\pi_w}{\gamma} \right)_+ . \tag{3.13}$$

Эта формула имеет прозрачную интерпретацию: если термин $w$ в документе $d$ встречается существенно чаще, чем предсказывают тематическая и фоновая компоненты модели, то его появление объясняется особенностями данного документа, и тогда $\pi_{dw} > 0$.

Обучение робастной модели осуществляется с помощью Алгоритма 4. *Регуляризация* вводится заменой частотных оценок (3.6)–(3.7) параметров $\phi_{wt}$, $\theta_{td}$ на шагах 5, 6, 15 сглаженными оценками (3.9). *Сэмплирование* вводится заменой распределения $\tilde{H}_{dw}$ его эмпирической оценкой, аналогичной (3.8), при вычислении переменных $\delta_{\text{т}}$, $\delta_{\text{ш}}$, $\delta_{\text{ф}}$ (шаги 11, 12, 13).

Зависимость перплексии от параметров $\gamma$ и $\varepsilon$, как правило, монотонная, причём параметр $\gamma$ гораздо сильнее влияет на перплексию, чем $\varepsilon$, см. таблицу 3.1. С ростом $\gamma$ перплексия уменьшается, так как компонента шума близка к униграммной модели документа, $\pi_{dw} \approx n_{dw}/n_d$, которая наиболее точно предска-

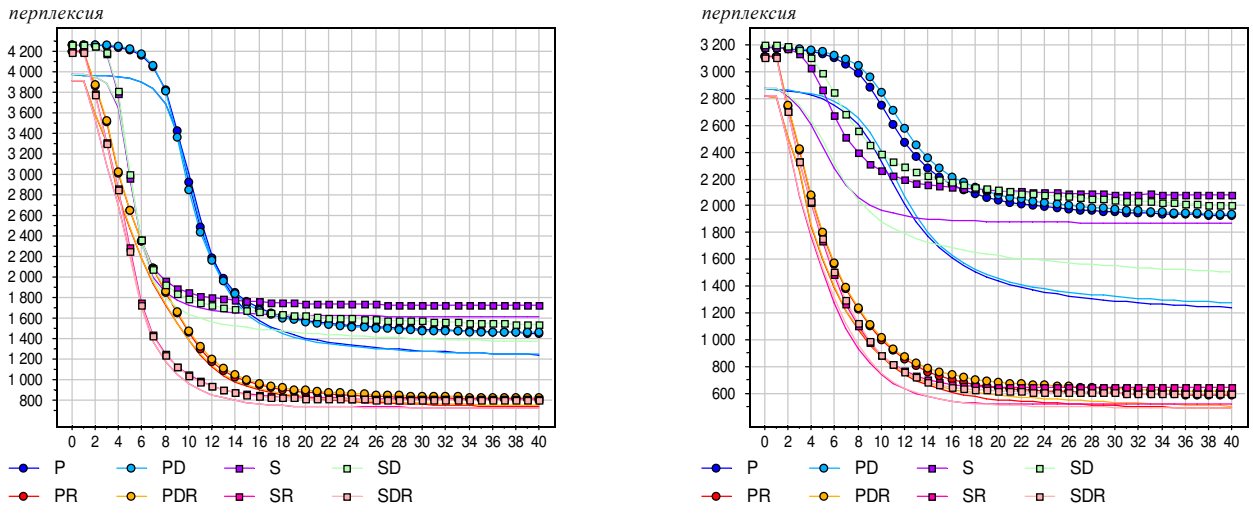



Рис. 3.4. Зависимость контрольной перплексии от числа итераций для всевозможных сочетаний эвристик: D — регуляризация Дирихле ($\alpha_t = 0.5$, $\beta_w = 0.01$); R — робастность ($\gamma = 0.3$, $\varepsilon = 0.01$); S — сэмплирование ($s = n_{dw}$), P — пропорциональное распределение; $|T| = 100$. Тонкие кривые без точек — перплексия обучающей выборки.

зывает вероятности слов $p(w \mid d)$, однако не является тематической. С ростом $\varepsilon$ перплексия увеличивается, так как компонента фона близка к униграммной модели коллекции, $\pi_w \approx n_w/n$, которая хуже предсказывает вероятности слов $p(w \mid d)$, чем тематическая модель. Оценки апостериорных вероятностей шума $\hat{p}_{\text{ш}} = \nu/n$ и фона $\hat{p}_{\text{ф}} = \nu'/n$ также зависят от $\gamma$ и $\varepsilon$ монотонно. Следовательно, оптимальные значения параметров $\gamma$ и $\varepsilon$ должны определяться по внешним критериям качества той прикладной задачи, для решения которой строится тематическая модель.

Наиболее важным выводом исследования робастной модели стало то, что она справляется с переобучение более эффективно, чем регуляризация Дирихле, и не нуждается в дальнейшей регуляризации (см. рис. 3.5).

**Комбинирование сэмплирования, регуляризации и робастности.** Эвристики сэмплирования/пропорционального распределения (S/P), регуляризации Дирихле (D) и робастности (R) могут рассматриваться независимо и образо-

вывать 8 различных алгоритмов. Сочетание SD соответствует LDA-GS. Сочетание P соответствует PLSA-EM. Сочетание SDR соответствует SWB-GS. Однако возможны и другие новые гибридные модели.

Сравнение всех восьми алгоритмов представлено на рис. 3.4. Оно позволяет сделать следующие выводы:

1) для обеих задач робастные алгоритмы существенно превосходят неробастные и гораздо меньше переобучаются;

2) сэмплирование (3.8) немного хуже пропорционального учета всех элементов распределения $p_{tdw}$;

3) сэмплирование без сглаживания может приводить к увеличению перплексии.

Величина переобучения (разность перплексии на обучающей и контрольной выборке) больше зависит от задачи, чем от алгоритма. Сравнение перплексии различных алгоритмов на обучении приводит к тем же качественным выводам, что и сравнение перплексии на контроле. Это приводит к выводу о том, что в данном случае для сравнения алгоритмов не нужна столь сложная методика разделения контрольных документов для вычисления перплексии; достаточно вычислять перплексию только на обучающей выборке.

**Упрощённая робастная модель.** Недостатком предыдущей модели является необходимость подбирать параметры $\gamma$, $\varepsilon$ и хранить параметры $\pi_{dw}$, число которых сопоставимо с размером коллекции. В качестве альтернативы рассмотрим упрощённую робастную модель, в которой фоновая компонента отсутствует, а шумовая компонента $\pi_{dw}$ включается только когда $Z_{dw} = 0$, то есть когда термин $w$ в документе $d$ не является тематическим:

$$p(w \mid d) = \nu_d Z_{dw} + \big[Z_{dw} = 0\big]\pi_{dw}, \tag{3.14}$$

где параметр $\nu_d$ определяется из условия нормировки $\sum_{w \in W} p(w \mid d) = 1$.

Максимизация правдоподобия (3.3) снова приводит к частотным оценкам

условных вероятностей (3.6)–(3.7), но теперь $p_{tdw}$ и $\hat{n}_{dwt}$ оцениваются только по тематическим терминам:

$$\hat{n}_{dwt} = \big[Z_{dw} > 0\big] n_{dw} p_{tdw}.$$

Оптимальное значение $\pi_{dw}$ достаточно определять только для тех $(d, w)$, при которых $Z_{dw} = 0$. Оно также выражается аналитически и совпадает с несмещённой частотной оценкой условной вероятности $p(w \mid d)$, называемой также *униграммной оценкой*:

$$\pi_{dw} = n_{dw} / n_d.$$

Нормировочный множитель $\nu_d$ равен доле тематических терминов в документе:

$$\nu_d = \sum_{w \in W} \big[Z_{dw} > 0\big] \pi_{dw} = \frac{1}{n_d} \sum_{w \in d} \big[Z_{dw} > 0\big] n_{dw}.$$

Заметим, что параметры $\pi_{dw}$ и $\nu_d$ не нужны для вычисления тематической компоненты модели — матриц $\Phi$ и $\Theta$, но могут понадобиться при вычислении перплексии (3.2), которая непосредственно зависит от $p(w \mid d)$.

Упрощённая робастная модель не требует дополнительных затрат памяти или времени. Поэтому в дальнейшем она будет использоваться во всех случаях, за исключением робастной модели (3.10), когда возможно обнуление тематической компоненты $Z_{dw}$.

**Эвристики разреживания.** Гипотеза разреженности предполагает, что коллекция порождается дискретными распределениями $\phi_{wt} = p(w \mid t)$ и $\theta_{td} = p(t \mid d)$, в которых подавляющее большинство вероятностей равны нулю. Следствием этого является также и разреженность распределений $p_{tdw} = p(t \mid d, w)$. Обнуление значительной доли вероятностей $\phi_{wt}$ и $\theta_{td}$ позволяет ускорить ЕМ-алгоритм и хранить тематическую модель в более сжатом виде, открывая возможности для обработки очень больших коллекций.

*Модель PLSA* не оптимизирует структуру разреженности распределений и требует задавать её через начальное приближение. Отдельные значения $\phi_{wt}$

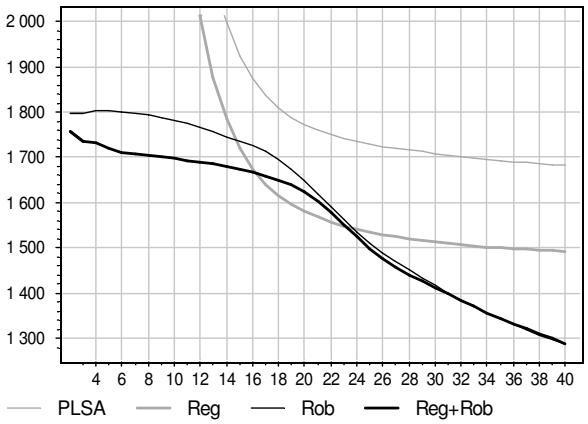



Рис. 3.5. Робастность сильнее уменьшает перплексию PLSA, чем регуляризация. Регуляризация далее не улучшает робастную модель.

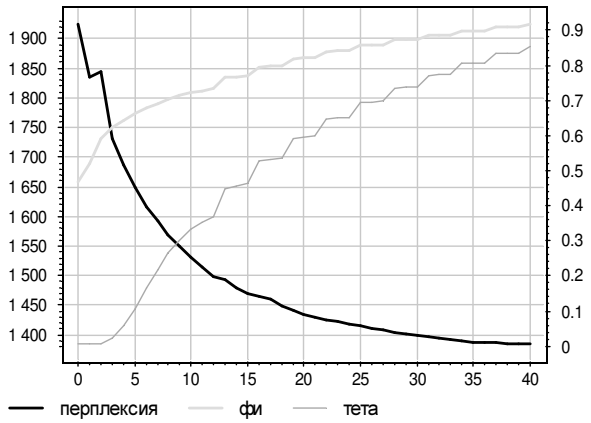



Рис. 3.6. При разреживании доля нулевых $\phi_{wt}$ и $\theta_{td}$ (отложена по правой оси) увеличивается при монотонном уменьшении перплексии.

и $\theta_{td}$ могут в ходе итераций сами собой приближаться к нулю, но, как правило, их доля недостаточна для получения выигрыша в производительности.

*Модель LDA* также не является разреженной — априорные распределения Дирихле запрещают вероятностям $\phi_{wt}$ и $\theta_{td}$ и гиперпараметрам $\beta_w$ и $\alpha_t$ принимать нулевые значения. При стремлении гиперпараметров к нулю распределения Дирихле порождают векторы $\phi_t$ и $\theta_d$, компоненты которых стремятся к нулю, но никогда не обращаются в нуль.

Известные подходы к разреживанию LDA требуют введения дополнительных параметров и усложнения EM-алгоритма. В [84] предлагается хранить не сами значения $\phi_{wt}$ и $\theta_{td}$, а только их разности с фоновыми распределениями. В [85] предполагается, что каждая тема описывается распределением Дирихле на подмножестве слов, заданном бинарными переменными $b_{wt}$ из распределения Бернулли. Сглаженность и разреженность регулируется независимо параметрами распределения Дирихле и распределения Бернулли. Недостатком данной модели является большое число дополнительных скрытых переменных, которые усложняют обучение. В [86] вводится распределение псевдо-Дирихле, которое строится путём расширения области определения распределения Дири-

хле и имеет ограниченную плотность, в то время как распределение Дирихле не ограничено в случае $\alpha < 1$, что и приводит к запрету нулевых значений $\phi_{wt}$ и $\theta_{td}$.

В данной работе исследуются различные стратегии *принудительного разреживания*, когда в конце каждой итерации (полного прохода всей коллекции $D$) обнуляется некоторое количество наименьших значений $\phi_{wt}$ и $\theta_{td}$.

*Робастные модели* допускают разреженность тематической компоненты модели и одновременно исключают ситуацию бесконечной перплексии, так как нулевое значение $Z_{dw}$ компенсируется ненулевым значением шумовой компоненты $p_{\text{ш}}(w \mid d)$. Чем больше $\gamma$, тем более разреженной может быть тематическая компонента модели. В первом эксперименте с робастным PLSA на каждой итерации принудительно обнулялись 5% наименьших значений $\theta_{td}$ и $\phi_{wt}$. При этом разреженность матриц $\Theta$ и $\Phi$ достигала порядка 90% без существенной потери качества модели (рис. 3.6).

Далее исследовались следующие стратегии разреживания.

*Простая стратегия*: в каждом из распределений $\phi_t$, $\theta_d$ обнуляется заданная доля $r$ наименьших *ненулевых* значений. После обнуления производится перенормировка распределений. Число обнуляемых значений сокращается от итерации к итерации, поскольку доля берётся от числа ненулевых значений. Обнуления прекращаются, когда в распределении остаётся $\lfloor r^{-1} \rfloor$ ненулевых значений. Недостатком этой стратегии является стремление к выравниванию доли ненулевых значений во всех распределениях, что представляется довольно странным ограничением.

*Сложная стратегия* устраняет этот недостаток. В каждом из распределений $\phi_t$, $\theta_d$ обнуляется максимальное число наименьших значений, так, чтобы оно не превышало $r|W|$ и $r|T|$ соответственно, и сумма обнуляемых значений не превышала заданного порога $R_\phi$ или $R_\theta$ для распределений $\phi_t$ или $\theta_d$ соответственно. В экспериментах эта стратегия показала лучшие результаты.

Разреживания включаются, начиная с итерации $i_0$, чтобы в распределениях правильно выделились малые вероятности, и делаются не на каждой итера-

ции, чтобы модель успевала восстановить адекватность. В экспериментах разреживания включались на итерациях с номерами $i = i_0 + k\delta,\ \ k = 1, 2, \ldots$, где $i_0$ и $\delta$ — параметры стратегии разреживания.

Разреживание может приводить к обнулению распределения $p(t \mid d, w)$, тогда термин $w$ интерпретируется как нетематический для документа $d$. Поэтому разреживание применяется совместно с робастной моделью (3.10), либо с упрощённой робастной моделью (3.14).

Результаты экспериментов приведены на рис. 3.7, 3.8.

При совмещении упрощённой робастной модели, сэмплирования и разреживания достигается наименьшая перплексия и одновременно наибольшая разреженность матрицы $\Phi$ — до 99.4% для RuDis и 99.6% для NIPS, рис. 3.7.

В робастных алгоритмах с шумом и фоном разреживание почти не влияет на перплексию и позволяет достигать сопоставимой разреженности, рис. 3.8.

Под «агрессивным» разреживанием понимается уменьшение $\delta$ до 1 или уменьшение $i_0$ до 1 или применение сложной стратегии, когда доля обнуляемых значений не уменьшается с итерациями. При агрессивном разреживании или при использовании стохастического EM-алгоритма возможно разреживание распределений $\phi_t$ до 99%. При числе тем $T = 100$ это означает, что каждый термин в среднем относится только к одной теме.

## 3.3. Обсуждение и выводы

Описан широкий класс методов тематического моделирования на базе обобщённого EM-алгоритма и эвристик сглаживания, сэмплирования, частого обновления параметров, робастности и разреживания, которые могут сочетаться в различных комбинациях. В экспериментах на двух текстовых коллекциях получены следующие выводы.

1. Робастные алгоритмы с разреживанием являются лучшими по критерию контрольной перплексии и не требует введения априорных распределений

Дирихле.

2. Контрольная перплексия LDA лучше, чем у PLSA не потому, что PLSA переобучается, а потому, что LDA завышает оценки вероятности редких слов. При корректном сравнении на больших коллекциях перплексии PLSA и LDA практически не различаются.

3. Принудительное разреживание в робастных моделях PLSA позволяет обнулять до 99% параметров без ухудшения контрольной перплексии.

4. Упрощённая робастная модель с разреживанием, в отличие от модели с фоном и шумом, чётко выделяет в документах нетематические термины, не требует хранения параметров $\pi_{dw}$, не требует задания параметров $\gamma$ и $\varepsilon$, и почти не увеличивает объём вычислений.

5. Наряду с сэмплированием Гиббса возможны и другие стратегии разреживания распределений $p(t \mid d, w)$, в частности, сэмплирование небольшого фиксированного числа $s$ тем и постепенное разреживание путём обнуления небольшой доли наименьших вероятностей.

6. На достаточно больших коллекциях ($10^6$ терминов и более) обучающая и контрольная перплексия ведут себя практически одинаково и приводят к одинаковым качественным выводам. Таким образом, нет необходимости вычислять контрольную перплексию.

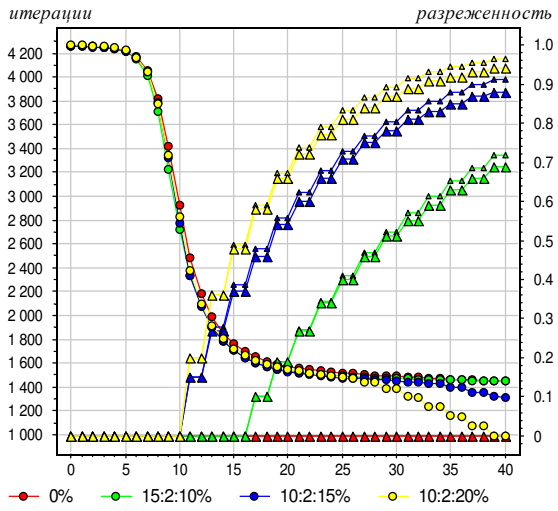
Результаты данной главы опубликованы в работах [22], [27], [28].

**Алгоритм 4** Робастный PLSA-GEM.

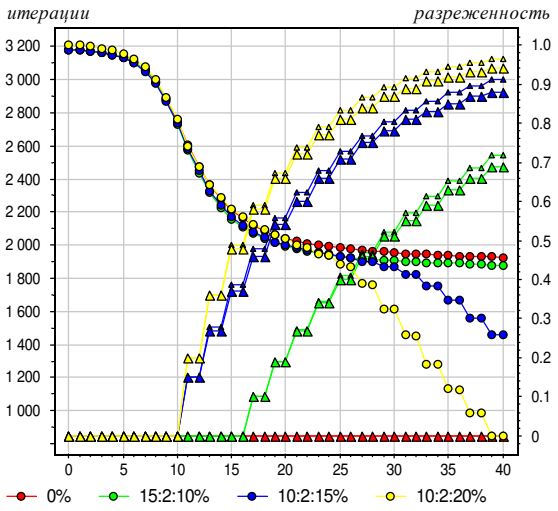
**Вход:** коллекция $D$, число тем $|T|$, начальные приближения $\Theta$, $\Phi$, параметры $\gamma$, $\varepsilon$;

**Выход:** распределения $\Phi$, $\Theta$, $\Pi$;

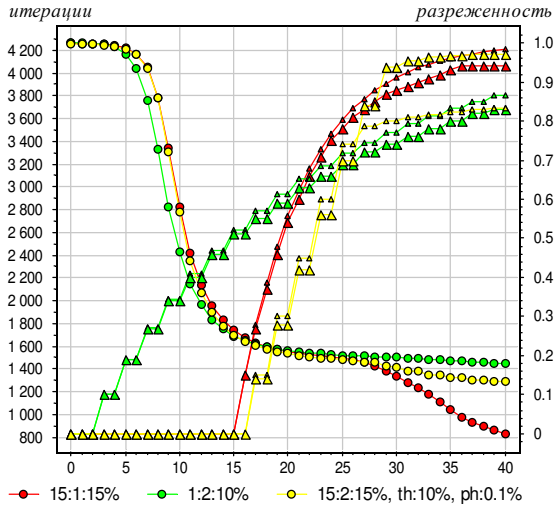
инициализировать $\forall d \in D,\ \forall w \in W,\ \forall t \in T$:
$\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$, $n_{dwt}$, $\nu_{dw}$, $\nu_d$, $\nu$, $\nu'_{dw}$, $\nu'_w$, $\nu' := 0$;
$\pi_{dw} := n_{dw}/n_d$; $\pi_w := n_w/n$;
**повторять**
  **для всех** $d \in D,\ w \in d$
    **если** не первый проход коллекции **то**
      $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t;\ \forall t \in T$;
      $\theta_{td} := \hat{n}_{dt}/\hat{n}_d;\ \forall t \in T$;
      $\pi_w := \nu'_w/\nu'$;
      $\pi_{dw} := (n_{dw}/\nu_d - Z_{dw}/\gamma - \varepsilon\pi_w/\gamma)_+$;
    $Z := Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w$;
    **для всех** $t \in T$: $n_{dwt} > 0$ или $\phi_{wt}\theta_{td} > 0$
      $\delta_т := n_{dw}\phi_{wt}\theta_{td}/Z$; увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$ на $(\delta_т - n_{dwt})$; $n_{dwt} := \delta_т$;
    $\delta_ш := n_{dw}\gamma\pi_{dw}/Z$; увеличить $\nu_d$, $\nu$ на $(\delta_ш - \nu_{dw})$; $\nu_{dw} := \delta_ш$;
    $\delta_ф := n_{dw}\varepsilon\pi_w/Z$; увеличить $\nu'_w$, $\nu'$ на $(\delta_ф - \nu'_{dw})$; $\nu'_{dw} := \delta_ф$;
**пока** $\Phi$, $\Theta$, $\Pi$ не стабилизируются.
обновить $\phi_{wt}$, $\theta_{td}$, $\pi_w$, $\pi_{dw}$ для всех $d, w, t$;



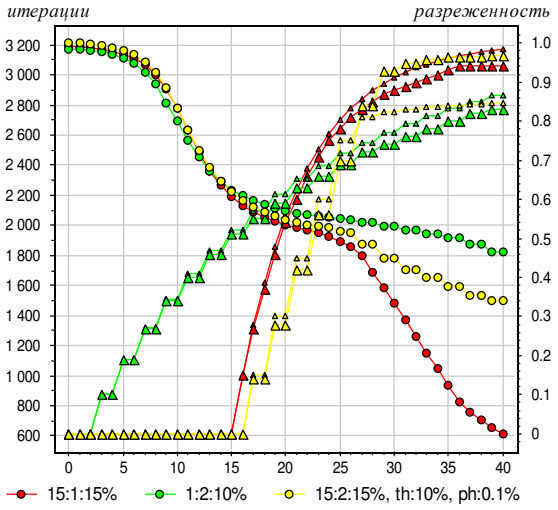

RuDis, P, разреживание через 2 итерации



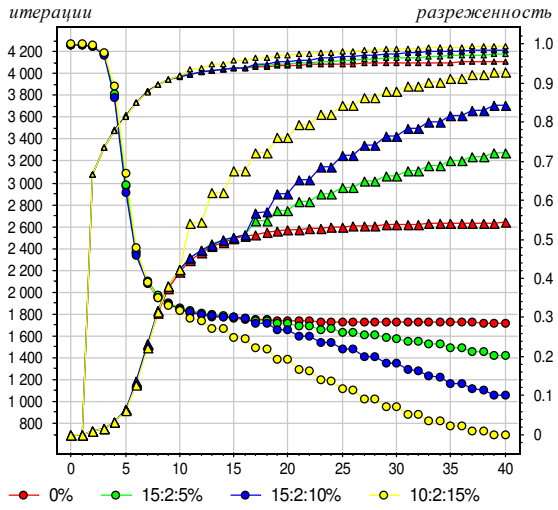

NIPS, P, разреживание через 2 итерации



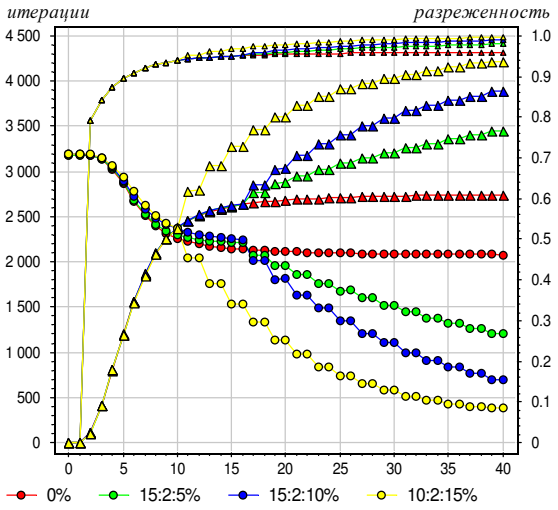

RuDis, P, агрессивное разреживание




NIPS, P, агрессивное разреживание

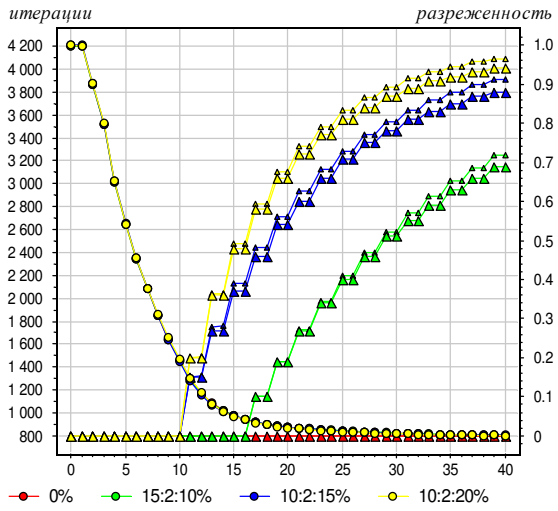



RuDis, S, через 2 итерации

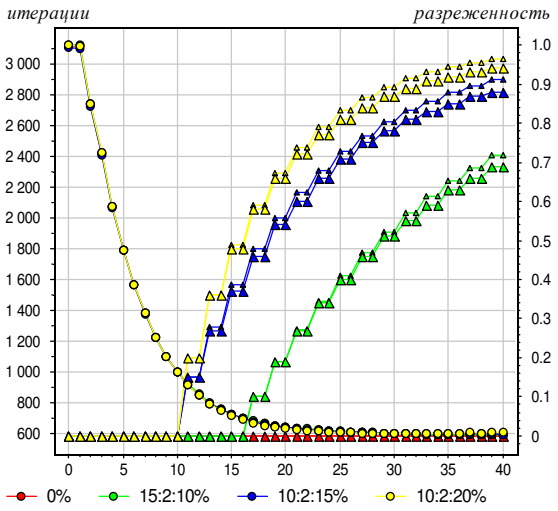



NIPS, S, через 2 итерации

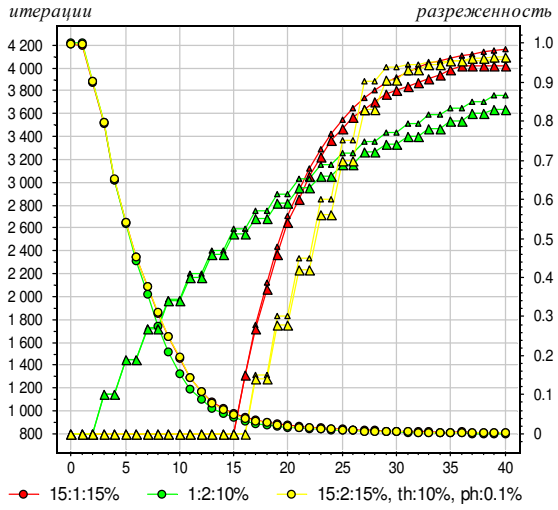
Рис. 3.7. Зависимость перплексии (○) и разреженности матриц $\Phi$ (△) и $\Theta$ (△) от числа итераций для EM-алгоритма с сэмплированием и без при различных параметрах разреживания, обозначаемых $i_0{:}\delta{:}r$, $\mathsf{th}{:}R_\theta$, $\mathsf{ph}{:}R_\phi$. Число тем $|T| = 100$.



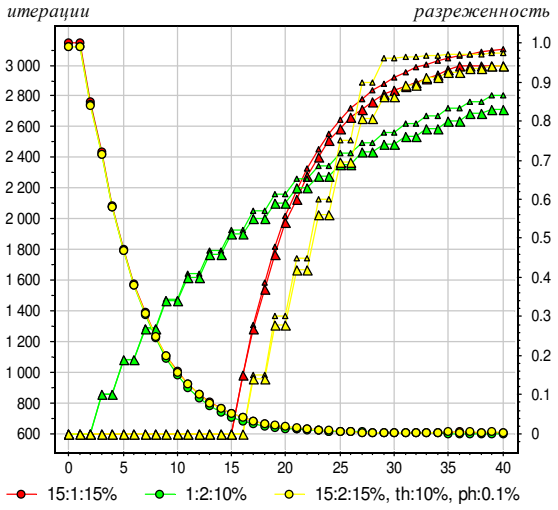

RuDis, PR, разреживание через 2 итерации



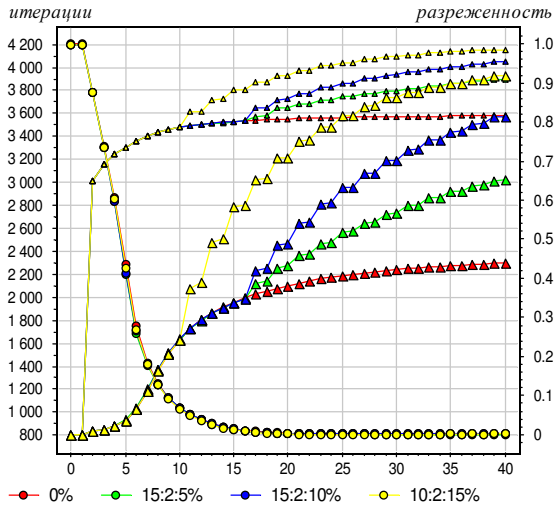

NIPS, PR, разреживание через 2 итерации



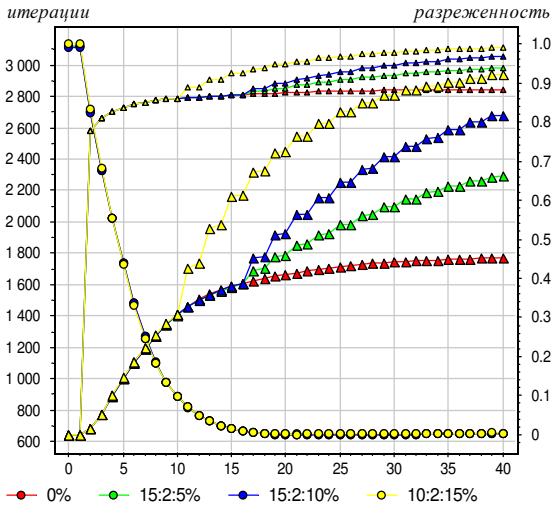

RuDis, PR, агрессивное разреживание




NIPS, PR, агрессивное разреживание




RuDis, SR, через 2 итерации




NIPS, SR, через 2 итерации

Рис. 3.8. Зависимость перплексии (○) и разреженности матриц $\Phi$ (▵) и $\Theta$ (△) от числа итераций робастного EM-алгоритма с сэмплированием и без с параметрами робастности $\gamma = 0.3$, $\varepsilon = 0.01$ и параметрами разреживания $i_0{:}\delta{:}r$, $\mathsf{th}{:}R_\theta$, $\mathsf{ph}{:}R_\phi$. Число тем $|T| = 100$.

# Глава 4

# Аддитивная регуляризация тематических моделей

В данной главе рассматривается альтернатива байесовскому подходу в вероятностном тематическом моделировании — *аддитивная регуляризация тематических моделей* (Additive Regularization of Topic Models, ARTM) [20]. Это приложение классической теории регуляризации некорректно поставленных задач [87] к тематическому моделированию. Обычно построение тематической модели сводится к задаче стохастического матричного разложения. В общем случае она имеет бесконечно много решений, то есть является некорректно поставленной. Для её регуляризации к логарифму правдоподобия добавляются штрафные слагаемые, формализующие дополнительные требования к модели. В частности, предлагается модель предметных и фоновых тем, позволяющая увеличивать интерпретируемость тем с помощью комбинации трех регуляризаторов, воздействующих на блоки матриц $\Phi$ и $\Theta$. Предлагается методика многокритериального оценивания качества модели и экспериментального подбора оптимальной траектории регуляризации.

ARTM имеет несколько принципиальных отличий от байесовского подхода, который в настоящее время является основным в вероятностном тематическом моделировании. Во-первых, не ставится задача построения чисто вероятностной модели порождения текста. Многие лингвистические ограничения легче формализуются с помощью оптимизационных критериев, чем через априорные распределения. Распределение Дирихле утрачивает роль «главного регуляризатора» и уступает место разнообразным проблемно-ориентированным регуляризаторам. Во-вторых, вместо байесовского вывода используется более простой подход — регуляризованный EM-алгоритм. Построение многоцелевых тематических моделей [88] существенно упрощается благодаря аддитивности

регуляризаторов. Добавление регуляризатора требует его дифференцирования по параметрам и небольшой модификации M-шага в готовом EM-подобном алгоритме.

ARTM отличается также и от ранее предлагавшихся методов регуляризации [86, 89–91]. В каждом из них использовался какой-либо конкретный регуляризатор: KL-дивергенция, распределение Дирихле, $L_1$- или $L_2$-норма.

В данной главе предлагаются общие методы регуляризации многоцелевых тематических моделей и описывается набор регуляризаторов, полезных для решения прикладных задач.

В разделе 4.1 вводится аддитивная регуляризация тематических моделей, и обосновывается формула регуляризованного M-шага. Рассматриваются регуляризаторы для сглаживания, разреживания, частичного обучения, декоррелирования, улучшения когерентности.

В разделе 4.2 иллюстрируется применение ARTM для улучшения интерпретируемости тематической модели. Рассматривается проблема комбинирования регуляризаторов и вводится понятие *траектории регуляризации*. Эксперименты показывают, что комбинирование регуляризаторов приводит к улучшению тематической модели по совокупности критериев.

В разделе 4.3 предлагается регуляризатор автоматического отбора тем. Исследуется качество его работы на синтетичесих и реальных данных, а также в комбинации с ранее рассмотренными регуляризаторами повышения интерпретируемости.

В разделе 4.4 приводится обсуждение результатов и основные выводы.

## 4.1. Подход аддитивной регуляризации

Правдоподобие (3.3) зависит только от произведения $\Phi\Theta$, которое определено с точностью до линейного преобразования: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$, при условии, что матрицы $\Phi' = \Phi S$ и $\Theta' = S^{-1}\Theta$ также стохастические. Выбор преоб-

разования $S$ в EM-подобных алгоритмах никак не контролируется и зависит от случайного начального приближения.

Допустим, что наряду с правдоподобием (3.3) требуется максимизировать ещё $r$ критериев $R_i(\Phi,\Theta),\ i=1,\ldots,r$, называемых *регуляризаторами* [87]. Для многокритериальной оптимизации будем максимизировать линейную комбинацию критериев $L(\Phi,\Theta)$ и $R_i(\Phi,\Theta)$ с неотрицательными *коэффициентами регуляризации* $\tau_i$:

$$R(\Phi,\Theta)=\sum_{i=1}^{r}\tau_i R_i(\Phi,\Theta),\quad L(\Phi,\Theta)+R(\Phi,\Theta)\ \to\ \max_{\Phi,\Theta}; \tag{4.1}$$

$$\sum_{w\in W}\phi_{wt}=1,\quad \phi_{wt}\geqslant 0; \qquad \sum_{t\in T}\theta_{td}=1,\quad \theta_{td}\geqslant 0. \tag{4.2}$$

Введем оператор norm, который преобразует произвольный заданный вектор $(x_i)_{i\in I}$ в вектор вероятностей $(p_i)_{i\in I}$ дискретного распределения с помощью обнуления отрицательных элементов и последующей нормировки:

$$p_i=\operatorname*{norm}_{i\in I}(x_i)=\frac{\max\{x_i,0\}}{\sum\limits_{i\in I}\max\{x_i,0\}}.$$

Если $x_i\leqslant 0$ для всех $i\in I$, то результатом оператора norm по определению считается нулевой вектор.

В работе [24] доказана теорема о необходимых условиях локального экстремума задачи (4.1), (4.2).

**Теорема 2.** *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Тогда точка $(\Phi,\Theta)$ локального экстремума задачи (4.1), (4.2), удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}=p(t|d,w)$:*

$$p_{tdw}=\frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}}; \tag{4.3}$$

$$\phi_{wt}=\operatorname*{norm}_{w\in W}\left(n_{wt}+\phi_{wt}\frac{\partial R}{\partial\phi_{wt}}\right); \qquad n_{wt}=\sum_{d\in D}n_{dw}p_{tdw}; \tag{4.4}$$

$$\theta_{td}=\operatorname*{norm}_{t\in T}\left(n_{td}+\theta_{td}\frac{\partial R}{\partial\theta_{td}}\right); \qquad n_{td}=\sum_{w\in d}n_{dw}p_{tdw}. \tag{4.5}$$

*за исключением нулевых столбцов $\Phi$, $\Theta$ в решении данной системы.*

*Доказательство.* Выведем уравнения (4.4) для $\phi_{wt}$. Уравнения (4.5) для $\theta_{td}$ выводятся аналогично.

Запишем необходимые условия локального экстремума Каруша-Куна-Таккера для задачи (4.1), (4.2):

$$\underbrace{\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w\,|\,d)} + \frac{\partial R}{\partial \phi_{wt}}}_{X_{wt}} = \lambda_t - \lambda_{wt}; \quad \lambda_{wt} \geqslant 0; \quad \lambda_{wt}\phi_{wt} = 0; \tag{4.6}$$

где $\lambda_t$ and $\lambda_{wt}$ — множители Лагранжа для ограничений нормировки и неотрицательности соответственно. Умножим обе части первого равенства на $\phi_{wt}$ и выделим вспомогательные переменные $p_{tdw}$:

$$\phi_{wt}\lambda_t = \phi_{wt}X_{wt} = \sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}. \tag{4.7}$$

Фиксируем некоторую тему $t$. Если для всех $w \in W$ значение $X_{wt} \leqslant 0$, то будем считать такую тему вырожденной и исключим ее из модели, положив $\phi_{wt} = 0$, $\forall w \in W$.

Иначе существует слово $u$ такое, что значение $X_{ut} > 0$. Далее рассмотрим два случая для некоторого слова $w \in W$. Если $X_{wt} \leqslant 0$, то $\lambda_{wt} = \lambda_t - X_{wt} > 0$, и из условия дополняющей нежесткости $\phi_{wt} = 0$. Если $X_{wt} > 0$, то из (4.7) имеем $\phi_{wt}\lambda_t = \phi_{wt}X_{wt}$. Объединяя два случая, запишем:

$$\phi_{wt}\lambda_t = \max\left(0, n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right). \tag{4.8}$$

Просуммируем обе части уравнения по всем словам $w \in W$:

$$\lambda_t = \sum_{w\in W} \max\left(0, n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right). \tag{4.9}$$

Наконец, получим формулу М-шага (4.4), выражая $\phi_{wt}$ из (4.8) и (4.9).

□

Решение системы уравнений (4.3)–(4.5) методом простых итераций соответствует регуляризованному ЕМ-алгоритму. В нём сохраняется Е-шаг (3.5), а формулы М-шага заменяются регуляризованными уравнениями (4.4)–(4.5). Таким

образом, EM-алгоритм для обучения регуляризованной модели может быть реализован путём незначительной модификации любого EM-подобного алгоритма. В частности, в Алгоритме 1 достаточно заменить шаги 7 и 8 в соответствии с уравнениями (4.4)–(4.5).

**Замечание о вырожденности.** Будем называть тему $t$ *вырожденной*, если

$$n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} \leqslant 0 \text{ для всех } w \in W.$$

Уравнение (4.4) даёт в этом случае нулевой вектор, который не удовлетворяет требованию нормировки, не является условным распределением $\phi_{wt} = p(w \mid t)$, и формально не может являться решением задачи (4.1), (4.2). Поэтому вырожденную тему приходится исключать из модели. Сокращение числа тем может быть желательным побочным эффектом регуляризации.

Аналогично, будем называть документ $d$ *вырожденным*, если

$$n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \leqslant 0 \text{ для всех } t \in T.$$

Вырожденность документа может означать, что модель не в состоянии его описать, например, если он слишком короткий или не соответствует тематике коллекции. Вырожденный документ фактически исключается из коллекции.

Вырожденность является следствием чрезмерного разреживающего воздействия регуляризатора $R$ на параметры модели. На практике вырожденность возникает довольно редко. Если вырожденность нежелательна, то её можно избежать путём уменьшения коэффициента регуляризации. При постепенном его уменьшении наступает момент, когда условие вырожденности темы перестаёт выполняться хотя бы для одного терма (или, соответственно, условие вырожденности документа перестаёт выполняться хотя бы для одной темы), и нулевой столбец в матрице решения переходит в ненулевой, удовлетворяющий необходимым условиям экстремума по теореме 2.

**Примеры регуляризаторов.** Далее пересматриваются тематические модели, ранее разработанные в рамках байесовского подхода. Для каждой из них удаётся найти соответствующий регуляризатор, который по Теореме 2 приводит к тому же самому или очень похожему алгоритму обучения модели. По сравнению с байесовским подходом, ARTM радикально упрощает вывод алгоритма и позволяет комбинировать регуляризаторы в произвольных сочетаниях.

Мы будем использовать дивергенцию Кульбака–Лейблера (относительную энтропию) как меру различия двух дискретных распределений $(p_i)_{i=1}^n$ и $(q_i)_{i=1}^n$:

$$\mathrm{KL}(p\|q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}.$$

Минимизация KL-дивергенции эквивалентна максимизации правдоподобия модели распределения $q$ по эмпирическому распределению $p$.

**Сглаживающий регуляризатор и модель LDA.** Потребуем, чтобы распределения $\phi_t$ и $\theta_d$ были близки по дивергенции Кульбака–Лейблера к заданным распределениям $\beta = (\beta_w)_{w\in W}$ и $\alpha = (\alpha_t)_{t\in T}$ соответственно:

$$\sum_{t\in T} \mathrm{KL}_w(\beta_w\|\phi_{wt}) \to \min_{\Phi}, \qquad \sum_{d\in D} \mathrm{KL}_t(\alpha_t\|\theta_{td}) \to \min_{\Theta}.$$

Складывая два функционала с коэффициентами $\beta_0, \alpha_0$ и удаляя из суммы константы, получим регуляризатор

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T}\sum_{w\in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d\in D}\sum_{t\in T} \alpha_t \ln \theta_{td} \to \max.$$

Применение общих формул (4.4) и (4.5) даёт то же выражение для М-шага (3.9), что и модель LDA:

$$\phi_{wt} = \underset{w\in W}{\mathrm{norm}}\big(n_{wt} + \beta_0\beta_w\big), \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}\big(n_{td} + \alpha_0\alpha_t\big),$$

если в качестве гиперпараметров взять дискретные распределения $\beta$ и $\alpha$, умноженные на коэффициенты регуляризации: $(\beta_0\beta_t)_{t\in T}$, $(\alpha_0\alpha_w)_{w\in W}$.

Интерпретация регуляризатора через KL-дивергенцию представляется не менее естественной, чем через априорное распределение Дирихле.

**Разреживающий регуляризатор.** Предположим, что каждый документ и каждый термин связан с небольшим числом тем. Тогда среди вероятностей $\phi_{wt}$ и $\theta_{td}$ должно быть много нулевых. При построении тематических моделей больших коллекций с большим числом тем сильная разреженность матриц $\Phi, \Theta$ помогает сократить затраты памяти и времени.

Чем сильнее разрежено распределение, тем меньше его энтропия. Максимальной энтропией обладает равномерное распределение. Поэтому будем максимизировать KL-дивергенцию между модельными распределениями $\phi_t$, $\theta_d$ и заданными распределениями $\beta = (\beta_w)_{w\in W}$, $\alpha = (\alpha_t)_{t\in T}$, например, равномерными:

$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in T}\sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in T} \alpha_t \ln \theta_{td} \to \max. \qquad (4.10)$$

Формулы M-шага, согласно (4.4) и (4.5), отличаются от сглаживающего регуляризатора знаком параметра и приводят к разреживанию распределений:

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\big(n_{wt} - \beta_0\beta_w\big), \qquad \theta_{td} = \underset{t\in T}{\operatorname{norm}}\big(n_{td} - \alpha_0\alpha_t\big).$$

Идея энтропийной регуляризации была предложена в динамической тематической модели PLSA для разреживания распределений тем во времени при обработке видеопотоков [92]. В данной задаче документами являются видеозаписи, терминами — признаки на изображениях, темами — появление определённого объекта в течение определённого времени, например, проезд автомобиля. Однако возможность применения этой же техники для разреживания распределений $\phi_t$ и $\theta_d$ осталась незамеченной.

Многие исследования, направленные на разреживание модели LDA, приводят к чрезмерно сложным конструкциям [84–86, 93, 94], поскольку существует внутреннее противоречие между требованием разреженности и свойством распределения Дирихле не допускать нулевых вероятностей. Наш подход к разреживанию представляется более простым и естественным. Заметим также, что

сглаживание и разреживание описываются одинаково, если не вводить ограничений на знаки параметров $\beta_w, \alpha_t$.

**Сглаживающий регуляризатор для частичного обучения.** Для улучшения интерпретируемости тематической модели могут задаваться обучающие данные для части документов или части тем.

Пусть для документов $d \in D_0$ известно, что они относятся к темам $T_d \subset T$, для тем $t \in T_0$ известно, что к ним относятся термины $W_t \subset W$. Введём регуляризатор, минимизирующий сумму KL-дивергенций между $\phi_{wt}$ и равномерными распределениями на подмножествах терминов $\beta_{wt} = \frac{1}{|W_t|}[w \in W_t]$, а также между $\theta_{td}$ и равномерными распределениями на подмножествах тем $\alpha_{td} = \frac{1}{|T_d|}[t \in T_d]$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \to \max.$$

Формулы M-шага, согласно (4.4) и (4.5), принимают вид

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}}\big(n_{wt} + \beta_0 \beta_{wt}\big)\,[t \in T_0];$$
$$\theta_{td} = \underset{t \in T}{\operatorname{norm}}\big(n_{td} + \alpha_0 \alpha_{td}\big)\,[d \in D_0].$$

Это тоже вариант сглаживания, и ещё одно обобщение LDA, но теперь векторы $\beta, \alpha$ различны для распределений $\phi_t, \theta_d$ и зависят от обучающих данных.

**Декоррелирующий регуляризатор для тем.** Считается, что повышение различности тем улучшает интерпретируемость модели [95]. Регуляризатор, минимизирующий ковариации между вектор-столбцами $\phi_t, \phi_s$,

$$R(\Phi) = -\gamma \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max,$$

приводит к формуле M-шага

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}}\Big(n_{wt} - \gamma \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws}\Big).$$

Согласно этой формуле, вероятности $\phi_{wt}$ наиболее значимых тем слова $w$ в ходе итераций становятся ещё больше. Вероятности менее значимых тем постепенно уменьшаются и могут обращаться в нуль. Таким образом, данный регуляризатор также является разреживающим. Кроме того, он обладает дополнительным полезным свойством группировать стоп-слова в отдельные темы [95].

**Максимизация когерентности.** Тема называется *когерентной*, если термины, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции [96, 97]. Когерентность может оцениваться как по самой коллекции $D$ [98], так и по сторонней коллекции, например, по Википедии [99]. Средняя когерентность тем считается хорошей мерой интерпретируемости тематической модели [97].

Пусть заданы оценки совместной встречаемости $C_{wv} = \hat{p}(w \,|\, v)$ для пар терминов $(w, v) \in W^2$. Обычно $C_{wv}$ оценивают как долю документов, содержащих термин $v$, в которых термин $w$ встречается не далее чем через 10 слов от $v$.

Запишем по формуле полной вероятности условную вероятность $\hat{p}(w \,|\, t)$ через условные вероятности $\phi_{vt} = p(v \,|\, t)$ всех терминов $v$, когерентных с $w$:

$$\hat{p}(w \,|\, t) = \sum_{v \in W \setminus w} C_{wv} \phi_{vt} = \sum_{v \in W \setminus w} \frac{C_{wv} n_{vt}}{n_t}.$$

Введём регуляризатор, требующий, чтобы оценка $\hat{p}(w \,|\, t)$ была согласована с тематической моделью, то есть близка к $\phi_{wt}$ по KL-дивергенции:

$$R(\Phi) = \tau \sum_{t \in T} n_t \sum_{w \in W} \hat{p}(w \,|\, t) \ln \phi_{wt} \to \max.$$

Формула M-шага, согласно (4.4), принимает вид

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}} \Big( n_{wt} + \tau \sum_{v \in W \setminus w} C_{wv} n_{vt} \Big).$$

Эта же формула предлагалась в [98] для модели LDA и алгоритма сэмплирования Гиббса, с более сложным обоснованием через обобщённую урновую

схему Пойя, и более сложной эвристической оценкой $C_{wv}$.

В работе [99] предлагалось использовать другой регуляризатор:

$$R(\Phi) = \tau \sum_{t \in T} \ln \sum_{u,v \in W} C_{uv} \phi_{ut} \phi_{vt} \to \max,$$

и другую оценку совместной встречаемости $C_{uv} = N_{uv} \big[\mathrm{PMI}(u,v) > 0\big]$, где $N_{uv}$ — число документов, в которых термины $u, v$ хотя бы один раз встречаются рядом (не далее, чем через 10 слов), $\mathrm{PMI}(u,v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация (pointwise mutual information), $N_u$ — число документов, в которых термин $u$ встречается хотя бы один раз.

Таким образом, в литературе пока отсутствует единый подход к оптимизации когерентности. Известные подходы легко формализуются в рамках ARTM и не требуют введения априорных распределений Дирихле.

## 4.2. Разреженность и интерпретируемость тем

Интерпретируемость тематической модели является плохо формализуемым требованием. Содержательно оно означает, что по спискам наиболее частотных слов и документов темы эксперт может понять, о чём эта тема, и дать ей адекватное название. Свойство интерпретируемости важно в информационно-поисковых системах для систематизации и визуализации результатов тематического поиска или категоризации документов.

Большинство существующих методов оценивания интерпретируемости основано на привлечении экспертов-асессоров. В [100] экспертам предлагалось непосредственно оценивать полезность тем по трёхбалльной шкале. В методе интрузий [101] для каждой найденной темы составляется список из 10 наиболее частотных слов, в который внедряется одно случайное слово. Тема считается интерпретируемой, если подавляющее большинство экспертов правильно указывают лишнее слово. Экспертные подходы необходимы на стадии исследований, но они затрудняют автоматическое построение тематических моделей.

В серии работ [96, 96, 98, 100] удалось найти величину, которая вычисляется по коллекции автоматически и хорошо коррелирует с экспертными оценками интерпретируемости. Это когерентность (coherence), оценивающая, насколько часто наиболее вероятные слова темы встречаются рядом в документах данной коллекции или во внешней политематической коллекции, такой, как Википедия. Когерентность на сегодняшний день остается основной мерой интерпретируемости тематических моделей, вычисляемой автоматически.

В данной работе предлагается другой подход к формализации понятия интерпретируемости и вводятся дополнительные меры интерпретируемости, также не требующие привлечения асессоров. Предполагается, что интерпретируемая тема должна содержать лексическое ядро — множество слов, характерных для определённой предметной области, которые часто употребляются рядом в документах, с большой вероятностью употребляются в данной теме и практически не употребляются в других темах. Отсюда следует, что из бесконечного множества стохастических матричных разложений $F \approx \Phi\Theta$ нас больше всего интересуют те, в которых матрицы $\Phi$ и $\Theta$ обладают следующей структурой разреженности. Множество тем разбивается на два подмножества, $T = S \sqcup B$: предметные темы $S$ и фоновые темы $B$.

*Предметные темы* $t \in S$ содержат термины предметных областей. Их распределения $p(w\,|\,t)$ разрежены и существенно различны (декоррелированы). Распределения $p(d\,|\,t)$ также разрежены, так как каждая предметная тема присутствует в относительно небольшой доле документов.

*Фоновые темы* $t \in B$ содержат слова общей лексики, которых не должно быть в предметных темах. Их распределения $p(w\,|\,t)$ и $p(d\,|\,t)$ сглажены, так как эти слова присутствуют в большинстве документов. Тематическую модель с фоновыми темами можно рассматривать как обобщение робастных моделей, рассмотренных в третьей главе, в которых использовалось только одно фоновое распределение.

**Комбинирование регуляризаторов.** Для обеспечения требуемой структуры разреженности матриц $\Phi$ и $\Theta$ с предметными и фоновыми темами предлагается комбинация из пяти регуляризаторов: сглаживание фоновых тем в матрицах $\Phi$ и $\Theta$, разреживание предметных тем в матрицах $\Phi$ и $\Theta$, и декоррелирование предметных тем в матрице $\Phi$:

$$\begin{aligned} R(\Phi,\Theta) = &- \beta_0 \sum_{t\in S}\sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in S} \alpha_t \ln \theta_{td} \\ &+ \beta_1 \sum_{t\in B}\sum_{w\in W} \beta_w \ln \phi_{wt} + \alpha_1 \sum_{d\in D}\sum_{t\in B} \alpha_t \ln \theta_{td} \\ &- \gamma \sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W} \phi_{wt}\phi_{ws} \to \max. \end{aligned}$$

где в качестве фонового распределения $\beta$ можно брать либо равномерное распределение, либо частоты слов в коллекции $\beta_w = n_w/n$; в качестве $\alpha$ естественно использовать равномерное распределение. Формулы М-шага для комбинированной модели выписываются согласно (4.4), (4.5):

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big( n_{wt} - \beta_0 \underbrace{\beta_w[t\in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} + \beta_1 \underbrace{\beta_w[t\in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \underbrace{\gamma\,[t\in S]\,\phi_{wt}\sum_{s\in S\setminus t}\phi_{ws}}_{\text{декоррелирование}} \Big);$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big( n_{td} - \alpha_0 \underbrace{\alpha_t[t\in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} + \alpha_1 \underbrace{\alpha_t[t\in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} \Big).$$

**Траектории регуляризации.** При линейном комбинировании регуляризаторов $R_i$ возникает проблема выбора вектора коэффициентов $\tau = (\tau_i)_{i=1}^r$. Эффективный способ их оптимизации применяется в эластичных сетях (elastic net) для задач регрессии и классификации [102], однако он подходит только для комбинирования $L_1$ и $L_2$-регуляризаторов. В тематическом моделировании разнообразие регуляризаторов гораздо больше. При чрезмерно больших значениях коэффициентов некоторые регуляризаторы могут конфликтовать друг с другом, ухудшать сходимость или приводить к вырождению модели. С другой стороны, при чрезмерно низких значениях коэффициентов регуляризаторы могут

утраичивать свое влияние на модель. В теории решения некорректно поставленных обратных задач [87] известно, что для достижения множества решений коэффициенты регуляризации должны в ходе итераций сходиться к нулю. Однако оптимальный темп этой сходимости существенно зависит от конкретной задачи, и на практике его приходится подбирать экспериментально.

Будем называть *траекторией регуляризатора* функцию его коэффициента регуляризации от номера итерации. Будем подбирать траектории регуляризаторов экспериментальным путём, анализируя их влияние на критерии качества модели в ходе итераций.

**Измерение качества модели.** Поскольку задача построения тематической модели является многокритериальной, то и измерение качества модели должно вестись по совокупности критериев. Не претендуя на полноту, перечислим критерии, которые мы использовали в наших экспериментах.

*Контрольная перплексия* $\mathscr{P}(D', p_D)$ вычисляется по контрольной выборке документов $D'$ для модели $p_D$, построенной по обучающей выборке документов $D$, не пересекающейся с $D'$. В наших экспериментах использовалось случайное разбиение коллекции в пропорции $|D| : |D'| = 9 : 1$. Каждый контрольный документ $d$ разбивался случайным образом на две половины: по первой оценивались параметры $\theta_d$, по второй вычислялась перплексия. Если во второй половине оказывались термины, которых не было в обучающей коллекции $D$, то они игнорировались. Параметры $\phi_t$ оценивались только по обучающей коллекции.

*Разреженность* модели измерялась долей $\mathscr{S}_\Phi$ и $\mathscr{S}_\Theta$ нулевых элементов, соответствующих предметным темам в матрицах $\Phi$ и $\Theta$,

*Доля фоновых слов* во всей коллекции

$$\mathscr{B} = \frac{1}{n} \sum_{d \in D} \sum_{w \in d} \sum_{t \in B} n_{dw} p(t \,|\, d, w)$$

принимает значения от 0 до 1. Значения, близкие к 0, говорят о том, что мо-

дель не способна отделять слова общей лексики от специальной терминологии. Значения, близкие к 1, свидетельствуют о вырождении тематической модели, например, в результате чрезмерного разреживания.

*Интерпретируемость* тематической модели оценивалась несколькими критериями. Определим ядро $W_t$ темы $t$ как множество терминов, которые имеют высокую условную вероятность $p(t\,|\,w) = \phi_{wt}\frac{n_t}{n_w}$ для данной темы:

$$W_t = \big\{w \in W \;\big|\; p(t\,|\,w) > 0.25\big\}.$$

По ядру определим три показателя интерпретируемости темы $t$:

$\mathrm{pur}_t = \sum\limits_{w\in W_t} p(w\,|\,t)$ — чистота темы (чем выше, тем лучше);

$\mathrm{con}_t = \frac{1}{|W_t|}\sum\limits_{w\in W_t} p(t\,|\,w)$ — контрастность темы (чем выше, тем лучше);

$\mathrm{ker}_t = |W_t|$ — размер ядра (ориентировочный оптимум $\frac{|W|}{|T|}$).

*Когерентность темы* $t$ измерялась как средняя *поточечная взаимная информация* по всем парам $k$ наиболее вероятных слов темы $t$ [96]:

$$\mathscr{C}_t^k = \frac{2}{k(k-1)}\sum_{i=1}^{k-1}\sum_{j=i}^{k}\mathrm{PMI}(w_i, w_j),$$

где $w_i$ — $i$-й термин в порядке убывания $\phi_{wt}$. Число $k$ в большинстве работ полагают равным 10. Интересно оценить когерентность более глубоко, поэтому мы вычисляли ещё две оценки когерентности модели: при $k = 100$ и по ядрам тем.

Показатели когерентности, размера ядра, чистоты и контрастности модели определим как средние по всем предметным темам $t \in S$.

**Исходные данные.** Эксперименты проводились на коллекции NIPS, которая содержит $|D| = 1566$ текстов статей научной конференции Neural Information Processing Systems на английском языке. Суммарная длина коллекции $n \approx 2.3 \cdot 10^6$ слов. Объём словаря $|W| \approx 1.3 \cdot 10^4$. Контрольная коллекция $D'$ содержит 174 документа. Предварительная обработка текстов включала приведе-

ние к нижнему регистру, удаление пунктуации, удаление стоп-слов с помощью библиотеки BOW toolkit [103].

**Результаты экспериментов.** Во всех экспериментах фиксировалось число тем $|T| = 100$, из них фоновых тем $|B| = 10$, число итераций 40.

В таблице 4.1 приводятся результаты сравнения тематических моделей. Первые две строки соответствуют стандартным моделям PLSA и LDA, остальные строки — регуляризованным моделям ARTM. Первые три колонки задают комбинации регуляризаторов сглаживания, разреживания и декоррелирования. Остальные колонки соответствуют введённым выше критериям качества.

Для оценивания LDA использовался регуляризованный EM-алгоритм с параметрами сглаживания $\alpha = 0.5,\ \beta = 0.01$, соответствующими симметричному распределению Дирихле.

Для сглаживания фоновых тем использовались равномерные распределения при коэффициентах регуляризации $\alpha = 0.8,\ \beta = 0.1$.

Для разреживания предметных тем в столбцах матрицы $\Phi$ использовалось одно из двух распределений: равномерное $\beta_w = \frac{1}{|W|}$ или фоновое $\beta_w = \frac{n_w}{n}$.

Основной вывод заключается в том, что комбинирование регуляризаторов позволяет улучшить все критерии качества при незначительном ухудшении перплексии. Разреживание обнуляет до 96% элементов матрицы $\Phi$ и до 87% элементов матрицы $\Theta$. Декоррелирование повышает чистоту и когерентность тем. Сглаживание фоновых тем помогает им очистить предметные темы от слов общей лексики. Все эти улучшения сопровождаются незначительной потерей перплексии, что согласуется с наблюдениями и выводами из [101] о том, что модели, имеющие лучшую перплексию, то есть лучше предсказывающие появление слов в документах, часто демонстрируют худшую интерпретируемость латентных тем.

Для мониторинга процесса построения модели и подбора траекторий регуляризации строились графики зависимости показателей качества модели от но-

Таблица 4.1. Сравнение регуляризованных тематических моделей со сглаживанием (Sm), разреживанием (Sp) по равномерному (u) или фоновому (b) распределению и декоррелированием (Dc). Критерии: $\mathscr{P}$ — контрольная перплексия, $\mathscr{B}$ — доля фоновых слов в коллекции, $\mathscr{S}_\Phi$ и $\mathscr{S}_\Theta$ — разреженность матриц $\Phi$ и $\Theta$, con — контрастность, pur — чистота, ker — размер ядра, $\mathscr{C}^{\text{ker}}$ — когерентность ядра, $\mathscr{C}^{10}$ и $\mathscr{C}^{100}$ — когерентность 10 и 100 наиболее вероятных слов. Выделены лучшие значения в каждой колонке.

| Sm | Sp | Dc | $\mathscr{P}$ | $\mathscr{B}$ | $\mathscr{S}_\Phi$ | $\mathscr{S}_\Theta$ | con | pur | ker | $\mathscr{C}^{\text{ker}}$ | $\mathscr{C}^{10}$ | $\mathscr{C}^{100}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| – | – | – | 1923 | 0.00 | 0.000 | 0.000 | 0.43 | 0.14 | 100 | 0.84 | 0.25 | 0.17 |
| + | – | – | 1902 | 0.00 | 0.000 | 0.000 | 0.42 | 0.12 | 82 | 0.93 | 0.26 | 0.17 |
| – | u | – | 2114 | 0.24 | 0.957 | 0.867 | 0.53 | 0.20 | 71 | 0.91 | 0.25 | 0.18 |
| – | b | – | 2507 | 0.51 | 0.957 | 0.867 | 0.46 | 0.56 | 151 | 0.71 | 0.60 | 0.58 |
| – | – | + | 2025 | 0.57 | 0.561 | 0.000 | 0.46 | 0.38 | 109 | 0.82 | 0.94 | 0.56 |
| + | u | – | 1961 | 0.25 | 0.957 | 0.867 | 0.51 | 0.20 | 64 | **0.97** | 0.26 | 0.18 |
| + | b | – | 2025 | 0.49 | 0.957 | 0.867 | 0.45 | 0.52 | 128 | 0.77 | 0.55 | 0.55 |
| + | – | + | 1985 | 0.59 | 0.582 | 0.000 | 0.46 | 0.39 | 97 | 0.87 | 0.93 | 0.57 |
| + | u | + | 2010 | 0.73 | **0.980** | 0.867 | **0.56** | 0.73 | 78 | **0.94** | 0.94 | 0.62 |
| + | b | + | 2026 | **0.80** | **0.979** | 0.867 | 0.52 | **0.89** | 111 | 0.81 | **0.96** | **0.83** |

мера итерации. На рис. 4.1–4.2 модель PLSA без регуляризаторов (серые линии) сравнивается с регуляризованной моделью ARTM (чёрные линии). Критерии качества откладываются на трёх графиках по вертикальным осям: на верхнем графике по левой оси — контрольная перплексия $\mathscr{P}$, по правой оси — разреженности матриц $\mathscr{S}_\Phi$ и $\mathscr{S}_\Theta$ и доля фоновых слов $\mathscr{B}$; на среднем графике по левой оси — размер ядра ker, по правой оси — контрастность con и чистота pur; на нижнем графике по левой оси — когерентности $\mathscr{C}^{\text{ker}}$, $\mathscr{C}^{10}$ и $\mathscr{C}^{100}$. Такие графики дают понимание эффектов каждого регуляризатора в отдельности и в комбинации с остальными.

Интересно отметить, что критерии качества могут существенно изменяться после достижения сходимости правдоподобия модели, то есть при неизменной перплексии или при незначительном её ухудшении.

Рис. 4.1 показывает совокупное влияние разреживания предметных тем (по фоновому распределению $\beta_w$) и сглаживания фоновых тем. В частности, видно, что модель PLSA не разреживает матрицы $\Phi$ и $\Theta$ и даёт очень низкую

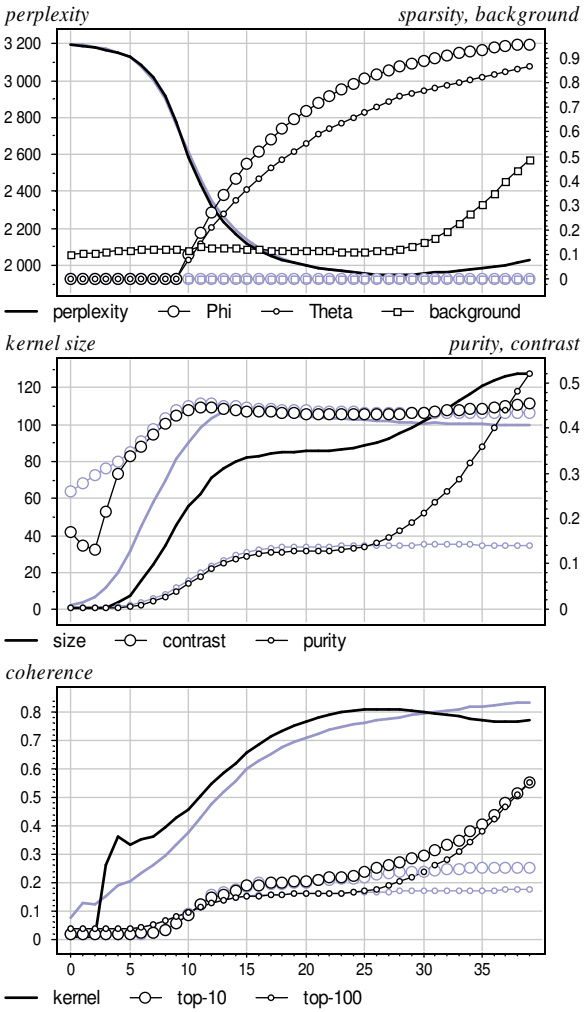



Рис. 4.1. Серый: PLSA. Чёрный: сглаживание, разреживание. Увеличивается разреженность (sparsity) и чистота тем (purity) при небольшом ухудшении перплексии (perplexity).

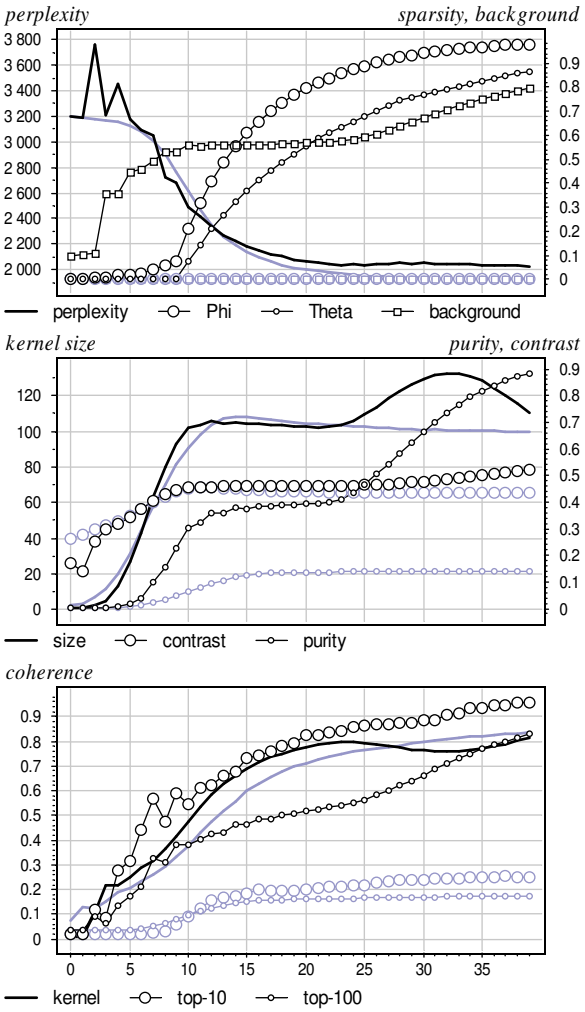



Рис. 4.2. Серый: PLSA. Чёрный: сглаживание, разреживание, декоррелирование. Улучшается когерентность (coherence) по спискам top-10 и top-100 слов в темах, чистота и контрастность.

чистоту тем. Рис. 4.2 позволяет увидеть дополнительные эффекты декоррелирования. В частности, видно, что декоррелирование увеличивает чистоту и когерентность тем, очищает темы от слов общей лексики, при этом доля фоновых слов во всей коллекции достигает почти 80%.

По результатам экспериментов можно дать следующие рекомендации по выбору траекторий регуляризации. Коэффициенты регуляризации для разреживания предметных тем рекомендуется включить только после того, как итерационный процесс начал сходиться и определились близкие к нулю элементы

Таблица 4.2. Сравнение тем в моделях PLSA и ARTM, предметные темы.

| PLSA, тема 50 | ARTM, тема 50 |
|---|---|
| **face** | **face** |
| images | **faces** |
| **faces** | **facial** |
| recognition | **cottrell** |
| set | **pentland** |
| image | **gesture** |
| based | **lane** |
| **hme** | **emotion** |
| **facial** | **person** |
| representation | **steering** |
| view | **appearance** |
| figure | **baluja** |
| model | **setpoint** |
| experts | camera |
| network | tracking |
| human | **pose** |
| expert | **pomerleau** |
| space | **mouth** |
| examples | **darrell** |
| system | **lighting** |

| PLSA, тема 32 | ARTM, тема 32 |
|---|---|
| **query** | **mlp** |
| set | **query** |
| **queries** | **queries** |
| data | **cart** |
| algorithm | **documents** |
| learning | retrieval |
| **documents** | relevant |
| number | **document** |
| performance | rank |
| words | sampling |
| mlp | instances |
| **cart** | **splits** |
| values | collection |
| cluster | gibbs |
| experiments | **sex** |
| results | **ranking** |
| relevant | ordering |
| **retrieval** | recursive |
| classification | text |
| algorithms | axis |

матриц $\Phi$ и $\Theta$. Более раннее или более резкое разреживание может ухудшать перплексию. Мы включали разреживание, начиная с 10-й итерации, обнуляя на каждой итерации 8% ненулевых значений в каждом векторе $\theta_d$ и 10% в каждом векторе $\phi_t$.

Декоррелирование предметных тем включалось с первой итерации, коэффициент регуляризации был выбран постоянным и наибольшим, при котором ещё не происходило существенного увеличения перплексии, для данной коллекции было подобрано значение $\gamma = 2 \cdot 10^5$.

Сглаживание фоновых тем также оказалось лучше включать с первой итерации, не меняя коэффициент регуляризации в ходе итераций.

Таблица 4.3. Сравнение тем в моделях PLSA и ARTM, фоновая тема.

| PLSA, тема 2 | PLSA, тема 55 | ARTM, тема 2 | ARTM, фон |
|---|---|---|---|
| model | **music** | **estimator** | model |
| **prediction** | rules | **music** | data |
| series | note | **musical** | models |
| neural | representation | **notes** | parameters |
| models | neural | mozer | noise |
| data | events | **melody** | neural |
| estimation | net | **composition** | mixture |
| time | set | **bach** | prediction |
| function | time | **chorales** | set |
| method | **musical** | **melodic** | gaussian |
| nonlinear | figure | **jackknife** | likelihood |
| based | network | cooperative | networks |
| point | **notes** | **subnet** | test |
| points | input | **gem** | figure |
| **estimator** | **melody** | **melodies** | training |
| parameters | structure | **icl** | performance |
| error | harmony | **tonal** | network |
| algorithm | **tau** | **accent** | number |
| estimate | **pitch** | augmented | input |
| linear | temporal | piece | results |

**Качественный анализ интерпретируемости тем.** Интересно проанализировать подробнее словарный состав тем. В таблицах 4.2-4.3 представлены по 20 наиболее вероятных слов в некоторых темах базовой модели PLSA и предлагаемой регуляризованной модели. Слова упорядочены по убыванию вероятностей $n_{wt}$. Тематические термины, вошедшие в ядро темы, выделены жирным. В данной коллекции темы интерпретируются как задачи, подходы, методы, отвечающие тематике научной конференции NIPS.

Первое наблюдение состоит в том, что выделенные слова (ядра тем) являются определяющими для темы, в то время как по остальным словам понять тему гораздо труднее. Например, тема 50 посвящена распознаванию лиц. Для модели PLSA в ядре оказываются слова с корнем face (лицо). Остальные слова — recognition (распознавание), representation (представление), figure (рисунок),

model (модель) и т.д. — могут относиться ко многим другим темам конференции NIPS. Сравнение PLSA и регуляризованной модели показывает, что в ARTM слова с корнем face получают наибольшие вероятности. Среди наиболее вероятных слов темы также появляется гораздо больше слов ядра: Cottrell, Pentland (фамилии двух ученых, занимающихся распознаванием лиц), gesture (жестикуляция), lane (морщина), emotion (эмоции) и т.д. Аналогичные выводы можно сделать по теме 32, посвященной задаче ранжирования в информационном поиске, и по большинству других тем.

Модели строились из одного и того же начального приближения, поэтому в большинстве случаев темы с одинаковыми номерами похожи друг на друга. Это позволило избежать сложностей сопоставления тем в двух моделях (например, с помощью венгерского алгоритма). Не совсем так происходит с темой 2. В модели ARTM она посвящена обработке музыкальных сигналов. В модели PLSA тема состоит из слов model (модель), prediction (предсказание), series (серия), neural (нейронный), data (данные) и других слов, нейтральных для коллекции NIPS и не позволяющих интерпретировать тему. Единственная тема в PLSA, содержащая в топе слова music (музыка) или melody (мелодия) – это тема 55. Тем не менее, она снова содержит много общеупотребительных слов коллекции, затрудняющих интерпретацию. Соответствующая тема в модели ARTM свободна от таких слов и целиком состоит из предметных терминов, относящихся к анализу музыки.

В таблице 4.3 представлена одна из фоновых тем, выделенных моделью ARTM. Все фоновые темы содержат термины, широко употребляемые во всей коллекции NIPS. При этом в некоторых фоновых темах доминируют слова, имеющие отношение к классификации, в некоторых – к вероятностным моделям, в некоторых – к нейронным сетям, и т. д.

## 4.3. Автоматический отбор тем

Для постепенного отбора тем предлагается инициализировать модель большим число тем, а затем воздействовать регуляризатором разреживания распределения $p(t) = \sum_d p(d)\theta_{td}$, максимизируя дивергенцию Кульбака-Лейблера между $p(t)$ и равномерным распределением на темах:

$$R(\Theta) = -\tau \frac{n}{|T|} \sum_{t \in T} \ln \sum_{d \in D} p(d)\theta_{td} \;\to\; \max.$$

Встраивание этого регуляризатора в модель согласно теореме 2 приводит к следующим формулам М-шага:

$$\theta_{td} = \underset{t \in T}{\text{norm}}\Big(n_{td} - \tau \frac{n}{|T|} \frac{n_d}{n_t} \theta_{td}\Big).$$

Если заменить $\theta_{td}$ несмещенной оценкой $\frac{n_{td}}{n_d}$, получим разреживание строк матрицы $\Theta$:

$$\theta_{td} = \underset{t \in T}{\text{norm}}\Big(n_{td} - \tau \frac{n_{td}\, n}{|T| n_t}\Big).$$

Если счетчик $n_t$ в знаменателе мал, то все элементы строки получат нулевое значение, а соответствующая тема $t$ будет выведена из модели.

**Эксперименты.** В работе [25] представлены результаты экспериментов на синтетических данных, демонстрирующие корректное восстановление числа тем в данных с помощью предложенного регуляризатора отбора тем. В экспериментах на реальных данных демонстрируется возможность встраивания нового регуляризатора в рассмотренную ранее модель с разреженными и различными предметными темами. На графиках 4.3 и 4.4 представлена зависимость введенных ранее метрик качества от итераций обучения моделей. Для регуляризованной модели осуществляется сглаживание фоновых тем с параметрами $|S| = 10$, $\alpha_t = 0.8$, $\beta_w = 0.1$. Одновременно с этим производится разреживание, декоррелирование и отбор предметных тем. Коэффициент декоррелирования растет линейно в течение первых 60 итераций до максимального значения $\gamma = 200000$,

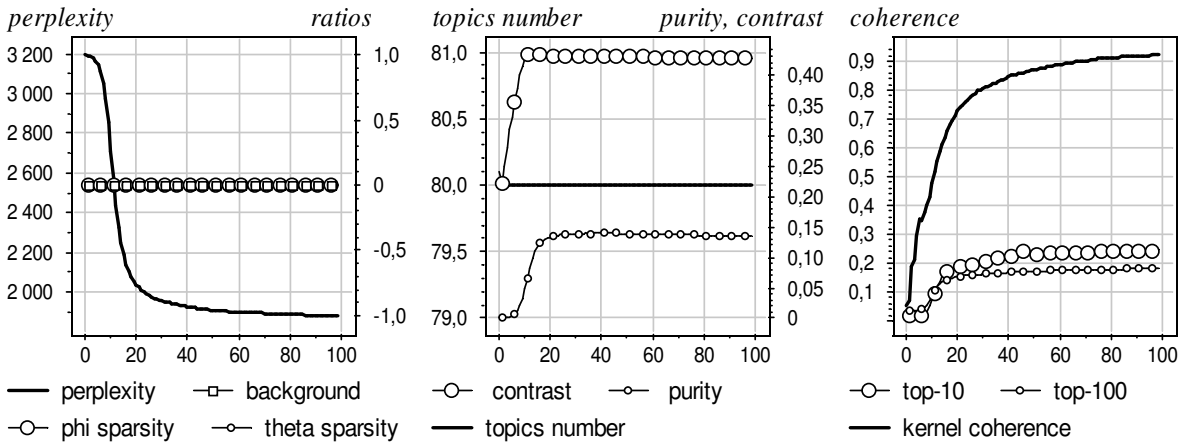



Рис. 4.3. Стандартная тематическая модель LDA.

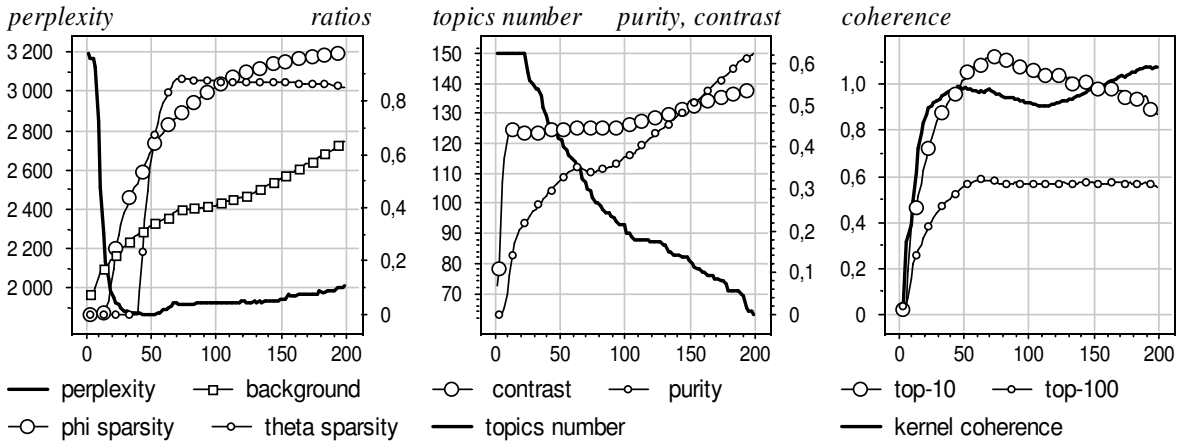



Рис. 4.4. ARTM с регуляризаторами разреживания, декоррелирования и отбора тем.

не разрушающего модель. Отбор тем с коэффициентом $\tau = 0.3$ включается после 15 итераций, чтобы избежать отбора среди почти случайных и похожих тем на ранних итерациях обучения. Отбор тем и декоррелирование применяются через итерацию, т.к. их совместное воздействие может приводить к нежелательным эффектам. На графике отображены замеры качества после итераций декоррелирования. Разреживание подключается с 40 итерации с постепенно возрастающими коэффициентами, так что 2% элементов $\Theta$ и 9% элементов $\Phi$ зануляются на каждой итерации. В результате удается построить последовательность моделей с уменьшающимся числом тем, при этом предметные темы сильно разреженны и интерпретируемы, что видно из сравнения показателей чистоты, контрастности и когерентности с показателями модели LDA.

## 4.4. Обсуждение и выводы

Развивается полу-вероятностный подход к моделированию тематики текстовых коллекций — *аддитивная регуляризация тематических моделей* [21, 24], основанный на максимизации взвешенной суммы критериев регуляризации. Построение тематической модели рассматривается как задача многокритериальной оптимизации, которая сводится к однокритериальной задаче путём скаляризации критериев. Для решения оптимизационной задачи используется регуляризованный EM-алгоритм, в который можно подставлять любые регуляризаторы или их линейные комбинации.

Формализуется понятие интерпретируемости тематической модели и предлагается модель фоновых и предметных тем [23]. В данной модели комбинируются регуляризаторы разреживания, сглаживания и деккоррелирования, что приводит к очищению предметных тем от слов общей лексики и повышению их различности. Демонстрируется улучшение набора критериев качества модели (когерентность тем, чистота и контрастность ядер тем, разреженность) при несущественном ухудшении перплексии модели.

Предлагается разреживающий регуляризатор отбора незначимых тем [25]. Его работа исследуется в комбинации с регуляризаторами сглаживания, разреживания и декоррелирования. В рамках одного цикла обучения строится набор моделей с постепенно сокращающимся числом тем и повышающейся разреженностью распределений. Выбор модели может быть обусловлен внешними критериями качества, описывающими требования конечного приложения.

# Глава 5

# Тематические векторные модели семантики

В главе 1 были рассмотрены модели, решающие задачу определения семантической близости слов. Тематическое моделирование обычно считается непригодным для данной цели. Однако в литературе эксперименты чаще всего проводятся только с тематической моделью LDA, которая действительно не позволяет обучить качественные векторные представления слов.

В данной главе с помощью тематического моделирования строятся векторные представления слов, которые решают задачу определения семантической близости на уровне модели SGNS, ставшей стандартным выбором для этой задачи. Кроме того, удается добиться интерпретируемости и разреженности, что невозможно в большинстве других моделей. Слова с максимальной вероятностью внутри каждой компоненты объединяются в темы, которые человек может проинтерпретировать и поименовать. С помощью аддитивной регуляризации тематической модели достигается дополнительная разреженность векторных представлений.

Построение тематических представлений отдельных слов обобщается на произвольные сегменты текста, в частности, предложения и документы. Кроме того, учитываются мета-данные, связанные с документами, такие как автор, категория, временная метка и другие. Предлагается алгоритм построения единого векторного пространства, где расстояния между данными различной природы хорошо интерпретируются. Например, ближайшими к временной метке «9 мая» оказываются слова военной тематики.

Возникает возможность вводить в модель дополнительные требования и строить векторные представления текста, удовлетворяющие специфичным требованиям прикладных задач.

## 5.1. Тематические векторные представления слов

Согласно дистрибутивной гипотезе смысл слова определяется распределением над множеством слов, совместно встречающихся с ним в локальных контекстах (на практике, в скользящем окне фиксированной ширины). Эта гипотеза противоречит гипотезе о представлении документа в виде «мешка слов», широко используемой в вероятностном тематическом моделировании. Таким образом, в стандартных моделях информация о локальных со-встречаемостях слов теряется, что приводит к низкому качеству предсказания семантической близости слов.

Для формализации дистрибутивной гипотезы в рамках вероятностного тематического моделирования будем для каждого слова $w_i$ в корпусе текстов предсказывать слова $w_j$ из локальной окрестности $H_i$ с помощью смеси тем:

$$p(w_j|w_i) = \sum_{t\in T} p(w_j|t)p(t|w_i) = \sum_{t\in T} \phi_{w_j t}\theta_{t w_i}. \tag{5.1}$$

Сделаем предположения о независимости слов внутри каждой окрестности, а также о независимости окрестностей. Тогда можно записать следующую задачу максимизации правдоподобия по корпусу:

$$\sum_{i=1}^{N}\sum_{j\in H_i} \ln \sum_{t\in T} \phi_{w_j t}\theta_{t w_i} + R(\Phi,\Theta) \to \max_{\Phi,\Theta} \tag{5.2}$$

$$\forall u, t \quad \phi_{ut} \geqslant 0; \quad \sum_{u\in W} \phi_{ut} = 1; \tag{5.3}$$

$$\forall t, v \quad \theta_{tv} \geqslant 0; \quad \sum_{t\in T} \theta_{tv} = 1. \tag{5.4}$$

Обучение модели возможно с помощью итераций EM-алгоритма. Обозначим через $n_{vu}$ агрегированный счетчик совместной встречаемости слов в локальных окрестностях $H_i$, $i = 1, \ldots, N$. Докажем следующую теорему.

**Теорема 3.** *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Тогда точка $(\Phi,\Theta)$ локального экстремума задачи (5.2)-(5.4) удовлетворяет системе*

*уравнений со вспомогательными переменными* $p_{tvu} = p(t|v,u)$*:*

$$\begin{cases} p_{tvu} = \dfrac{\phi_{ut}\theta_{tv}}{\sum\limits_{s\in T}\phi_{us}\theta_{sv}}; \\ \phi_{ut} = \underset{u\in W}{\operatorname{norm}}\left(n_{ut} + \phi_{ut}\dfrac{\partial R}{\partial \phi_{ut}}\right); \quad n_{ut} = \sum\limits_{v\in W} n_{vu}p_{tvu}; \\ \theta_{tv} = \underset{t\in T}{\operatorname{norm}}\left(n_{tv} + \theta_{tv}\dfrac{\partial R}{\partial \theta_{tv}}\right); \quad n_{tv} = \sum\limits_{u\in W} n_{vu}p_{tvu}, \end{cases}$$

*за исключением нулевых столбцов* $\Phi$, $\Theta$ *в решении данной системы.*

*Доказательство.* Введем понятие *контейнера*, порожденного словом $v$:

$$C(v) = \cup_{i:w_i=v} H_i, \quad \forall v \in W, \tag{5.5}$$

где объединение понимается в смысле мульти-множества, индексация $i$ сквозная по корпусу, $H_i$ содержит слова из окна некоторой ширины для позиции $i$. Таким образом, каждый контейнер $C(v)$ является мешком слов, составленным из локальных окрестностей всех вхождений слова $v$ в корпус, причем частота слов $u \in W$ определяется счетчиками $n_{vu}$.

Перегруппируем слагаемые в логарифме правдоподобия:

$$\sum_{i=1}^{N}\sum_{j\in H_i} \ln p(w_j \,|\, w_i) = \sum_{v\in W}\sum_{u\in W} n_{vu} \ln p(u \,|\, v). \tag{5.6}$$

Заметим, что выражение (5.6) совпадает с правдоподобием модели PLSA, примененной к описанному *корпусу контейнеров*. Таким образом, для доказательства теоремы достаточно повторить рассуждения из теоремы 2, где вместо документов рассматривать контейнеры. □

Полученная модель названа вероятностными тематическими представлениями слов *(Probabilistic Word Embeddings, PWE)* и впервые предложена автором данной работы в [26].

Заметим, что модель PWE похожа на тематическую модель коротких текстов WNTM, описанную в главе 1. Однако в модели WNTM используются априорные распределения Дирихле на параметры, а обучение производится с помощью сэмплирования Гиббса. Кроме того, модель WNTM не рассматривалась

как способ обучения векторных представлений слов и не исследовалась в задачах определения семантической близости. Также не устанавливалась ее связь с моделью Skip-Gram, изложенная в следующем параграфе.

**Связь с другими моделями.** Предлагаемая модель PWE использует для обучения ту же информацию о локальной со-встречаемости слов, что и многие векторные модели семантики, в частности, модель Skip-Gram, которая уже обсуждалась в главе 1. Так, правдоподобие в обеих моделях имеет вид (5.6). Ключевое и единственное различие между моделями Skip-Gram и PWE заключается в способе параметризации отдельных вероятностей.

Параметры тематической модели PWE удовлетворяют ограничениям неотрицательности и нормировки, т.е. являются вероятностными распределениями. Поэтому каждая вероятность слова в контексте представляется *вероятностной смесью распределений*:

$$p(u|v) = \langle \phi_u, \theta_v \rangle \,, \tag{5.7}$$

где $\phi_u = (\phi_{ut})_{t \in T}$ — вектор слова $u$, $\theta_v = (\theta_{tv})_{t \in T}$ — вектор контекста $v$.

Модель Skip-Gram, напротив, обучается без ограничений на параметры, и каждая вероятность моделируется с помощью функции softmax:

$$p(u|v) = \operatorname{softmax} \langle \phi_u, \theta_v \rangle = \frac{\exp \langle \phi_u, \theta_v \rangle}{\sum_{w \in W} \exp \langle \phi_w, \theta_v \rangle}. \tag{5.8}$$

Подсчет знаменателя, т.е. нормировка распределения, несет вычислительные трудности, поэтому на практике ее избегают.

Таблица 5.1 подытоживает сравнение моделей. В ней логичным образом появляется еще одна модель, известная как модель doc2vec, — расширение модели word2vec для документов. В терминологии оригинальной статьи это архитектура DBOW модели paragraph2vec [1]. Как и модель PLSA, она моделирует вероятности слов в документах и использует счетчики встречаемости слов в документах при обучении. Как и модель Skip-Gram, каждая вероятность моде-

| | *слова-слова* | *слова-документы* |
|---|---|---|
| *softmax* | word2vec (Skip-Gram) | doc2vec (DBOW) |
| *смесь распределений* | PWE | PLSA |

Таблица 5.1. Сопоставление походов по типу данных (слова-слова или слова-документы) и типу вероятностной модели (softmax или вероятностная смесь распределений).

лируется через softmax:

$$p(w|d) = \text{softmax}\,\langle \phi_w, \theta_d \rangle\,. \tag{5.9}$$

Модель параметризована матрицами $\Phi^{W\times T}$ и $\Theta^{T\times D}$, на которые не накладывается никаких дополнительных ограничений.

Итак, существует два независимых аспекта моделирования:

1) тип со-встречаемостей: «слова-слова» или «слова-документы»;

2) параметризация вероятностей: смесь распределений или softmax.

Опция моделирования со-встречаемостей «слова-слова» с помощью вероятностной смеси распределений остается в литературе мало исследованной и исследуется нами далее в этой главе.

**EM-алгоритм для модели PWE.** Предложенная модель вероятностных тематических представлений слов может быть обучена стандартной схемой EM-алгоритма 1. На вход поступает коллекция контейнеров $C(v),\ v \in W$. На выходе строятся матрицы слов $\Phi$ и контейнеров (контекстов) $\Theta$. Недостатком данного алгоритма является необходимость хранить матрицу $\Theta$, а также долгая сходимость. Это связано с тем, что обновления параметров производятся раз в коллекцию, таким образом необходимы десятки эпох (проходов по коллекции).

Альтернативой может быть онлайновая версия EM-алгоритма, предложенная в [104] для модели LDA. Ее особенность заключается во внутреннем цикле по каждому документу до сходимости $\theta_d$, что позволяет избежать хранения матрицы $\Theta$. В данной работе мы предлагаем онлайновый алгоритм обучения

**Алгоритм 5** Online EM-алгоритм для модели PWE.

**Вход:** коллекция контейнеров $C(v)$, $\forall v \in W$; число тем $|T|$; коэффициент $\rho$;

**Выход:** векторные представления слов $\Phi^B$;

**повторять**
  инициализировать матрицу $\Phi$ случайно;
  обнулить $n_{ut} := 0$, $n_{ut}^{batch} := 0$, $\forall u \in W$, $\forall t \in T$;
  **для всех** $v \in W$
    инициализировать вектор $\theta_v$ случайно;
    обнулить $n_{tv} := 0$, $\forall t \in T$;
    **повторять**
      **для всех** $u \in C(v)$
        $p_{tvu} := \underset{t \in T}{\text{norm}}(\phi_{ut}\theta_{tv})$, $\forall t \in T$;
        накопить $n_{tv} := n_{tv} + n_{vu}p_{tvu}$, $\forall t \in T$;
      обновить $\theta_{tv} := \underset{t \in T}{\text{norm}}\left(n_{tv} + \theta_{tv}\frac{\partial R}{\partial \theta_{td}}\right)$;
    **пока** вектор контекста $\theta_v$ не стабилизируется;
    накопить $n_{ut}^{batch} := n_{ut}^{batch} + n_{vu}p_{tvu}$, $\forall u \in C(v)$, $\forall t \in T$;
    **если** обработан пакет контейнеров, **то**
      обновить $n_{ut} := \rho\, n_{ut} + n_{ut}^{batch}$; $\phi_{ut} := \underset{u \in W}{\text{norm}}\left(n_{ut} + \phi_{ut}\frac{\partial R}{\partial \phi_{ut}}\right)$; $n_{ut}^{batch} := 0$;
**пока** $\Phi$ не стабилизируются;
**вернуть** $\Phi^B = (\phi_{tu}^B)$, где $\phi_{tu}^B := \underset{t \in T}{\text{norm}}\left(n_{ut} + \phi_{ut}\frac{\partial R}{\partial \phi_{ut}}\right)$.

модели PWE (см. Алгоритм 5). Коллекция делится на пакеты, и обновления происходят после обработки каждого пакета. Коэффициент $\rho$ усреднения глобальных счетчиков $n_{ut}$ и счетчиков последнего пакета является гиперпараметром и подбирается в экспериментах. Как правило, для сходимости $\Phi$ достаточно от 1 до 3 проходов по коллекции в зависимости от размеров коллекции. Алгоритм возвращает векторные представления слов $\phi_u^B = (\phi_{tu}^B)_{t \in T}$, $\phi_{tu}^B = p(t|u)$, которые являются вероятностными распределениям на множестве тем $T$. Заметим, что такие вектора можно получить из матрицы $\Phi$, применив форму-

лу Байеса. Для получения оценок близости слов в экспериментах следующего раздела сравнивается несколько способов, из которых наилучшим оказывается скалярное произведение обученных векторов.

## 5.2. Задачи семантической близости и аналогий слов

**Оценивание качества на задаче близости.** Основное свойство, полезное для приложений, – близость похожих слов в векторном пространстве представлений. Как уже отмечалось в главе 1, существуют различные подходы к тому, какие слова считать похожими, и это сильно затрудняет оценивание качества моделей. Тем не менее, известно несколько стандартных датасетов, на которых сравнивается качество векторных представлений слов. Каждый такой датасет состоит из списка пар слов с экспертными оценками близости по некоторой шкале. Модель считается хорошей, если она присваивает парам слов такие оценки близости, что ранжирование списка получается схожим с экспертным. Чтобы это проверить, как правило, подсчитывается корреляция Спирмена. Среди стандартных датасетов можно назвать WordSim353 [105], разделенный на семантически *близкие* и *связанные* пары слов [38], MEN [106], SimLex-999 [107] и Mechanical Turk [108].

В дальнейших экспериментах предлагаемая модель сравнивается с моделью SGNS, для которой оценки близости слов находятся с помощью косинусного расстояния в построенном векторном пространстве. Стандартное решение в случае тематического моделирования — это получение векторов $p(t|w)$ с помощью модели LDA и подсчет расстояния Хеллингера между ними [6]. Как видно из таблицы 5.2, такой подход существенно проигрывает модели SGNS. Наше дальнейшее исследование нацелено на преодоление этого различия в качестве.

**Детали предобработки данных.** В экспериментах использовалась версия англоязычной Википедии 2016-01-13. Предобработка проводилась скриптами

из [12], чтобы гарантировать одинаковые условия для обучения SGNS и тематических представлений слов. Была проведена фильтрация словаря: удалены 25 наиболее популярных слов, сохранены следующие 100000, удалены пары слов, которые встретились совместно менее 5 раз. Эксперименты проводились для ширины окна 2, 5 и 10. Здесь приводятся результаты только для ширины окна 5, так как выводы для других значений аналогичны. Для всех моделей использовался прием выравнивания частотности слов, активно использующийся при обучении модели SGNS, однако не применявшийся ранее для тематического моделирования. Он заключается в исключении из корпуса слововхождений (subsampling) с вероятностью:

$$p_{\text{remove}}(u) = 1 - \sqrt{\frac{\tau}{p(u)}}. \tag{5.10}$$

Коэффициент $\tau = 10^{-5}$ был выбран согласно стандартным рекомендациям. По нашим наблюдениям этот прием незначительно улучшает интерпретируемость тем за счет их очищения от частотных слов общей лексики.

Взвешивание слов внутри скользящего окна согласно их расстоянию до целевого слова не дает существенных преимуществ.

**Задача близости.** Результаты проведенных экспериментов представлены в таблице 5.2. Предлагаемая модель представлений PWE существенно превосходит качество тематической модели LDA и выходит на сопоставимое качество с моделью SGNS, которая является одним из стандартных методов решения задачи близости слов.

По сравнению с моделью LDA первого существенного прироста качества удается добиться, если перейти от моделирования частот $n_{wd}$ слов в документах к частотам $n_{uv}$ со-встречаемости слов, как было описано в предыдущем разделе.

Следующим важным вопросом становится способ подсчета близости векторов в пространстве вероятностных представлений слов. Были протестированы следующие варианты: расстояние Хеллингера, дивергенция Кульбака-Лей-

Таблица 5.2. Корреляция Спирмена на задачах близости слов. Варьируются следующие характеристики. Раскладываемые статистики (Data): $n_{wd}$ — встречаемости слов в документах, $n_{uv}$ — со-встречаемости слов, $n_{uv}/n_v$ — нормированные со-встречаемости, s/pPMI — shifted/positive PMI. Схемы итераций (Optimization): offline — 30 оффлайновых итераций без хранения $\Theta$, mixed — 2 онлайновых и 30 оффлайновых итераций, online — 6 итераций. Расстояния в пространстве ембедингов (Metric): cos - косинусное расстояние, hel — расстояние Хеллингера, dot — скалярное произведение.

| Model | Data | Optimization | Metric | WordSim Sim. | WordSim Rel. | WordSim | Bruni MEN | SimLex-999 |
|---|---|---|---|---|---|---|---|---|
| SGNS | *sPMI* | SGD | cos | **0.752** | **0.632** | **0.666** | **0.745** | **0.384** |
| LDA | $n_{wd}$ | online EM | hel | 0.530 | 0.455 | 0.474 | 0.583 | 0.220 |
| LDA | $n_{wd}$ | online EM | dot | 0.580 | 0.516 | 0.532 | 0.599 | 0.230 |
| PWE | $n_{uv}$ | offline EM | dot | 0.709 | 0.635 | 0.654 | 0.658 | 0.240 |
| PWE | $n_{uv}/n_v$ | offline EM | dot | 0.642 | 0.580 | 0.576 | 0.679 | 0.262 |
| PWE | *pPMI* | offline EM | dot | 0.701 | 0.615 | 0.647 | **0.707** | **0.276** |
| PWE | $n_{uv}$ | mixed EM | dot | **0.723** | **0.675** | 0.682 | 0.672 | 0.263 |
| PWE | $n_{uv}$ | online EM | dot | 0.718 | 0.673 | **0.685** | 0.669 | 0.263 |

блера, косинусное расстояние, скалярное произведение векторов. Несмотря на интуитивное предположение о том, что в данной задаче подходят расстояния, предназначенные для вероятностных распределений, наилучшее качество показало скалярное произведение. Вероятностные представления слов при этом были получены из матрицы параметров $\Phi$ по правилу Байеса.

В следующем эксперименте исследовались различные схемы итерационного процесса. Был предложен онлайновый EM-алгоритм для модели PWE, аналогичный алгоритму online-LDA [109], в котором матрица $\Phi$ обновляется инкрементально, а матрица $\Theta$ не хранится и каждый раз инициализируется случайно. Детали его реализации представлены в алгоритме 5. Он превзошел по качеству стандартный оффлайновый алгоритм 1 для коллекции контейнеров $C(v)$, а также оффлайновый алгоритм без хранения матрицы $\Theta$. Дальнейшего прироста качества удалось добиться при комбинированном подходе (*mixed* в таблице 5.2), где сначала осуществляется несколько онлайновых, а затем несколько

оффлайновых проходов.

Помимо этого, были исследованы различные оценки со-встречаемости, например, нормированные счетчики $n_{uv}/n_v$. Заметим, что это эквивалентно матричному разложению эмпирических вероятностей $p(u|v)$ по *невзвешенной* сумме дивергенций Кульбака-Лейблера. На качестве векторных представлений такая эвристика сказалась отрицательно. Прироста на некоторых тестовых выборках удалось добиться при разложении pPMI-матрицы.

**Задача аналогий.** Задача аналогий или *relational similarity* уже обсуждалась в главе 1. Утверждается, что для пар отношений слов $a : a'$ и $b : b'$ справедливо $b - a + a' \approx b'$ в векторном пространстве. Например, если представить слова «король», «королева», «мужчина», «женщина» векторами модели SGNS, то ближайший вектор для арифметического выражения «король - мужчина + женщина» будет соответствовать слову «королева». Именно благодаря этому примеру модель word2vec стала широко известна. Однако стоит сделать несколько оговорок.

В примере выше ближайшим словом, на самом деле, является «король». В оригинальной статье [1] из поиска удалялись три слова, по которым получен вектор. В более поздней статье [110] показано, что такой эффект затрагивает до 90% примеров (в зависимости от категории аналогий). При этом после удаления исходных векторов из поиска модель, действительно, достигает точности 0.7 на задаче предсказания четвертого слова в примере. Это явление требует объяснений.

В [110] показано, что точность модели высока лишь на тех примерах, у которых векторы для слов $b$ и $b'$ близки в пространстве. Это приводит к предположению о том, что оригинальные выборки четверок слов для тестирования моделей не покрывают все разнообразие аналогий, а содержат примеры, простые для векторных моделей семантики.

Семантический датасет [56], на котором производилось тестирование моде-

ли word2vec содержит лишь несколько категорий аналогий: страны и столицы, страны и валюты, профессии и пол. Чтобы превзойти ограничения исходного датасета, был предложен более полный датасет BATS (The Bigger Analogies Test Set) [111], разделенный на множество категорий. Качество задачи аналогий на нем оказалось существенно ниже.

Авторы [112] далее сопоставляют вычитание и сложение векторов word2vec с простым подходом, который находит ближайшего соседа для вектора одного из исходных слов. Этот подход проигрывает на специфичных категориях аналогий из [56], однако не уступает в качестве на расширенном наборе аналогий BATS. Таким образом, свойство контролируемого изменения смысла слова при произведении арифметических операций в построенном векторном пространстве оказывается не столь важным.

В связи с отсутствием приемлемой методологии оценки качества, а также неясной практической значимостью задачи аналогий, мы не включаем в данную работу сравнение методов по этому критерию.

## 5.3. Интерпретируемость и разреженность компонент

В предыдущем разделе было показано, что предложенный алгоритм построения вероятностных тематических представлений слов позволяет решать задачу близости. Рассмотрим дополнительные свойства полученных представлений, которыми не обладают стандартные SGNS вектора.

**Интерпретируемость компонент.** Важным вопросом при обучении векторных представлений слов становится интрерпретируемость. Можно ли говорить, что компоненты вектора соответствуют «атомам смысла», т.е. каким-либо семантическим признакам? Правда ли, что есть компонента, по которой все слова, относящиеся к мужскому полу, имеют большое значение, а все слова, относящиеся к женскому, — маленькое?

Таблица 5.3. Интерпретируемость компонент PWE и SGNS векторов.

| PWE | | | SGNS | | |
|---|---|---|---|---|---|
| art | arbitration | game | avg | transports | rana |
| painting | ban | games | hearth | recon | walnut |
| museum | requests | video | soc | grumman | rashid |
| painters | arbitrators | gameplay | protector | convoys | malek |
| gallery | noticeboard | multiplayer | decomposition | piloted | aziz |
| sculpture | block | puzzle | whip | stealth | khalid |
| painter | administrators | mario | stochastic | flotilla | yemeni |
| exhibition | arbcom | nintendo | sewer | convoy | andalusian |
| portraits | sanctions | player | splinter | supersonic | bien |
| drawings | mediation | gaming | accessory | bomber | gcc |

Классические SGNS вектора таким свойством не обладают. Это видно из таблицы 5.3. Каждая колонка соответствует некоторой компоненте. Выведены 10 слов с наибольшим значением. При этом неясно, по какому семантическому признаку они объединены. Существуют подходы, позволяющие повышать интерпретируемость части компонент векторов SGNS с помощью ортогональных преобразований матриц [113].

Тематические модели сразу же выделяют *темы* как семантически связанные группы слов. Поскольку в модели PWE именно темы являются компонентами векторов, то компоненты также приобретают интерпретируемость, и им можно дать названия. Например, в таблице 5.3 первая тема о художниках, вторая о юристах, третья о видео-играх.

Визуальное сопоставление списков слов субъективно и не раз подвергалось критике в литературе. Возможная альтернатива — это привлечение асессоров для оценивания по сценарию *word intrusion (найди лишнее слово)*. В такой постановке в список наиболее вероятных слов темы примешивают одно случайное слово. Считается, что чем точнее эксперты находят это слово, тем более интерпретируемой была исходная тема. Мы воспользуемся автоматически вычисляемой мерой интерпретируемости — *когерентностью*. В [114, 115] было показано, что когерентность лучше других альтернативных мер интерпретируемости кор-

релирует с экспертными оценками, при этом не требует привлечения асессоров. Согласно [116] когерентностью темы будем считать усредненное значение PMI по всем парам топ-слов темы:

$$\mathcal{C}(t) = \frac{2}{k(k-1)} \sum_{(w_i,w_j)\in \mathrm{Top}_k} \mathrm{PMI}(w_i, w_j), \tag{5.11}$$

где $\mathrm{Top}_k$ – это множество $k$ наиболее вероятных слов темы (компоненты). Когерентность модели получается усреднением когерентностей тем. В литературе предлагались и другие оценки когерентности, например, в [115] вместо PMI используются логарифмы вероятностей, а в [117] более сложные оценки дистрибутивной семантики. Многие подходы сопоставлены в обзоре [118]. Несмотря на свою популярность в тематическом моделировании, когерентность почти не используется в литературе по векторным представлениям слов. Тем не менее, она использовалась авторами модели неотрицательных разреженных представлений слов (Non-Negative Sparse Embeddings, NNSE) [6], а также авторами модели онлайновых интерпретируемых представлений слов (Online Interpretable Word Embeddings, OIWE) [119]. Мы также будем оценивать интерпретируемость векторных представлений именно по этой методике.

Замеры когерентности для моделей представлены на Рисунке 5.1. Для предлагаемой тематической модели PWE и стандартной тематической модели LDA сортировка слов в компоненте осуществляется по вероятностям $p(w|t)$. Для модели SGNS были протестированы два различных подхода: сортировка по исходным значениям и сортировка по вероятностям. Для получения вероятностей мы применяем softmax к каждой вектор-строке матрицы $\Phi$, а затем по правилу Байеса переходим от вероятностей $p(t|w)$ к вероятностям $p(w|t)$. Первый вариант исходных значений можно также трактовать как softmax, примененный к каждому вектор-столбцу исходной матрицы.

По графику видно, что модель PWE превосходит по когерентности стандартную тематическую модель LDA и модель SGNS.

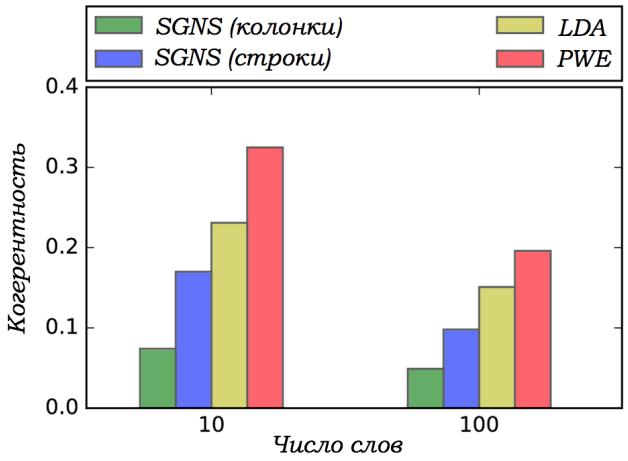



Рис. 5.1. Количественная оценка интерпретируемости: PMI-когерентность по спискам топ-10 и топ-100 слов в компонентах. Предлагаемая модель PWE превосходит другие подходы.

**Разреженность компонент.** Построение векторных представлений слов с помощью тематического моделирования позволяет подключить разрабатываемый подход аддитивной регуляризации и учесть дополнительные требования к модели. В частности, чтобы увеличить разреженность векторных представлений, к модели (5.2) был добавлен разреживающий регуляризатор (4.10), основанный на максимизации дивергенции Кульбака-Лейблера между искомыми тематическими распределениями и равномерным распределением. В результате эксперимента были получены вектора, содержащие **93%** нулей и показавшие такое же качество на задачах близости, как и неразреженные вектора (см. онлайновый EM-алгоритм в таблице 5.2).

## 5.4. Векторные представления мультимодальных данных

Часто вместе с документами известна дополнительная информация, например, авторы, даты, категории и т.д. Учет такой информации может улучшать векторное описание слов и документов, а также расширять область применимости методов. В частности, в рекомендательных системах возникает необходимость оценки семантической близости между товарами и покупателями; в рекламе — между объявлениями и пользователями; в анализе транзакционных

данных — между клиентами и категориями покупок. При этом, как правило, доступны большие объемы текстовых данных, описывающие каждую из взаимодействующих сущностей.

Такого рода данные могут быть включены в векторные модели семантики с помощью подхода аддитивной регуляризации тематических моделей. Будем называть типы доступной мета-информации дополнительными *модальностями*, при этом базовой модальностью будем считать текст. Далее мы предлагаем алгоритм построения единого векторного пространства для элементов (токенов) всех модальностей.

Пусть дана коллекция, где для каждого документа $d$ дополнительно известны токены модальностей $m \in M$. Пусть $m = 0$ соответствует базовой модальности слов. Напомним, что для построения вероятностных представлений PWE составлялась коллекция контейнеров, где каждый контейнер порождался некоторым словом-контекстом $v \in W^0$ и содержал объединение слов из окрестностей всех вхождений $v$ в корпус. Другими словами, учитывались счетчики $n_{vu}$ *локальной со-встречаемости* слов $u \in W^0$ и $v \in W^0$ в скользящем окне фиксированной ширины.

Пополним контейнер $v$ токеном $w$ новой модальности $m$, если он приписан какому-либо документу, где встречается слово $v$. Более точно, для $v \in W^0$ и $w \in W^m$, $m \neq 0$, подсчитаем *документную со-встречаемость*:

$$n_{vw} = \sum_{d:w \in d} n_{dv},$$

где $n_{dv}$ — частота слова $v$ в документе $d$.

Таким образом, взаимодействия слов, как и прежде, определяются счетчиками локальной со-встречаемости, а взаимодействия слов с токенами других модальностей определяются счетчиками документной со-встречаемости. Каждый контейнер при этом порождается словом, но содержит токены всех модальностей.

Для обучения модели будем оптимизировать взвешенную сумму слагае-

мых, каждое из которых является правдоподобием, записанным относительно соответствующей модальности. Аналогичный подход был предложен в [120] для включения дополнительных модальностей в модель PLSA.

Рассмотрим задачу:

$$\sum_{m\in M}\lambda_m\sum_{v\in W^0}\sum_{u\in W^m} n_{vu}\ln\sum_{t\in T}\phi_{ut}\theta_{tv} + R(\Phi,\Theta)\to\max_{\Phi,\Theta}; \tag{5.12}$$

$$\forall u,t \quad \phi_{ut}\geqslant 0;\quad \sum_{u\in W^m}\phi_{ut}=1,\ \forall m\in M; \tag{5.13}$$

$$\forall t,v \quad \theta_{tv}\geqslant 0;\quad \sum_{t\in T}\theta_{tv}=1. \tag{5.14}$$

где $\lambda_m > 0$ — веса модальностей, $W^m$ — словари модальностей; $m = 0$ соответствует базовой модальности слов; $n_{vu}$ — локальная со-встречаемость, если $u \in W^0$, и документная со-встречаемость иначе.

В такой модели матрица параметров $\Phi$ разбивается на блоки по словарям различных модальностей, и нормировка производится в рамках каждого отдельного блока. Таким образом, каждая тема описывается несколькими альтернативными распределениями. Матрица $\Theta$ сохраняет прежнюю размерность и интерпретацию. Возможна модель, в которой токены дополнительных модальностей также порождают свои контейнеры. Это восстанавливает симметричность раскладываемой матрицы и расширяет матрицу $\Theta$, которая в таком случае содержит вектор-столбцы для токенов всех модальностей.

Учет дополнительных модальностей не противоречит введению регуляризаторов разреживания и любых других. Обучение по-прежнему производится EM-алгоритмом с регуляризованными формулами M-шага.

**Теорема 4.** *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Тогда точка $(\Phi,\Theta)$ локального экстремума задачи (5.12)-(5.14) удовлетворяет системе*

*уравнений со вспомогательными переменными* $p_{tvu} = p(t|v,u)$*:*

$$p_{tvu} = \underset{t\in T}{\operatorname{norm}}\big(\phi_{ut}\theta_{tv}\big); \tag{5.15}$$

$$\phi_{ut} = \underset{u\in W^m}{\operatorname{norm}}\left(n_{ut} + \phi_{ut}\frac{\partial R}{\partial \phi_{ut}}\right); \quad n_{ut} = \sum_{v\in W^0} n_{vu}p_{tvu}; \tag{5.16}$$

$$\theta_{tv} = \underset{t\in T}{\operatorname{norm}}\left(n_{tv} + \theta_{tv}\frac{\partial R}{\partial \theta_{tv}}\right); \quad n_{tv} = \sum_{m\in M}\sum_{u\in W^m} \lambda_m n_{vu}p_{tvu}. \tag{5.17}$$

*за исключением нулевых столбцов* $\Phi$, $\Theta$ *в решении данной системы.*

*Доказательство.* По параметрам $\Phi$ задача распадается на независимые для каждой модальности. Поэтому вывод формул М-шага (5.16) аналогичен рассмотренному в теореме 2 с учетом того, что вместо документов $d$ используются контейнеры $v$.

Выведем формулы для параметров $\Theta$. Запишем необходимые условия локального экстремума по теореме Каруша–Куна–Таккера:

$$\underbrace{\sum_{m\in M}\lambda_m \sum_{u\in W^m} n_{vu}\frac{\phi_{ut}}{p(u|v)} + \frac{\partial R}{\partial \theta_{tv}}}_{X_{tv}} = \mu_v - \mu_{tv}; \quad \mu_{tv} \geqslant 0; \quad \mu_{tv}\theta_{tv} = 0; \tag{5.18}$$

где множители Лагранжа $\mu_v$ и $\mu_{tv}$ соответствует ограничениям нормировки и неотрицательности соответственно. Домножим обе части равенства на $\theta_{tv}$ и выделим вспомогательные переменные $p_{tvu}$:

$$\theta_{tv}\mu_v = \theta_{tv}X_{tv} = \sum_{m\in M}\lambda_m \sum_{u\in W^m} n_{vu}p_{tvu} + \theta_{tv}\frac{\partial R}{\partial \theta_{tv}} = n_{tv} + \theta_{tv}\frac{\partial R}{\partial \theta_{tv}}. \tag{5.19}$$

Фиксируем некоторый контейнер $v$. Если для всех $t \in T$ значение $X_{tv} \leqslant 0$, то будем считать такой контейнер вырожденным и исключим его из модели, положив $\theta_{tv} = 0$, $\forall t \in T$.

Иначе существует тема $s$ такая, что значение $X_{sv} > 0$. Далее рассмотрим два случая для некоторой темы $t \in T$. Если $X_{tv} \leqslant 0$, то $\mu_{tv} = \mu_v - X_{tv} > 0$, и из условия дополняющей нежесткости $\theta_{tv} = 0$. Если $X_{tv} > 0$, то из (5.19) имеем $\theta_{tv}\mu_v = \theta_{tv}X_{tv}$. Объединяя два случая, запишем:

Таблица 5.4. Корреляция Спирмена на задачах близости. Модели обучена на русскоязычном мультимодальном корпусе Lenta.ru. Учет меток времени и категорий улучшает качество.

| Model | WordSim Sim | WordSim Rel | MC | RG | HJ | SimLex |
|---|---|---|---|---|---|---|
| SGNS | 0.630 | 0.530 | 0.377 | 0.415 | 0.567 | **0.243** |
| CBOW | 0.625 | 0.513 | 0.403 | 0.370 | 0.551 | 0.170 |
| PWE | 0.649 | 0.565 | 0.605 | **0.594** | 0.604 | 0.123 |
| Multi-PWE | **0.682** | **0.58** | **0.607** | 0.584 | **0.611** | 0.144 |

$$\theta_{tv}\mu_v = \max\left(0, n_{vu} + \theta_{tv}\frac{\partial R}{\partial \theta_{tv}}\right). \tag{5.20}$$

Суммируя левую и правую часть по $t$, получаем выражение для $\mu_v$:

$$\mu_v = \sum_{t\in T} \max\left(0, n_{vu} + \theta_{tv}\frac{\partial R}{\partial \theta_{tv}}\right). \tag{5.21}$$

Подставляя его обратно в (5.20), получаем формулу М-шага (5.17). □

**Эксперимент на новостной коллекции Lenta.ru.** В проведенном эксперименте рассматривалась коллекция новостного сайта на русском языке Lenta.ru. Она содержит 100033 новостей общим объемом 10050714 токенов. Дополнительно, известны временные метки документов (825 уникальных), категории (22 уникальных) и под-категории (97 уникальных). Размер словаря – 54963 слов. При предобработке коллекции для обучения вероятностных представлений использовалось окно ширины 5 и прием выравнивания частот слов (5.10).

Для оценивания качества использовался набор данных HJ [121] с экспертными оценками 398 пар слов, которые являются переводами на русский следующих наборов данных: MC [122], RG [123] и WordSim353 [105]. Также мы использовали перевод набора данных SimLex-999 [124].

Таблица 5.4 показывает, что вероятностные представления слов PWE превосходят модель SGNS на большинстве датасетов даже без использования дополнительных модальностей времени и категорий. Возможной причиной низкого качества модели SGNS является размер корпуса. Мы также пробовали

использовать модель CBOW [1], следуя общей рекомендации использовать эту архитектуру для небольших данных, однако ее качество оказалось еще ниже. В этом и других экспериментах мы замечаем, что тематическое моделирование требует данные меньших объемов для получения качественных представлений, чем модели семейства word2vec.

С использованием дополнительных модальностей качество улучшается (см. Multi-PWE в таблице 5.4). Интересно, что метки времени и категорий способствуют более точным оценкам близости слов (базовой модальности). Далее мы рассматриваем два различных режима. В первом модальности используются только как токены (несимметричный случай), а во втором еще и порождают свои контейнеры (симметричный случай). Близости слов оказываются лучше в несимметричном подходе, а близости между токенами различных модальностей выигрывают от симметризации. В таблице 5.5 приведено несколько примеров временных меток и ближайших к ним слов. Результаты очень хорошо интерпретируются: первая колонка соответствует выходу фильма «Звездные войны», вторая – премии Оскар, а третья – Дню Победы.

Таким образом, введение модальностей в рамках аддитивной регуляризации позволяет не только улучшить качество представлений слов, но и получить осмысленные близости между токенами различных модальностей. Результаты данного раздела опубликованы в [26].

## 5.5. О связывании векторов слов и контекстов

**Матрица слов и матрица контекстов.** Интересным наблюдением является то, что во всех рассмотренных моделях квадратная матрица раскладывается в произведение в двух не связанных друг с другом матриц:

$$F^{W\times W} \approx \Phi^{W\times T}\,\Theta^{T\times W}. \tag{5.22}$$

В литературе по векторным представлениям слов принято называть первую

Таблица 5.5. Ближайшие слова к датам в векторном пространстве модели PWE.

| 2015-12-18<br>Премьера Звездных Войн | 2016-02-29<br>Вручение Оскара | 2015-05-09<br>День Победы |
|---|---|---|
| джедай | статуэтка | великий |
| ситх | кинонаграда | годовщина |
| фетт | номинироваться | летие |
| энакин | кинопремия | нормандия |
| чубакка | Линклейтер | парад |
| киносага | Оскар | демонстрация |
| хэмилл | Бёрдмен | шествие |
| кэрри | удостоиться | Владимир |
| приквел | award | празднование |
| соло | критики | концентрационный |
| пробуждение | отрочество | освенцим |
| бойега | оператор | марш |
| трилогия | Любецки | фотопортрет |
| абрамс | режиссёр | труженник |

матрицу матрицей слов, а вторую – матрицей контекстов, подчеркивая различные роли одних и тех же элементов словаря $W$. Возникает вопрос, нельзя ли уменьшить пространство параметров, положив:

$$\Theta = \Phi^T; \quad F \approx \Phi\,\Phi^T. \tag{5.23}$$

С этим вопросом связано несколько аргументов. Во-первых, в ряде задач все же возникает потребность раскладывать неквадратную матрицу $F$, когда множество *слов* и множество *контекстов* не совпадают. Например, в качестве *контекста* может выступать пара (слово, сдвиг относительно другого слова). Так, в работе [125] такой подход показывал высокое качество. Тем не менее, в подавляющем большинстве задач исходная матрица квадратная.

Во-вторых, с лингвистической точки зрения понятия *слова* и *контекста* различны даже для квадратной матрицы. Вектор *слова* $w$ должен отражать особенности употребления самого слова, в то время как вектор *контекста* $w$ должен отражать особенности употребления других слов рядом с ним. Тем не менее, на практике это различие стирается. Так, в моделях GloVe и SGNS успеш-

ной эвристикой оказывается усреднение двух векторов: $\frac{1}{2}(\phi_w+\theta_w)$. Именно такое представление слова дает наилучшее качество в прикладных задачах.

В-третьих, разложение вида (5.23) возможно только для неотрицательно определенной матрицы $F$.

**Неединственность решения.** Важной проблемой задачи матричного разложения является неединственность решения:

$$F \approx \Phi\,\Theta = (\Phi\,S)\,(S^{-1}\,\Theta) = \Phi'\,\Theta', \tag{5.24}$$

для любой невырожденной матрицы $S$. Таким образом, при одном и том же значении оптимизируемого функционала мы можем получать различные матрицы. Особенно важно, что для подсчета близости слов обычно используется скалярное произведение строк матрицы $\Phi$, которое также изменится в случае использования матрицы $\Phi'$.

Рассмотрим теперь случай поиска разложения в виде (5.23):

$$F \approx \Phi\,\Phi^T = (\Phi\,S)\,(S^{-1}\,\Phi^T) = (\Phi\,S)\,(\Phi\,S)^T = \Phi'\,\Phi'^T. \tag{5.25}$$

В этом случае матрица $S$ должна быть ортогональной, а значит, скалярные произведения строк матрицы $\Phi$ не изменятся. Таким образом, связывание параметров обеспечивает единственность разложения в смысле оценок близости слов.

Аналогичное связывание входных и выходных представлений слов применяется в языковых моделях [126, 127]. Однако в стандартных моделях векторных представлений слов (word2vec, GloVe) этого не происходит.

**Связывание параметров в тематических моделях.** Рассмотрим тематическую модель битермов (Biterm Topic Model, BTM) [128]. Она была предложена для моделирования корпусов коротких текстов и превзошла на таких данных стандартный подход с использованием LDA. В модели BTM моделируется совместное распределение слов $u$ и контекстов $v$ как взвешенное скалярное

произведение двух строк матрицы $\Phi$:

$$p(u,v) = \sum_{t \in T} \phi_{ut}\phi_{vt}\pi_t, \tag{5.26}$$

где $\pi$ — вектор, задающий распределение на темах и не зависящий от контекста. Словари слов и контекстов совпадают.

В данной модели, так же как и в модели WNTM, дополнительно используются априорные распределения Дирихле. Далее будем это опускать и рассматривать аналоги обеих моделей без регуляризации.

Введем обозначение для матрицы условных вероятностей, полученных из матрицы $\Phi$ по формуле Байеса:

$$\Phi^B = (\phi^B_{tw}), \quad \phi^B_{tw} = p(t|w) = \frac{p(w|t)p(t)}{p(w)} = \frac{\phi_{wt}p(t)}{p(w)}. \tag{5.27}$$

**Теорема 5.** *Если при инициализации модели WNTM положить $\Theta = \Phi^B$, то данная связь матриц $\Phi$ и $\Theta$ сохраняется в течение EM-итераций, а полученная модификация WNTM в точности совпадает с моделью BTM.*

*Доказательство.* Выпишем формулы EM-алгоритма для модели WNTM:

- **E-шаг:**

$$p(t|v,u) = \frac{\phi_{ut}\theta_{tv}}{\sum_t \phi_{ut}\theta_{tv}}; \tag{5.28}$$

- **M-шаг:**

$$\phi_{ut} = \frac{n_{ut}}{n_t}; \quad n_{ut} = \sum_{v \in W} n_{vu}p(t|v,u); \quad n_t = \sum_{u \in W} n_{ut}; \tag{5.29}$$

$$\theta_{tv} = \frac{n_{tv}}{n_v}; \quad n_{tv} = \sum_{u \in W} n_{vu}p(t|v,u); \quad n_v = \sum_{t \in T} n_{tv}. \tag{5.30}$$

Если функция $p(t|v,u)$ симметрична относительно аргументов $u$ и $v$, то матрицы счетчиков $n_{ut}$ и $n_{tv}$ совпадают с точностью до транспонирования, а матрицы $\Phi$ и $\Theta$ отличаются транспонированием и нормировкой:

$$\theta_{tv} = \frac{n_{tv}}{n_v} = \frac{n_{vt}n_t}{n_t n_v} = \phi_{vt}\frac{n_t}{n_v}. \tag{5.31}$$

Таблица 5.6. Корреляция Спирмена для WNTM и BTM на задачах близости слов.

| Model | WS-353 Sim | WS-353 Rel | WS-353 | SimLex Hill et al. | MEN Bruni et. al | RareWords Luong et al. | Radinsky M. Turk |
|---|---|---|---|---|---|---|---|
| BTM | **0.68** | **0.59** | **0.61** | 0.24 | 0.65 | 0.32 | 0.54 |
| WNTM | 0.67 | 0.58 | 0.60 | 0.24 | **0.66** | **0.33** | **0.55** |

В таком случае очередной итеративный пересчет по формуле (5.28) снова приведет к симметричной $p(t|v,u)$.

Инициализируем матрицу $\Theta$ согласно (5.31). Соотношение (5.31) будет инвариантом итерационного процесса. Рассмотрим, как при этом изменится вероятностная генеративная модель WNTM:

$$p(u|v) = \sum_{t\in T} \phi_{ut}\theta_{tv} = \sum_{t} \phi_{ut}\phi_{vt}\frac{n_t}{n_v}. \tag{5.32}$$

По определению условной вероятности запишем:

$$p(u,v) = p(u|v)p(v) = \sum_{t\in T} \phi_{ut}\phi_{vt}\frac{n_t}{n_v}\frac{n_v}{n} = \sum_{t} \phi_{ut}\phi_{vt}\pi_t; \quad \pi_t = \frac{n_t}{n}. \tag{5.33}$$

Мы пришли в точности к формуле (5.26), определяющей модель BTM. □

**Эксперименты.** Модели WNTM и BTM были реализованы в библиотеке BigARTM [129]. Сравнение проводилось на задаче близости слов по стандартным тестовым наборам. Обучение моделей производилось на англоязычной Википедии. По результатам в Таблице 5.6 видно, что качество моделей практически не отличается. При этом число параметров в модели BTM в два раза меньше, чем в модели WNTM. Это интересный результат, который показывает, что в данном случае выгодно производить связывание параметров модели.

## 5.6. Представления предложений и документов

Помимо представления слов в семантическом пространстве низкой размерности, часто возникает задача представления более длинных фрагментов текста, таких как предложения, параграфы или целые документы. Такие модели

активно развиваются. Часть из них основана на выражении вектора предложения через взвешенное среднее векторов слов. В [130] это происходит в пост-обработке. Выбираются веса, штрафующие слишком частотные слова, в результате взвешенное среднее предобученных векторов слов дает высокое качество. В модели Sent2vec [131] этот этап переносится из пост-обработки в обучение. Максимизируется правдоподобие наблюдаемых слов в предложениях, при этом предложения моделируются усреднением векторов слов. Еще одна известная модель StarSpace [132] обучается по выборке пар семантически близких предложений, при этом каждое предложение по-прежнему раскладывается в сумму слов. Семантически близкими могут считаться последовательные предложения в тексте (обучение без учителя) или предложения-дубликаты согласно разметке асессоров (обучение с учителем).

Подходы другой группы [133–136] работают с предложением целиком, уходя от гипотезы мешка слов. Так, в модели SkipThought [133] предложение моделируется рекуррентной нейронной сетью GRU для предсказания следующего и предыдущего предложений в тексте. Предполагается, что представление нейронной сети, получаемое из последнего состояния кодировщика (encoder), содержит всю необходимую семантическую информацию о предложении. В модели InferSent [135] исследуются несколько различных способов агрегации состояний кодировщика: усреднение векторов, взятие максимума по каждой компоненте, механизм внимания, несколько слоев сверток. Обучение производится по набору пар предложений SNLI [137], размеченных для задачи предсказания логического следствия (entailment). Более полный обзор векторных представлений предложений можно найти в [138], а сравнение качества работы на различных прикладных задачах в [139].

Такие представления, как правило, не являются интерпретируемыми. В нашей работе [29] было показано, что увеличение разреженности представлений может приводить к повышению интерпетируемости. Исследовалась простейшая архитектура автокодировщика (autoencoder), в которой предложение представ-

Таблица 5.7. Качество векторов предложений: корреляция Пирсона/Спирмена на задачах близости STS-2014 и SICK relatedness (связанность); точность на задаче SICK entailment (следование).

| Model | STS-2014 | | | | | SICK | |
|---|---|---|---|---|---|---|---|
| | Forum | News | Headlines | Images | Tweets | Rel | Ent |
| BOW (ours) | 0.41/ 0.42 | 0.70/ 0.62 | 0.60/ 0.53 | 0.76/ 0.71 | 0.68/ 0.63 | 0.77/ 0.70 | 76.27 |
| Fitted (ours) | **0.45**/ **0.46** | **0.70**/ 0.62 | 0.61/ 0.55 | **0.76**/ 0.71 | 0.68/ 0.62 | 0.78/ **0.71** | **76.96** |
| BOW (w2v) | 0.39/ 0.46 | 0.67/ **0.66** | **0.64**/ **0.60** | 0.76/ **0.72** | **0.70**/ **0.69** | **0.79**/ 0.69 | 75.62 |

ляется последним слоем LSTM сети, а затем с помощью другой LSTM сети происходит предсказание того же предложения. Рассматривались несколько вариантов разреживающих слоев и проверялось качество восстановления предложения, а также интерпретируемость компонент получаемых представлений. Интерпретируемость измерялась с помощью когерентности, адаптированной для случая предложений. В результате было показано, что разреженные модели существенно улучшают интерпретируемость, однако могут ухудшать качество восстановления предложений.

**Тематические представления предложений.** В подходе, предлагаемом в данной работе, вектора слов являются вероятностными распределениями на множестве тем $p(t|w)$ и получаются из параметров модели $\phi_{wt} = p(w|t)$ по формуле Байеса. Представления предложений могут быть получены с помощью усреднения векторов слов, что соответствует одной итерации EM-алгоритма с фиксированными предобученными параметрами $\Phi$. При такой интерпретации интересной возможностью представляется уточнение векторов предложений $\Theta$ повторением EM-итераций до сходимости. В таком случае вхождения слов в документы получают контекстно-зависимые распределения $p(t|w, s)$, где $t$ – тема, $w$ – слово, $s$ – предложение, а вектора предложений являются усреднением этих распределений. Такой подход представляется предпочтительным, т.к. позволяет уточнять тематические вектора многозначных слов исходя из контекста, что должно приводить к более специфичным векторам предложений.

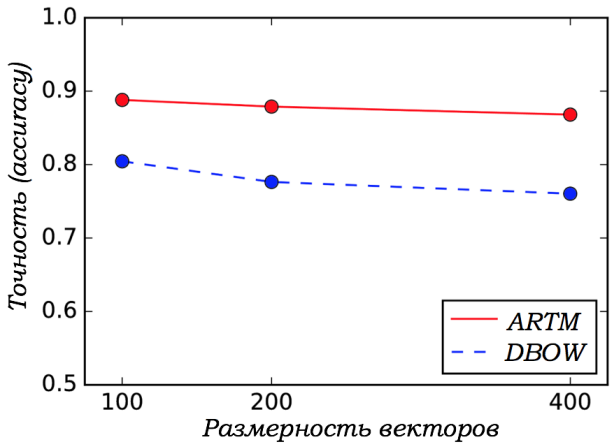



Рис. 5.2. Точность моделей ARTM и doc2vec-DBOW для нескольких размерностей пространства (100, 200, 400) на задаче предсказания близости в тройках статей ArXiv.

Это было подтверждено в нашем эксперименте [33]. Оценивалось качество решения задач семантической близости предложений на данных STS-2014 [140], а также задач семантической связанности (relatedness) и определения логического следования предложений (entailment) на данных SICK [141]. Векторные представления предложений использовались в линейных моделях, обученных с помощью инструмента SentEval [142]. Сравнивался стандартный подход, усредняющий вектора word2vec, а также два новых подхода: усреднение тематических векторов слов (BOW) и уточнение этих векторов по контексту с помощью 10 проходов EM-алгоритма (fitted). Во всех случаях вектора слов были предобучены на текстах английской Википедии. В таблице 5.7 показано, что контекстное уточнение векторов улучшает качество на всех наборах данных. При этом качество оказывается сопоставимым с качеством векторов, полученных на основе модели word2vec.

Как и в случае отдельных слов, полученные тематические представления предложений имеют преимущества разреженности и интерпретируемости. Предлагаемый подход может быть обобщен для произвольных признаков последовательного текста [33].

**Тематические представления документов.** Подход тематического моделирования естественным образом позволяет обучать представления документов. В эксперименте на коллекции статей ArXiv исследовалось качество решения задачи семантической близости документов. Датасет [143] состоит из автоматически сгенерированных триплетов статей: запрос, статья с пересекающимися ключевыми слова (семантически близкая) и статья с непересекающимися ключевыми словами (семантически далекая). Качество моделей оценивалось как точность определения, какая из статей близка к запросу. Тексты 963564 статей общей длиной 1416554733 токенов были предобработаны[1], после чего размер словаря составил 122596 слов. В нашем корпусе статей нашлось 15853 триплетов из оригинальных 20000 триплетов в тестового набора[2].

Тематические представления документов обучались одной эпохой онлайнового EM-алгоритма. Матрица $\Theta$ не хранилась, таким образом, алгоритм PWE не требовал объема памяти, линейного по числу документов. Представления статей из тестовой выборки были найдены с помощью 10 проходов по каждому документу. Тематическая модель сравнивалась с классической для этой задачи моделью DBOW [143] семейства doc2vec. Она тренировалась 15 эпохами с линейным затуханием градиентного шага с 0.025 до 0.001. Векторные представления тестовых документов были получены за 5 эпох. В отличие от EM-алгоритма, модель DBOW требует хранения векторных представлений всех документов в памяти и обучается дольше (несколько часов вместо 30 минут на одной машине). На Рисунке 5.2 видно, что предлагаемая ARTM модель существенно превосходит модель DBOW на широком диапазоне размерностей векторного пространства.

[1] https://github.com/romovpa/arxiv-dataset

[2] http://cs.stanford.edu/~quocle/triplets-data.tar.gz

## 5.7. Обсуждение и выводы

Проведено теоретическое и экспериментальное сравнение тематических моделей и векторных моделей семантики. Предложена формализация гипотезы дистрибутивной семантики в рамках подхода аддитивной регуляризации тематических моделей. Доказана теорема, на основе которой предложен новый алгоритм построения векторных представлений слов PWE. В экспериментах показано, что с его помощью удается определять семантическую близость слов наравне с известным методом SGNS и интерпретировать координаты векторного пространства как содержательные темы коллекции.

Показана связь модели PWE с другими тематическими моделями совместной встречаемости слов. Доказана теорема об эквивалентности моделей WNTM и BTM при связывании входных и выходных представлений слов. Теоретический результат подтвержден в экспериментах.

Предложен способ одновременного учета гипотезы дистрибутивной семантики и данных дополнительных модальностей. С помощью полученного расширения подхода APTM построено единое векторное пространство для токенов различных модальностей, и продемонстрированы интерпретируемые кроссмодальные близости. Показано улучшение качества предсказания семантической близости слов в результате введения в модель модальностей времени и категорий.

Предлагаемый подход представлений слов обобщен на случай сегментов текста, в частности, отдельных предложений и документов. В экспериментах получено качество, сопоставимое или превосходящее стандартные подходы семейства word2vec.

Стоит заметить, что в модели PWE не используется информация о частях слова (морфемах или буквенных $n$-граммах). Использование такой информации может повышать качество, как показано в последних работах по векторным представлениям слов. В частности, на этом основана модель FastText [3]. Второе

направление недавних исследований связано с обучением контексто-зависимых представлений слов. Модель ELMo [4], предложенная в 2018 году, превосходит другие модели на большом числе прикладных задач. Расширение разрабатываемого подхода тематических векторных представлений слов для учета частей слов и слов контекста представляется перспективной темой дальнейшего исследования. Первые эксперименты по контекстно-зависимым представлениям в модели PWE проведены в [33] и подтверждают, что отказ от гипотезы мешка слов приводит к повышению качества тематических векторных представлений.

# Заключение

Отправной точкой данного исследования стали методы вероятностного тематического моделирования, позволяющие описывать коллекцию текстовых документов с помощью набора скрытых тем. Обобщение известных алгоритмов построения таких моделей (PLSA-EM, LDA-GS и др.) позволило получить семейство EM-алгоритмов, в котором эвристики робастности к шуму и фону, сэмплирования тем и регуляризации Дирихле могут включаться независимо. Была предложена упрощенная робастная модель, в которой фоновая компонента отсутствует, а вес компоненты шума, предназначенной для описания редких нетематических терминов документов, настраивается автоматически. В экспериментах было показано, что в таких моделях удается строить сильно разреженные распределения, а дополнительная регуляризация Дирихле не требуется.

Далее модель PLSA, свободная от регуляризации, была использована как основа для построения *аддитивно регуляризованных тематических моделей*, учитывающих дополнительные критерии на матрицы параметров $\Phi$ и $\Theta$. В рамках данного подхода была предложена модель фоновых и предметных тем, являющаяся логичным продолжением робастных моделей. В данной модели сглаженные фоновые темы описывают общую лексику языка, а также наиболее частотные слова конкретной коллекции. Это позволяет освободить предметные темы от неспецифичных фоновых слов под действием комбинации регуляризаторов разреживания и декоррелирования. В экспериментах подтверждается повышение интерпретируемости, разреженности и различности предметных тем.

Аппарат построения интерпретируемых разреженных тематических моделей применяется для получения *семантических векторных представлений слов*, ключевым свойством которых является сохранение семантических расстояний. Это становится возможным после перехода от моделирования частот слов в документах к моделированию совместных частот слов в локальных контекстах. Предлагается модель тематических векторных представлений слов, в

которой комбинируются преимущества тематических моделей и моделей дистрибутивной семантики. В экспериментах показывается высокое качество решения задачи близости слов, а также сохранение интерпретируемости компонент векторов. С помощью регуляризатора разреживания удается получить более 90% нулей в построенных векторах без потери качества.

В последней части работы подход аддитивной регуляризации используется для построения единого векторного пространства для токенов различных модальностей (слов, дат, категорий и т.д.). Предложенная модель превосходит другие подходы по качеству решения задачи близости слов, а также демонстрирует интерпретируемые кросс-модальные близости. Расширение модели позволяет строить тематические векторные представления сегментов текста, в частности, предложений или целых документов. Данный подход показывает высокое качество на задаче определения семантической близости научных статей arXiv.

В дальнейшем планируется продолжение исследования и применение полученных интерпретируемых векторных представлений для построения систем разведочного информационного поиска. В данном приложении требуется не только определить семантическую близость слов или сегментов текста, но и проинтерпретировать темы, которые обеспечивают эту близость. Совмещение этих требований возможно в рамках предложенного подхода.

# Список литературы


1. Distributed Representations of Words and Phrases and their Compositionality. / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // NIPS / Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, Kilian Q. Weinberger. — 2013. — Pp. 3111–3119.
2. *Pennington Jeffrey, Socher Richard, Manning Christopher D.* Glove: Global Vectors for Word Representation. // EMNLP. — Vol. 14. — 2014. — Pp. 1532–1543.
3. Enriching Word Vectors with Subword Information / Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov // *Transactions of the Association for Computational Linguistics*. — 2017. — Vol. 5. — Pp. 135–146.
4. Deep Contextualized Word Representations / Matthew Peters, Mark Neumann, Mohit Iyyer et al. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Association for Computational Linguistics, 2018. — Pp. 2227–2237.
5. *Schunn C. D.* The Presence and Absence of Category Knowledge in LSA // Proceedings of the 21st Annual Conference of the Cognitive Science Society. — Mahwah. Erlbaum., 1999.
6. *Murphy Brian, Talukdar Partha Pratim, Mitchell Tom M.* Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. // COLING / Ed. by Martin Kay, Christian Boitet. — Indian Institute of Technology Bombay, 2012. — Pp. 1933–1950.
7. *Harris Zellig.* Distributional structure // *Word*. — 1954. — Vol. 10, no. 23. — Pp. 146–162.
8. *Firth J.R.* A synopsis of linguistic theory 1930-55 // *Studies in linguistic analysis. The Philological Society, Oxford*. — 1957. — Pp. 1–32.
9. Indexing by latent semantic analysis. / Scott Deerwester, Susan T. Dumais,

George W. Furnas et al. // *Journal of the American Society for Information Science 41.* — 1990. — Pp. 391–407.

10. *Lund Kevin, Burgess Curt.* Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence // *Behavior Research Methods, Instruments, & Computers.* — 1996. — Vol. 28. — Pp. 203–208.
11. *Turney Peter D., Pantel Patrick.* From Frequency to Meaning: Vector Space Models of Semantics // *Journal of Artificial Intelligence Research, (2010), 37, 141-188.* — 2010.
12. *Levy Omer, Goldberg Yoav, Dagan Ido.* Improving Distributional Similarity with Lessons Learned from Word Embeddings. // *TACL.* — 2015. — Vol. 3. — Pp. 211–225.
13. *Hofmann Thomas.* Probabilistic Latent Semantic Analysis // Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. — UAI'99. — San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. — Pp. 289–296.
14. *Blei David M., Ng Andrew Y., Jordan Michael I.* Latent dirichlet allocation // *Journal of Machine Learning Research.* — 2003. — Vol. 3. — Pp. 993–1022.
15. *Masada Tomonari, Kiyasu Senya, Miyahara Sueharu.* Comparing LDA with pLSI as a dimensionality reduction method in document clustering // Proceedings of the 3rd International Conference on Large-scale knowledge resources: construction and application. — LKR'08. — Springer-Verlag, 2008. — Pp. 13–26.
16. A comparative study of topic models for topic clustering of Chinese web news / Yonghui Wu, Yuxin Ding, Xiaolong Wang, Jun Xu // Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on. — Vol. 5. — 2010. — Pp. 236–240.
17. *Lu Yue, Mei Qiaozhu, Zhai ChengXiang.* Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA // *Information Retrieval.* — 2011. — Vol. 14, no. 2. — Pp. 178–203.

18. Knowledge discovery through directed probabilistic topic models: a survey / Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad // *Frontiers of Computer Science in China*. — 2010. — Vol. 4, no. 2. — Pp. 280–301.
19. *Blei David M.* Probabilistic topic models // *Communications of the ACM*. — 2012. — Vol. 55, no. 4. — Pp. 77–84.
20. *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // *Доклады РАН*. — 2014. — Т. 456, № 3. — С. 268–271.
21. *Vorontsov K. V., Potapenko A. A.* Additive Regularization of Topic Models // *Machine Learning Journal*. — 2015. — Vol. 101. — Pp. 303–323.
22. *Potapenko A. A., Vorontsov K. V.* Robust PLSA Performs Better Than LDA // 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. — Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 2013. — Pp. 784–787.
23. *Воронцов К. В., Потапенко А. А.* Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). — Вып. 13 (20). — М: Изд-во РГГУ, 2014. — С. 676–687.
24. *Vorontsov K. V., Potapenko A. A.* Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // Analysis of Images, Social networks and Texts (AIST 2014). — Vol. 436 of *Communications in Computer and Information Science*. — Springer International Publishing Switzerland, 2014. — Pp. 29–46.
25. *Vorontsov K. V., Potapenko A. A., Plavin A.V.* Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // The Third International Symposium On Learning And Data Sciences (SLDS 2015). — Vol. 9047. — Springer, A. Gammerman et al. (Eds.), LNAI, 2015. — P. 193–202.
26. *Potapenko A., Popov A., Vorontsov K.* Interpretable Probabilistic Embed-

dings: Bridging the Gap Between Topic Models and Neural Networks // AINL: Artificial Intelligence and Natural Language Conference / Ed. by Andrey Filchenkov, Lidia Pivovarova, Jan Žižka. — Vol. 789 of *Communications in Computer and Information Science*. — Springer International Publishing, 2017. — Pp. 167–180.

27. *Воронцов К. В., Потапенко А. А.* Регуляризация, робастность и разреженность вероятностных тематических моделей // *Компьютерные исследования и моделирование*. — 2012. — Т. 4, № 4. — С. 693–706.
28. *Воронцов К. В., Потапенко А. А.* Модификации EM-алгоритма для вероятностного тематического моделирования // *Машинное обучение и анализ данных*. — 2013. — Т. 1, № 6. — С. 657–686.
29. Learning and Evaluating Sparse Interpretable Sentence Embeddings / Valentin Trifonov, Octavian-Eugen Ganea, Anna Potapenko, Thomas Hofmann // EMNLP 2018 Workshop: Analyzing and interpreting neural networks for NLP. — Association for Computational Linguistics, 2018. — Pp. 200–210.
30. *Воронцов К. В., Потапенко А. А.* Робастные разреженные вероятностные тематические модели // Интеллектуализация обработки информации (ИОИ-2012): Докл. — Торус Пресс, 2012. — Pp. 605–608.
31. *Потапенко А. А.* Разреживание вероятностных тематических моделей // Математические методы распознавания образов: 16-ая Всеросс. конф.: Докл. — МАКС Пресс, 2013. — P. 89.
32. *Потапенко А. А.* Регуляризация вероятностной тематической модели для выделения ядер тем // Сборник тезисов XXI Международной научной конференции студентов, аспирантов и молодых ученых «Ломоносов-2014». — МАКС Пресс, 2014.
33. *Potapenko A.* Probabilistic approach for embedding arbitrary features of text // 7th International Conference - Analysis of Images, Social networks and Texts (AIST 2018). — LNCS, Springer, 2018 (to appear).
34. *Sahlgren Magnus, Lenci Alessandro.* The Effects of Data Size and Frequency

Range on Distributional Semantic Models // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. — 2016. — Pp. 975–980.

35. *Resnik Philip.* Using Information Content to Evaluate Semantic Similarity in a Taxanomy // International Joint Conference for Artificial Intelligence (IJCAI-95). — 1995. — Pp. 448–453.
36. *Budanitsky Alexander, Hirst Graeme.* Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures // Workshop on WordNet and other lexical resources, NAACL. — 2001.
37. *Schütze Hinrich, Pedersen Jan.* A Vector Model for syntagmatic and paradigmatic relatedness // Proc. of the 9th Annual Conference of the UW Centre for the New OED and Text Research. — Oxford, England: 1993. — Pp. 104–113.
38. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches / Eneko Agirre, Enrique Alfonseca, Keith Hall et al. // Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — NAACL '09. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. — Pp. 19–27.
39. *Gentner Dedre.* Structure-mapping: A theoretical framework for analogy // *Cognitive Science.* — 1983. — Vol. 7, no. 2. — Pp. 155–170.
40. *Mikolov Tomas, Yih Wen-tau, Zweig Geoffrey.* Linguistic Regularities in Continuous Space Word Representations // HLT-NAACL. — 2013. — Pp. 746–751.
41. *Turney Peter D.* Similarity of Semantic Relations // *Computational Linguistics.* — 2006. — Vol. 22, no. 3. — Pp. 379–416.
42. *Turney Peter D.* The Latent Relation Mapping Engine: Algorithm and Experiments // *Journal of Artificial Intelligence Research, (2008), 33, 615-655.* — 2008. — .
43. *Levy Omer, Goldberg Yoav.* Linguistic Regularities in Sparse and Explicit Word Representations. // CoNLL / Ed. by Roser Morante, Wen tau Yih. — ACL,

2014. — Pp. 171–180.

44. *Salton G., Wong A., Yang C. S.* A Vector Space Model for Automatic Indexing // *Commun. ACM.* — 1975. — . — Vol. 18, no. 11. — Pp. 613–620.
45. *G. Grefenstette., Tapanainen P.* What Is a Word, What Is a Sentence? Problems of Tokenization. // Proceedings of the 3rd International Conference on Computational Lexicography. — 1994. — Pp. 79–87.
46. *Donald Hindle.* Noun Classification from Predicate-Argument structures // 28th Annual Meeting of the Association for Computational Linguistics. — 1990.
47. *Riloff Ellen, Shepherd Jessica.* A Corpus-Based Approach for Building Semantic Lexicons // Second Conference on Empirical Methods in Natural Language Processing, EMNLP 1997, Providence, RI, USA, August 1-2, 1997. — 1997.
48. *Socher R., Biemann C., Osswald R.* Combining Contexts in Lexicon Learning for Semantic Parsing // Proceedings of NODALIDA-07. — Tartu, Estonia: 2007.
49. *Bisson Gilles, Nédellec Claire, Cañamero Dolores.* Designing Clustering Methods for Ontology Building: The Mo'K Workbench // Proceedings of the First International Conference on Ontology Learning - Volume 31. — OL'00. — Aachen, Germany, Germany: CEUR-WS.org, 2000. — Pp. 13–28.
50. *Pantel Patrick, Ravichandran Deepak, Hovy Eduard H.* Towards Terascale Semantic Acquisition // COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland. — 2004.
51. *Bullinaria J.A., Levy J.P.* Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study // *Behavior Research Methods.* — 2007. — Vol. 39. — Pp. 510–526.
52. Online Learning for Matrix Factorization and Sparse Coding / Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro // *Journal of Machine Learning Research.* — 2010. — Vol. 11. — Pp. 19–60.
53. *Hoyer P. O.* Non-negative sparse coding // NNSP. — 2002.

54. *Zuo Yuan, Zhao Jichang, Xu Ke.* Word network topic model: a simple but general solution for short and imbalanced texts. // *Knowl. Inf. Syst.* — 2016. — Vol. 48, no. 2. — Pp. 379–398.

55. A Neural Probabilistic Language Model / Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Janvin // *Journal of Machine Learning Research.* — 2003. — Vol. 3. — Pp. 1137–1155.

56. Efficient Estimation of Word Representations in Vector Space / Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean // *CoRR.* — 2013. — Vol. abs/1301.3781.

57. *Mnih Andriy, Hinton Geoffrey E.* A Scalable Hierarchical Distributed Language Model // NIPS. — 2008.

58. *Mimno David M., Thompson Laure.* The strange geometry of skip-gram with negative sampling // EMNLP. — 2017.

59. *Levy Omer, Goldberg Yoav.* Neural Word Embedding as Implicit Matrix Factorization // Advances in Neural Information Processing Systems 27 / Ed. by Z. Ghahramani, M. Welling, C. Cortes et al. — Curran Associates, Inc., 2014. — Pp. 2177–2185.

60. *Melamud Oren, Goldberger Jacob.* Information-Theory Interpretation of the Skip-Gram Negative-Sampling Objective Function // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — Association for Computational Linguistics, 2017. — Pp. 167–171.

61. *Turian Joseph Lev Ratinov, Bengio Yoshua.* Word representations: a simple and general method for semi-supervised learning. // Proceedings of the 48th annual meeting of the association for computational linguistics. — Association for Computational Linguistics, 2010.

62. *Hinton G. E. McClelland J. L., Rumelhart D. E.* Distributed representations. // *Rumelhart, D. E. and McClelland, J. L., editors, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foun-*

*dations.* — 1986.

63. *Marco Baroni, Georgiana Dinu Germán Kruszewski.* Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors // *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference.* — 2014. — Vol. 1. — Pp. 238–247.
64. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora / Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2010. — Pp. 1079–1088.
65. TextFlow: Towards Better Understanding of Evolving Topics in Text. / Weiwei Cui, Shixia Liu, Li Tan et al. // *IEEE transactions on visualization and computer graphics.* — 2011. — Vol. 17, no. 12. — Pp. 2412–2421.
66. Statistical topic models for multi-label document classification / Timothy N. Rubin, America Chambers, Padhraic Smyth, Mark Steyvers // *Machine Learning.* — 2012. — Vol. 88, no. 1-2. — Pp. 157–208.
67. Simultaneous image classification and annotation based on probabilistic model / Xiao-Xu Li, Chao-Bo Sun, Peng Lu et al. // *The Journal of China Universities of Posts and Telecommunications.* — 2012. — Vol. 19, no. 2. — Pp. 107–115.
68. *Feng Yansong, Lapata Mirella.* Topic models for image annotation and text illustration // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Association for Computational Linguistics, 2010. — Pp. 831–839.
69. *Yi Xing, Allan James.* A Comparative Study of Utilizing Topic Models for Information Retrieval // Advances in Information Retrieval. — Springer Berlin Heidelberg, 2009. — Vol. 5478 of *Lecture Notes in Computer Science.* — Pp. 29–41.
70. *Vulić Ivan, Smet Wim, Moens Marie-Francine.* Cross-language information retrieval models based on latent topic models trained with document-aligned

comparable corpora // *Information Retrieval.* — 2012. — Pp. 1–38.

71. *Krestel Ralf, Fankhauser Peter, Nejdl Wolfgang.* Latent dirichlet allocation for tag recommendation // Proceedings of the third ACM conference on Recommender systems. — ACM, 2009. — Pp. 61–68.
72. *Zavitsanos Elias, Paliouras Georgios, Vouros George A.* Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // *Journal of Machine Learning Research.* — 2011. — Vol. 12. — Pp. 2749–2775.
73. *Jameel Shoaib, Lam Wai.* An N-Gram Topic Model for Time-Stamped Documents // 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. — Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 2013. — Pp. 292–304.
74. *Blei David M., Ng Andrew Y., Jordan Michael I.* Latent Dirichlet allocation // *Journal of Machine Learning Research.* — 2003. — Vol. 3. — Pp. 993–1022.
75. *Hofmann Thomas.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. — Pp. 50–57.
76. *Bishop Christopher M.* Pattern Recognition and Machine Learning (Information Science and Statistics). — Berlin, Heidelberg: Springer-Verlag, 2006.
77. On Smoothing and Inference for Topic Models / A. Asuncion, M. Welling, P. Smyth, Y. W. Teh // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. — 2009. — Pp. 27–34.
78. *Vetrov D.P. Kropotov D.A.* Bayesian methods of machine learning. // Lecture notes. — 2014.
79. *McCallum A Mimno DM Wallach HM.* Rethinking LDA: Why Priors Matter. // Advances in Neural Information Processing Systems 22. — 2009. — Pp. 1973–1981.
80. *Dempster A. P., Laird N. M., Rubin D. B.* Maximum likelihood from incomplete data via the EM algorithm // *J. of the Royal Statistical Society, Series*

*B.* — 1977. — no. 34. — Pp. 1–38.

81. *Steyvers Mark, Griffiths Tom.* Finding scientific topics // *Proceedings of the National Academy of Sciences.* — 2004. — Vol. 101, no. Suppl. 1. — Pp. 5228–5235.
82. *Wang Yi.* Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. — 2008.
83. *Chemudugunta C., Smyth P., Steyvers M.* Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems. — MIT Press, 2007. — Vol. 19. — Pp. 241–248.
84. *Eisenstein Jacob, Ahmed Amr, Xing Eric P.* Sparse Additive Generative Models of Text // ICML'11. — 2011. — Pp. 1041–1048.
85. *Wang Chong, Blei David M.* Decoupling Sparsity and Smoothness in the Discrete Hierarchical Dirichlet Process // NIPS. — Curran Associates, Inc., 2009. — Pp. 1982–1989.
86. *Larsson Martin O., Ugander Johan.* A concave regularization technique for sparse mixture models // Advances in Neural Information Processing Systems 24 / Ed. by J. Shawe-Taylor, R.S. Zemel, P. Bartlett et al. — 2011. — Pp. 1890–1898.
87. *Tikhonov A. N., Arsenin V. Y.* Solution of ill-posed problems. — W. H. Winston, Washington, DC, 1977.
88. Multi-Objective Topic Modelling / O. Khalifa, D. Corne, M. Chantler, F. Halley // 7th International Conference Evolutionary Multi-Criterion Optimization (EMO 2013). — Springer LNCS, 2013. — Pp. 51–65.
89. *Si Luo, Jin Rong.* Adjusting Mixture Weights of Gaussian Mixture Model via Regularized Probabilistic Latent Semantic Analysis // Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) / Ed. by Tu Bao Ho, David Wai-Lok Cheung, Huan Liu. — Vol. 3518 of *Lecture Notes in Computer Science.* — Springer, 2005. — Pp. 622–631.
90. *Chien Jen-Tzung, Wu Meng-Sung.* Adaptive Bayesian Latent Semantic Analy-

sis // *IEEE Transactions on Audio, Speech, and Language Processing*. — 2008. — Vol. 16, no. 1. — Pp. 198–207.

91. Regularized latent semantic indexing / Quan Wang, Jun Xu, Hang Li, Nick Craswell // SIGIR. — 2011. — Pp. 685–694.
92. *Varadarajan Jagannadan, Emonet Rémi, Odobez Jean-Marc*. A Sparsity Constraint for Topic Models — Application to Temporal Activity Mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions. — 2010.
93. *Shashanka Madhusudana, Raj Bhiksha, Smaragdis Paris*. Sparse Overcomplete Latent Variable Decomposition of Counts Data // Advances in Neural Information Processing Systems, NIPS-2007 / Ed. by J. C. Platt, D. Koller, Y. Singer, S. Roweis. — Cambridge, MA: MIT Press, 2008. — Pp. 1313–1320.
94. *Chien Jen-Tzung, Chang Ying-Lan*. Bayesian Sparse Topic Model // *Journal of Signal Processessing Systems*. — 2013. — Pp. 1–15.
95. *Tan Yimin, Ou Zhijian*. Topic-weak-correlated Latent Dirichlet allocation // 7th International Symposium Chinese Spoken Language Processing (ISCSLP). — 2010. — Pp. 224–228.
96. Automatic evaluation of topic coherence / David Newman, Jey Han Lau, Karl Grieser, Timothy Baldwin // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — HLT '10. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. — Pp. 100–108.
97. Evaluating topic models for digital libraries / David Newman, Youn Noh, Edmund Talley et al. // Proceedings of the 10th annual Joint Conference on Digital libraries. — JCDL '10. — New York, NY, USA: ACM, 2010. — Pp. 215–224.
98. Optimizing semantic coherence in topic models / David Mimno, Hanna M. Wallach, Edmund Talley et al. // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — EMNLP '11. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. — Pp. 262–272.

99. *Newman David, Bonilla Edwin V., Buntine Wray L.* Improving Topic Coherence with Regularized Topic Models // Advances in Neural Information Processing Systems 24 / Ed. by J. Shawe-Taylor, R.S. Zemel, P. Bartlett et al. — 2011. — Pp. 496–504.

100. *Newman David, Karimi Sarvnaz, Cavedon Lawrence.* External Evaluation of Topic Models // Australasian Document Computing Symposium. — 2009. — December. — Pp. 11–18.

101. Reading Tea Leaves: How Humans Interpret Topic Models / Jonathan Chang, Sean Gerrish, Chong Wang et al. // Neural Information Processing Systems (NIPS). — 2009. — Pp. 288–296.

102. *Friedman Jerome H., Hastie Trevor, Tibshirani Rob.* Regularization Paths for Generalized Linear Models via Coordinate Descent // *Journal of Statistical Software.* — 2010. — Vol. 33, no. 1. — Pp. 1–22.

103. *McCallum Andrew Kachites.* Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. — http://www.cs.cmu.edu/∼mccallum/bow.

104. *Hoffman Matthew D., Blei David M., Bach Francis R.* Online Learning for Latent Dirichlet Allocation. // NIPS / Ed. by John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor et al. — Curran Associates, Inc., 2010. — Pp. 856–864.

105. Placing Search in Context: The Concept Revisited / Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias et al. // *ACM Trans. Inf. Syst.* — 2002. — . — Vol. 20, no. 1. — Pp. 116–131.

106. Distributional Semantics in Technicolor / Elia Bruni, Gemma Boleda, Marco Baroni, Nam-Khanh Tran // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1. — ACL '12. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. — Pp. 136–145.

107. *Hill Felix, Reichart Roi, Korhonen Anna.* Simlex-999: Evaluating Semantic

Models with Genuine Similarity Estimation // *Comput. Linguist.* — 2015. — Vol. 41, no. 4. — Pp. 665–695.

108. A Word at a Time: Computing Word Relatedness using Temporal Semantic Analysis / Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, Shaul Markovitch // Proceedings of the 20th International World Wide Web Conference. — Hyderabad, India: 2011. — March. — Pp. 337–346.
109. *Hoffman Matthew D., Blei David M., Bach Francis R.* Online Learning for Latent Dirichlet Allocation // NIPS. — Curran Associates, Inc., 2010. — Pp. 856–864.
110. *Rogers Anna, Drozd Aleksandr, Li Bofang.* The (too Many) Problems of Analogical Reasoning with Word Vectors // *SEM. — 2017.
111. *Gladkova Anna, Drozd Aleksandr, Matsuoka Satoshi.* Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't // SRW@HLT-NAACL. — 2016.
112. *Finley Gregory P., Farmer Stephanie, Pakhomov Serguei V. S.* What Analogies Reveal about Word Vectors and their Compositionality // *SEM. — 2017.
113. *Zobnin Alexey.* Rotations and Interpretability of Word Embeddings: The Case of the Russian Language // Analysis of Images, Social Networks and Texts. — Cham: Springer International Publishing, 2018. — Pp. 116–128.
114. Automatic Evaluation of Topic Coherence / David Newman, Jey Han Lau, Karl Grieser, Timothy Baldwin // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — HLT '10. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. — Pp. 100–108.
115. Optimizing Semantic Coherence in Topic Models / David Mimno, Hanna M. Wallach, Edmund Talley et al. // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — EMNLP '11. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. — Pp. 262–272.

116. *Newman David, Bonilla Edwin V., Buntine Wray L.* Improving Topic Coherence with Regularized Topic Models // NIPS. — 2011.

117. *Aletras Nikolaos, Stevenson Mark.* Evaluating Topic Coherence Using Distributional Semantics // IWCS. — 2013.

118. *Röder Michael, Both Andreas, Hinneburg Alexander.* Exploring the Space of Topic Coherence Measures // Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. — WSDM '15. — New York, NY, USA: ACM, 2015. — Pp. 399–408.

119. Online Learning of Interpretable Word Embeddings / Hongyin Luo, Zhiyuan Liu, Huan-Bo Luan, Maosong Sun // EMNLP. — 2015.

120. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections. / Konstantin Vorontsov, Oleksandr Frei, Murat Apishev et al. // TM@CIKM / Ed. by Nikolaos Aletras, Jey Han Lau, Timothy Baldwin, Mark Stevenson. — ACM, 2015. — Pp. 29–37.

121. Human and Machine Judgements for Russian Semantic Relatedness / Alexander Panchenko, Dmitry Ustalov, Nikolay Arefyev et al. // Analysis of Images, Social Networks and Texts (AIST'2016). — Springer, 2016.

122. *Miller George A., Charles Walter G.* Contextual correlates of semantic similarity // *Language and Cognitive Processes.* — 1991. — Vol. 6, no. 1. — Pp. 1–28.

123. *Rubenstein Herbert, Goodenough John B.* Contextual Correlates of Synonymy // *Commun. ACM.* — 1965. — . — Vol. 8, no. 10. — Pp. 627–633.

124. *Leviant Ira, Reichart Roi.* Judgment Language Matters: Towards Judgment Language Informed Vector Space Modeling // Preprint pubslished on arXiv (arxiv:1508.00106). — 2015.

125. *Levy Omer, Goldberg Yoav.* Dependency-Based Word Embeddings // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — Association for Computational Linguistics, 2014. — Pp. 302–308.

126. *Press Ofir, Wolf Lior.* Using the Output Embedding to Improve Language

Models // Proceedings of ACL: Volume 2, Short Papers. — ACL, 2017. — Pp. 157–163.

127. *Inan Hakan, Khosravi Khashayar, Socher Richard.* Tying Word Vectors and Word Classifiers: A Loss Framework for Language Modeling // *CoRR*. — 2016. — Vol. abs/1611.01462.
128. A biterm topic model for short texts. / Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng // Proceedings of WWW. — 2013. — Pp. 1445–1456.
129. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections / Konstantin Vorontsov, Oleksandr Frei, Murat Apishev et al. // AIST. — 2015.
130. *Arora Sanjeev, Liang Yingyu, Ma Tengyu.* A Simple but Tough-to-Beat Baseline for Sentence Embeddings // International Conference on Learning Representations. — 2017.
131. *Pagliardini Matteo, Gupta Prakhar, Jaggi Martin.* Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features // Proceedings of NAACL. — 2018.
132. StarSpace: Embed All The Things! / Ledell Wu, Adam Fisch, Sumit Chopra et al. // *CoRR*. — 2017. — Vol. abs/1709.03856.
133. Skip-thought Vectors / Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov et al. // Proceedings of the 28th International Conference on Neural Information Processing Systems. — NIPS'15. — Cambridge, MA, USA: MIT Press, 2015. — Pp. 3294–3302.
134. *Li Jiwei, Luong Minh-Thang, Jurafsky Dan.* A Hierarchical Neural Autoencoder for Paragraphs and Documents. // ACL (1). — The Association for Computer Linguistics, 2015. — Pp. 1106–1115.
135. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data / Alexis Conneau, Douwe Kiela, Holger Schwenk et al. // Proceedings of EMNLP. — Association for Computational Linguistics, 2017. — Pp. 670–680.

136. Universal Sentence Encoder / Daniel Cer, Yinfei Yang, Sheng-yi Kong et al. // *CoRR*. — 2018. — Vol. abs/1803.11175.

137. A large annotated corpus for learning natural language inference / Samuel R. Bowman, Gabor Angeli, Christopher Potts, Christopher D. Manning // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Association for Computational Linguistics, 2015.

138. *Hill Felix, Cho Kyunghyun, Korhonen Anna.* Learning Distributed Representations of Sentences from Unlabelled Data // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Association for Computational Linguistics, 2016. — Pp. 1367–1377.

139. *Perone Christian S., Silveira Roberto, Paula Thomas S.* Evaluation of sentence embeddings in downstream and linguistic probing tasks // *CoRR*. — 2018. — Vol. abs/1806.06259.

140. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity / Eneko Agirre, Carmen Banea, Claire Cardie et al. // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). — Association for Computational Linguistics, 2014. — Pp. 81–91.

141. A SICK cure for the evaluation of compositional distributional semantic models / Marco Marelli, Stefano Menini, Marco Baroni et al. // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014). — European Language Resources Association (ELRA), 2014.

142. *Conneau Alexis, Kiela Douwe.* SentEval: An Evaluation Toolkit for Universal Sentence Representations // *CoRR*. — 2018. — Vol. abs/1803.05449.

143. *Dai Andrew M., Olah Christopher, Le Quoc V.* Document Embedding with Paragraph Vectors // *CoRR*. — 2015. — Vol. abs/1507.07998.