На правах рукописи

Потапенко Анна Александровна

Семантические векторные представления текста на основе вероятностного тематического моделирования

05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва – 2018

Работа выполнена в департаменте больших данных и информационного поиска факультета компьютерных наук ФГАОУ ВО Национальный исследовательский университет «Высшая школа экономики».

Научный руководитель:	**Воронцов Константин Вячеславович**, доктор физико-математических наук, профессор РАН, руководитель лаборатории машинного интеллекта ФГАОУ ВО «Московский Физико-Технический Институт (ГУ)».
Официальные оппоненты:	**Бессмертный Игорь Александрович**, доктор технических наук, профессор факультета программной инженерии и компьютерной техники, ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики».
	Тутубалина Елена Викторовна, кандидат физико-математических наук, старший научный сотрудник НИЛ «Хемоинформатика и молекулярное моделирование» Химического института им. А.М. Бутлерова, ФГАОУ ВО «Казанский (Приволжский) федеральный университет».
Ведущая организация:	Федеральное государственное бюджетное учреждение науки Институт системного программирования им. В.П. Иванникова Российской академии наук.

Защита состоится «28» февраля 2019 года в 16:00 на заседании диссертационного совета Д 002.073.05 при Федеральном государственном учреждении «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН) по адресу 119333, г. Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке и на сайте ФИЦ ИУ РАН http://frccsc.ru.

Автореферат разослан « » декабря 2018 года.

Ученый секретарь
диссертационного совета Д 002.073.05,

д.ф.-м.н., профессор Рязанов В.В.

Общая характеристика работы

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 (Mikolov и др., 2013; Pennington и др., 2014; Bojanowski и др., 2017; Peters и др., 2018). 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 векторы низкой размерности, причем каждое слово представляется набором фиксированного числа признаков. Это противоречит гипотезе об экономном хранении (Murphy и др., 2012), согласно которой человеческий мозг представляет более специфичные концепты большим числом характеристик, а более общие — меньшим. 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 позволяет 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 (Harris, 1954; Firth, 1957). Модели векторного представления слов, основанные на частотных распределениях слов в контекстах, развиваются на протяжении последних десятилетий и хорошо изучены. Одними из первых таких моделей можно считать работы 1990-х годов Latent Semantic Analysis, LSA (Deerwester и др., 1990) и Hyperspace Analogue to Language, HAL (Lund, Burgess, 1996). Эти модели позволяют представлять слова векторами в некотором низкоразмерном пространстве, так что семантически близкие слова имеют близкие вектора. Для оценивания моделей существуют составленные вручную наборы пар слов с экспертными оценками близости. Подробный обзор представлен в (Turney, Pantel, 2010).

Недавно большую популярность получили модели *обучаемых* векторных представлений слов, в частности, семейство моделей word2vec (Mikolov и др., 2013). 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 (Levy и др., 2015). Недавно предложенная модель GloVe (Pennington и др., 2014) 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 идеями в данной области. При этом важным недостатком является отсутствие интерпретируемости компонент построенных векторов.

Вероятностное тематическое моделирование развивалось параллельно с данными подходами, начиная с модели вероятностного латентного семантического анализа (Probabilistic Latent Semantic Analysis, PLSA), которая была предложена в 1999 году (Hofmann, 1999). 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 популярной тематической моделью является латентное размещение Дирихле (Latent Dirichlet Allocation, LDA), в которой дополнительно предполагается, что параметры модели имеют априорное распределение Дирихле (Blei и др., 2003). Эта модель позиционируется 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 (Masada и др., 2008). Позднее были построены сотни расширений LDA, и предложены алгоритмы их обучения в рамках байесовского подхода (Daud и др., 2010; Blei, 2012). 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 метаданных, отказ от гипотезы мешка слов.

Альтернативный подход аддитивной регуляризации тематических моделей (АРТМ) предлагается в работе (Воронцов, 2014) 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.

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 АРТМ. 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 сущностей различных *модальностей* (авторы, даты и другие мета-данные документов).

Научная новизна. В данной работе объединяются преимущества вероятностного тематического моделирования и моделей векторного представления слов на основе их совместной встречаемости. Это позволяет строить векторное пространство с интерпретируемыми размерностями, с помощью которого успешно решается задача определения семантической близости слов или сегментов текста. Разрабатывается подход аддитивной регуляризации тематических моделей, позволяющий легко встраивать новые требования, мотивированные лингвисти-

ческими предположениями или специфичными свойствами конечных приложений. Удается добиться повышения разреженности, различности и интерпретируемости векторных представлений текста, а также построить векторные представления мета-данных документов.

Теоретическая и практическая значимость. Развивается подход аддитивной регуляризации тематических моделей, и предлагается комбинация регуляризаторов, 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 APTM. 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:

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

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

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

4. 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. Семинар Ассоциации по компьютерной лингвистики BlackBoxNLP: Analyzing and interpreting neural networks for NLP, октябрь 2018, Брюссель, Бельгия (стендовый доклад).

2. 7-ая международная конференция по анализу изображений, социальных сетей и текстов (AIST), июль 2018, Москва (устное выступление).

3. Семинар группы Томаса Хофманна, Высшая Техническая Школа Цюриха (ETH Zurich), ноябрь 2017, Цюрих (приглашенный доклад).

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

5. Семинар Ассоциации по компьютерной лингвистики RepL4NLP: 2nd Workshop on Representation Learning for NLP, август 2017, Ванкувер, Канада (стендовый доклад).

6. Семинар группы Криса Биманна по языковым технологиям, Технический Университет Дармштадта (TU Darmstadt), июль 2016, Дармштадт, Германия (приглашенный доклад).

7. Научный семинар по анализу текстов в компании Google, июнь 2016, Цюрих, Швейцария (приглашенный доклад).

8. Конференция школы Яндекса Machine Learning: Prospects and Applications, октябрь 2015, Берлин, Германия (стендовый доклад).

9. Научный семинар в компании Microsoft Research Cambridge, апрель 2015, Кэмбридж, Великобритания (приглашенный доклад).

10. 3-ий международный симпозиум On Learning And Data Sciences (SLDS), апрель 2015, Лондон, Великобритания (устное выступление).

11. Российская летняя школа по информационному поиску (RuSSIR), август 2014, Нижний Новгород (стендовый доклад).

12. Международная конференция по компьютерной лингвистике «Диалог», июнь 2014, Москва (устное выступление).

13. XXI Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов-2014», апрель 2014, Москва (устное выступление).

14. 16-ая Всероссийская конференция «Математические методы распознавания образов» (ММРО), сентябрь 2013, Казань (устное выступление).

15. 35-ая Европейская конференция по информационному поиску (ECIR), март 2013, Москва (стендовый доклад).

Публикации. Результаты диссертации опубликованы в 12 печатных работах, из них 7 в изданиях, рекомендованных ВАК для публикации основных научных результатов диссертаций на соискание ученой степени кандидата наук. Работы [1–6] индексируются в базе международного цитирований Scopus [1–6], работа [7] опубликована в русскоязычном журнале, входящем в перечень ВАК, работа [8] опубликована в рецензируемом научном журнале, работы [9–12] являются тезисами докладов.

Личный вклад автора. Подход аддитивной регуляризации тематических моделей разрабатывался в соавторстве с Воронцовым К.В. [1, 4–6]. Основные положения, выносимые на защиту, являются персональным вкладом автора в опубликованные работы. Результаты по комбинированию тематического моделирования с моделями дистрибутивной семантики получены автором лично, за исключением некоторых экспериментов, проведенных совместно с Поповым А.С. [3].

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.

Содержание работы

Во Введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, представлены выносимые на защиту научные положения.

В первой главе приводится обзор основных принципов дистрибутивной семантики. *Дистрибутивная семантика (distributional semantics)* изучает способы определения семантической близости слов на основе их распределения в большом корпусе текстов. Рассматривается общая схема обработки текста для получения оценок близости слов, и подробно изучается ее ключевой компонент – математические методы построения низкоразмерных векторов слов. В частности, приводится вывод модели неотрицательных разреженных представлений NNSE (Murphy и др., 2012), тематической модели коротких текстов WNTM (Zuo и др., 2014), модели глобальных векторов GloVe (Pennington и др., 2014), моделей CBOW,

Skip-Gram и Skip-Gram Negative Sampling (SGNS) (Mikolov и др., 2013).

Все методы излагаются в едином формализме и в большинстве случаев могут трактоваться как матричное разложение вида:

$$F^{W\times W} \approx \Phi^{W\times T} \cdot \Theta^{T\times W}, \tag{1}$$

где матрица F содержит статистики встречаемости слов в контекстах, а матрицы Φ и Θ содержат параметры модели. Строки матрицы Φ являются векторными представлениями слов, а столбцы матрицы Θ – векторными представлениями контекстов. Рассматривается симметричный случай, при котором словарь слов и контекстов совпадает и имеет размер W. Размерность векторных представлений T является гиперпараметром модели и обычно принимает значение порядка 100.

Методы построения векторных представлений различаются типом подсчитываемых статистик F, оптимизируемым функционалом при низкоранговом матричном разложении, дополнительными ограничениями на параметры, методом оптимизации. При систематичном анализе становится ясно, что методы, пришедшие из различных областей (языковое моделирование, тематическое моделирование, глубокие нейронные сети) обладают схожей структурой. Это понимание позволяет прийти к гибридным подходам пятой главы.

Вторая глава содержит обзор классических тематических моделей и алгоритмов их обучения. Введем некоторые обозначения.

Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) всех употребляемых в них слов, T — множество тем. Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, \ldots, w_{n_d})$ из словаря W. Предполагается, что существует конечное множество тем T, и каждое употребление слова w в каждом документе d связано с некоторой (скрытой) темой $t \in T$. Таким образом, 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 вероятностным распределениям $p(t \mid d)$ и $p(w \mid t)$:

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

Построение тематической модели — это обратная задача: по известной коллекции D

требуется восстановить породившие её $p(t\,|\,d)$ и $p(w\,|\,t)$. Обычно число тем $|T|$ много меньше $|D|$ и $|W|$, и задача сводится к поиску приближённого представления заданной матрицы частот

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

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

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

Матрицы F, Φ, Θ являются *стохастическими*, то есть имеют неотрицательные нормированные столбцы, представляющие дискретные распределения.

В *вероятностном латентном семантическом анализе* PLSA (Hofmann, 1999) для построения модели (2) максимизируется логарифм правдоподобия при ограничениях нормировки и неотрицательности:

$$L(\Phi,\Theta) = \ln\prod_{d\in D}\prod_{w\in d} p(w\,|\,d)^{n_{dw}} = \sum_{d\in D}\sum_{w\in d} n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}; \tag{3}$$

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

Модель латентного размещения Дирихле LDA (Blei и др., 2003) является расширением модели PLSA и делает дополнительное предположение о том, что столбцы матриц Φ, Θ имеют априорное распределение Дирихле.

Для обучения модели LDA используются методы байесовского подхода, при этом их детали в литературе по тематическому моделированию часто опускаются. В данной работе приводится описание EM-алгоритма в общем виде (Dempster и др., 1977), его применение для максимизации правдоподобия в модели PLSA и максимизации апостериорной вероятности в модели LDA. Для модели LDA рассматриваются два альтернативных способа обучения: вариационный вывод и сэмплирование Гиббса. Показывается взаимосвязь формул вариационного вывода в модели LDA с формулами E-шага обучения PLSA. Далее в работе рассматриваются ограничения байесовского подхода и предлагается альтернативный метод аддитивной регуляризации тематических моделей.

В третьей главе описанные схемы обучения моделей PLSA и LDA сопоставляются на уровне алгоритмов. Вводится 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.

Наиболее распространённым внутренним критерием качества тематической модели является *перплексия (perplexity)*. Это мера несоответствия или «удивлённости» модели $p(w \mid d)$ словам w, наблюдаемым в документах d. Перплексия определяется через логарифм правдоподобия (чем меньше, тем лучше) и подсчитывается по контрольной выборке документов.

В экспериментах на двух текстовых коллекциях были получены следующие выводы.

1. Робастные алгоритмы с разреживанием являются лучшими по критерию контрольной перплексии. Такие модели 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. Сэмплирование Гиббса может интерпретироваться как эвристика разреживания распределения тем на E-шаге EM-алгоритма и использоваться не только в модели LDA, но и в модели PLSA. Вместо него может быть также использована предлагаемая эвристика *экономного сэмплирования*.

4. Принудительное разреживание в робастных моделях PLSA путём обнуления небольшой доли наименьших вероятностей позволяет получать до 99% нулей в распределениях без ухудшения контрольной перплексии.

Результаты данной главы опубликованы в работах [2, 7], [8].

В четвертой главе эти результаты обобщаются в рамках подхода аддитивной регуляризации тематических моделей (АРТМ) (Воронцов, 2014). Это приложение классической теории регуляризации некорректно поставленных задач (Тихонов и Арсенин, 1977) 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, разрабатывается модель предметных и фоновых тем, позволяющая разделить специфичные термины предметных областей и фоновую лексику.

В разделе 4.1 вводится подход аддитивной регуляризации тематических моделей. Логарифм правдоподобия (3) в модели PLSA 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 преобразования S в EM-подобных алгоритмах никак не контролируется и зависит от случайного начального приближения. Поэтому наряду с правдоподобием (3) предлагается максимизировать r дополнительных кри-

териев $R_i(\Phi, \Theta)$, $i = 1, \ldots, r$, называемых *регуляризаторами*:

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta), \quad L(\Phi, \Theta) + R(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}; \tag{5}$$

$$\sum_{w \in W} \phi_{wt} = 1, \quad \phi_{wt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geqslant 0, \tag{6}$$

где τ_i — неотрицательные *коэффициенты регуляризации*, отвечающие за баланс требований в задаче многокритериальной оптимизации. В работе [5] доказана теорема о необходимых условиях локального экстремума задачи (5), (6), позволяющая обучать данную модель с помощью регуляризованного EM-алгоритма.

В разделе 4.2 рассматривается применение подхода APTM для повышения интерпретируемости тем [4]. 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. В данной работе предлагается новый подход к формализации данного понятия. Предполагается, что интерпретируемая тема должна содержать *лексическое ядро (kernel)* — множество слов, характерных для определённой предметной области, которые с большой вероятностью употребляются в данной теме и практически не употребляются в других темах. 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 (Chemudugunta и др., 2006; Потапенко и Воронцов, 2013), в которых использовалось только одно фоновое распределение.

Для обеспечения описываемой структуры матриц Φ и Θ предлагается комбинация из пяти регуляризаторов: сглаживание фоновых тем в матрицах Φ и Θ, разреживание предметных тем в матрицах Φ и Θ, и декоррелирование предметных тем в матрице Φ:

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

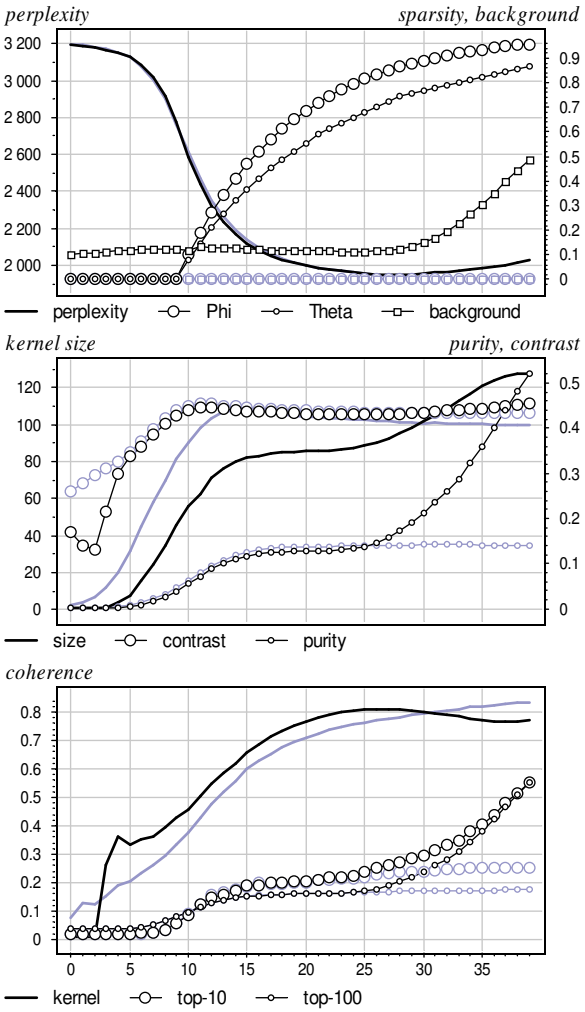

Рис. 1. Серый: PLSA. Чёрный: сглаживание, разреживание. Увеличивается разреженность (sparsity), чистота ядер тем (purity), размер ядер (kernel size), доля фоновых слов (background) при небольшом ухудшении перплексии (perplexity).

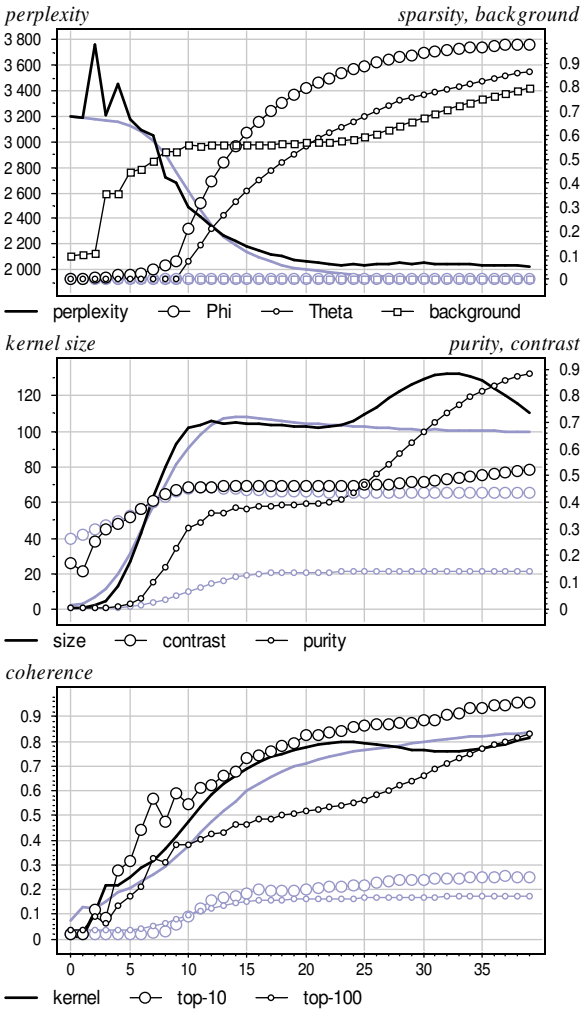

Рис. 2. Серый: PLSA. Чёрный: сглаживание, разреживание, декоррелирование. Улучшается когерентность тем (coherence), подсчитанная по 10 и 100 наиболее вероятным словам в темах, а также контастность ядер тем (contrast).

В качестве фоновых распределений α, β можно брать равномерные распределения. Коэффициенты регуляризации α_0, α_1, β_0, β_1, γ отвечают за баланс требований к модели и являются настраиваемыми гипепараметрами.

В проведенных экспериментах (Рис. 1, 2) помимо перплексии оцениваются критерии разреженности и интерпретируемости тематической модели. Общепринятой численной оценкой интерпретируемости, не требующей привлечения асессоров, является когерентность (Mimno и др., 2011; Newman и др., 2010). В данной работе когерентность оценивается по спискам наиболее вероятных слов в темах, а также по ядрам тем. Также вводятся новые меры интерпретируемости, основанные на понятии ядра темы: размер, чистота и контрастность ядер.

Основной вывод экспериментов заключается в том, что комбинирование регуляризаторов позволяет улучшить все критерии качества при незначительном ухудшении перплексии. Разреживание обнуляет до 96% элементов матрицы Φ и до 87% элементов матрицы Θ. Декорреляция повышает когерентность тем. Сглаживание фоновых тем помогает очистить предметные темы от слов общей лексики. Также повышаются критерии чистоты и контрастности ядер тем. Все эти улучшения сопровождаются незначительной потерей перплексии, что согласуется с выводами из (Chang и др., 2009) о том, что модели, имеющие лучшую перплексию, часто демонстрируют худшую интерпретируемость. В основном тексте работы приведено детальное описание всех используемых критериев качества, а также показаны примеры тем для стандартной модели PLSA и предложенной модели ARTM.

В разделе 4.3 предлагается регуляризатор разреживания распределения тем в коллекции $p(t) = \sum_d p(d)\theta_{td}$ для постепенного отбора тем. Максимизируется дивергенциия Кульбака-Лейблера между $p(t)$ и равномерным распределением на множестве тем:

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

В экспериментах на реальных данных демонстрируется возможность встраивания нового регуляризатора в рассмотренную ранее модель с разреженными и различными предметными темами, а также вырабатываются рекомендации по настраиванию коэффициентов регуляризации [6].

В пятой главе предлагается алгоритм построения вероятностных тематических представлений слов (Probabilistic Word Embeddings, PWE), 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. Материалы данной главы опубликованы в [3].

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:

$$p(w_j|w_i) = \sum_{t \in T} p(w_j|t)p(t|w_i) = \sum_{t \in T} \phi_{w_j t}\theta_{t w_i} = \left\langle \phi_{w_j}, \theta_{w_i} \right\rangle, \tag{7}$$

где $i = 1, \ldots, N$ индексирует позиции слов в корпусе, H_i содержит левые и правые контексты позиции i, ϕ_{w_j} — вектор слова w_j, θ_{w_i} — вектор слова-контекста w_i.

Сделаем предположения о независимости слов внутри каждой окрестности, а также о независимости окрестностей. Тогда можно записать следующую задачу максимизации регу-

ляризованного логарифма правдоподобия:

$$\sum_{i=1}^{N}\sum_{j\in H_i}\ln\sum_{t\in T}\phi_{w_j t}\theta_{tw_i} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \tag{8}$$

$$\sum_{u\in W}\phi_{ut} = 1, \quad \phi_{ut} \geqslant 0; \quad \sum_{t\in T}\theta_{tv} = 1, \quad \theta_{tv} \geqslant 0. \tag{9}$$

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:

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

Если $x_i \leqslant 0$ для всех $i \in I$, то результатом оператора norm по определению считается нулевой вектор. Обозначим через n_{vu} агрегированный счетчик совместной встречаемости слов в локальных окрестностях H_i, $i = 1, \ldots, N$. **В разделе 5.1** доказана следующая теорема.

Теорема 3. *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Тогда точка (Φ,Θ) локального экстремума задачи (8)-(9) удовлетворяет системе уравнений со вспомогательными переменными $p_{tvu} = p(t|v,u)$:*

$$p_{tvu} = \frac{\phi_{ut}\theta_{tv}}{\sum\limits_{s\in T}\phi_{us}\theta_{sv}}; \tag{10}$$

$$\phi_{ut} = \underset{u\in W}{\operatorname{norm}}\left(n_{ut} + \phi_{ut}\frac{\partial R}{\partial \phi_{ut}}\right); \quad n_{ut} = \sum_{v\in W} n_{vu}p_{tvu}; \tag{11}$$

$$\theta_{tv} = \underset{t\in T}{\operatorname{norm}}\left(n_{tv} + \theta_{tv}\frac{\partial R}{\partial \theta_{tv}}\right); \quad n_{tv} = \sum_{u\in W} n_{vu}p_{tvu}, \tag{12}$$

за исключением нулевых столбцов Φ, Θ в решении данной системы.

Решение системы уравнений (10)–(12) методом простых итераций соответствует регуляризованному EM-алгоритму. Нулевые столбцы матриц Φ, Θ в решении соответствуют *вырожденным* темам и документам, которые исключаются из модели. На практике это происходит редко и может говорить о необходимости понижения коэффициентов регуляризации.

В данной работе предлагается онлайновая версия EM-алгоритма, позволяющая избегать хранения матрицы Θ, а также сокращать число проходов по коллекции за счет более частого обновления параметров Φ.

Устанавливается связь предложенной модели с другими моделями векторных представлений (Таблица 1). Ключевое отличие модели PWE от популярной модели Skip-Gram заключается в использовании смеси распределений (7) вместо нормировки скалярного произведения с помощью операции softmax:

$$p(u|v) = \operatorname{softmax}\langle\phi_u, \theta_v\rangle = \frac{\exp\langle\phi_u, \theta_v\rangle}{\sum_w \exp\langle\phi_w, \theta_v\rangle}. \tag{13}$$

Таблица 1. Сопоставление походов по типу данных (слова-слова или слова-документы) и типу вероятностной модели (softmax или вероятностная смесь распределений).

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

В разделах 5.2 и 5.3 проводятся эксперименты на коллекции англоязычных статей Википедии. Для получения оценок близости слов сравниваются несколько способов, из которых наилучшим оказывается *скалярное произведение* векторов слов, составленных из условных вероятностей $\phi^B_{wt} = p(t|w)$. Качество определения семантической близости оценивается на стандартных тестовых наборах пар слов с экспертными оценками близости (WordSim-353, SimLex-999 и др.). В проведенных экспериментах демонстрируется качество, сопоставимое с моделью SGNS, а также ряд преимуществ предлагаемого подхода:

1. интерпретируемость компонент векторных представлений слов (Рис. 3);
2. высокая разреженность представлений (до 93%) без ухудшения качества модели;
3. подключение дополнительных регуляризаторов для учета специфичных требований.

Примером специфичных требований к модели может быть учет дополнительных данных, таких как время, категория, автор, или любая другая информация, доступная для документа. Будем называть такую мета-информацию дополнительными *модальностями*, при этом базовой модальностью будем считать текст.

В разделе 5.4 предлагается алгоритм построения единого векторного пространства для токенов различных модальностей, основанный на подходе мультимодального тематического моделирования (Vorontsov и др., 2015). Для этого рассматривается следующая оптимизационная задача:

$$\sum_{m\in M}\lambda_m \sum_{v\in W^0}\sum_{u\in W^m} n_{vu}\ln\sum_{t\in T}\phi_{ut}\theta_{tv} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \tag{14}$$

$$\forall u,t \quad \phi_{ut} \geqslant 0; \quad \sum_{u\in W^m}\phi_{ut} = 1, \forall m \in M; \tag{15}$$

$$\forall t,v \quad \theta_{tv} \geqslant 0; \quad \sum_{t\in T}\theta_{tv} = 1. \tag{16}$$

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

В такой модели матрица параметров Φ разбивается на блоки по словарям различных модальностей, и нормировка производится в рамках каждого отдельного блока. Таким обра-

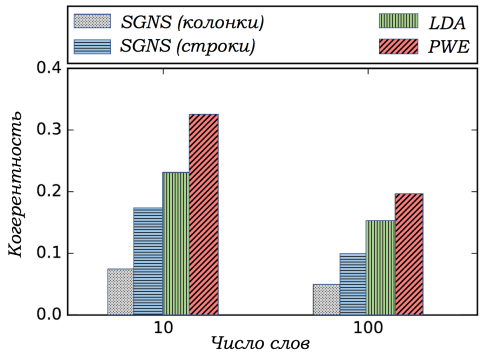

Рис. 3. Количественная оценка интерпретируемости (когерентность по спискам топ-слов в компонентах).

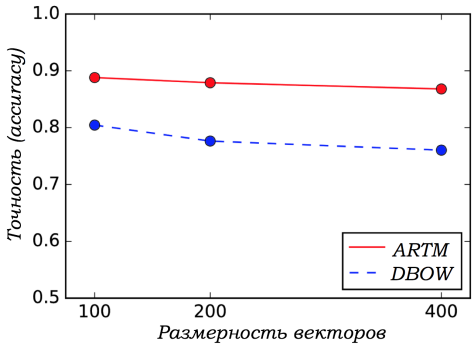

Рис. 4. Качество предсказания близости в тройках статей ArXiv для нескольких размерностей векторного пространства.

зом, каждая тема описывается несколькими альтернативными распределениями. Матрица Θ сохраняет прежнюю размерность и интерпретацию. Учет дополнительных модальностей не противоречит введению регуляризаторов разреживания и любых других. Обучение производится с помощью модификации EM-алгоритма, что обосновывается следующей теоремой.

Теорема 4. *Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Тогда точка (Φ, Θ) локального экстремума задачи (14)-(16) удовлетворяет системе уравнений со вспомогательными переменными $p_{tvu} = p(t|v, u)$:*

$$p_{tvu} = \underset{t \in T}{\operatorname{norm}}\big(\phi_{ut}\theta_{tv}\big); \tag{17}$$

$$\phi_{ut} = \underset{u \in W^m}{\operatorname{norm}}\left(n_{ut} + \phi_{ut}\frac{\partial R}{\partial \phi_{ut}}\right); \quad n_{ut} = \sum_{v \in W^0} n_{vu}p_{tvu}; \tag{18}$$

$$\theta_{tv} = \underset{t \in T}{\operatorname{norm}}\left(n_{tv} + \theta_{tv}\frac{\partial R}{\partial \theta_{tv}}\right); \quad n_{tv} = \sum_{m \in M}\sum_{u \in W^m} \lambda_m n_{vu}p_{tvu}. \tag{19}$$

за исключением нулевых столбцов Φ, Θ в решении данной системы.

В эксперименте на мультимодальной коллекции русскоязычных новостей Lenta.ru было продемонстрировано улучшение качества определения семантической близости слов при включении в модель модальностей времени и категорий (Таблица 2). Более того, даже базовая версия предлагаемой модели PWE превзошла подходы SGNS и CBOW, которые стали стандартными инструментами для таких задач. Детали предобработки коллекции, обучения моделей и тестовых выборок (столбцы таблицы) представлены в тексте работы.

Предлагаемый подход позволяет оценивать не только близости слов, но и близости сущностей различных модальностей. Например, ближайшими словами к дате *2016-02-29* (вручение «Оскара») оказываются слова *статуэтка, кинонаграда, номинироваться*.

Таблица 2. Корреляция Спирмена на задачах близости. Модели обучены на русскоязычном мультимодальном корпусе Lenta.ru. Учет меток времени и категорий улучшает качество векторов слов.

	WordSim-Sim	WordSim-Rel	MC	RG	HJ	SimLex-999
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**	**0.580**	**0.607**	0.584	**0.611**	0.144

В разделе 5.5 обсуждается связь предложенной модели с тематическими моделями коротких текстов Word Network Topic Model, WNTM (Zuo и др., 2014) и Biterm Topic Model, BTM (Yan и др., 2013). Доказывается теорема об эквивалентности моделей WNTM и BTM при определенной инициализации параметров.

Введем обозначение для матрицы условных вероятностей, полученных из матрицы Φ по формуле Байеса:

$$\Phi^B = (\phi^B_{wt}), \quad \phi^B_{wt} = p(t|w) = \frac{p(w|t)p(t)}{p(w)} = \frac{\phi_{wt}p(t)}{p(w)} \tag{20}$$

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

Данное утверждение подтверждается в эксперименте. Демонстрируется одинаковое качество моделей WNTM и BTM, несмотря на сокращение числа параметров вдвое. Аналогичное связывание векторных представлений слов и контекстов применяется в литературе для языковых моделей (Press и др., 2016; Inan и др., 2016). Однако в стандартных моделях векторных представлений слов (SGNS, GloVe) этого не происходит.

В разделе 5.6 предлагается алгоритм построения тематических векторных представлений текстовых фрагментов, в частности, отдельных предложений или целых документов. Ставится задача определения семантической близости документов. Качество оценивается на контрольной выборке триплетов статей arXiv, для которых эти близости известны (Dai и др, 2015). В проведенном эксперименте (Рис. 4) подход аддитивной регуляризации тематических моделей (ARTM) превосходит модель doc2vec (DBOW), являющуюся стандартным расширением модели word2vec для задачи определения близости документов.

В заключении перечисляются основные результаты работы.

Предложено семейство EM-алгоритмов, включающее как известные, так и новые алгоритмы обучения тематических моделей. Исследованы опции разреживания, робастности,

регуляризации и сэмплирования. В рамках подхода аддитивной регуляризации тематических моделей предложен набор из пяти регуляризаторов, повышающий интерпретируемость, разреженность и различность предметных тем модели. Предложен разреживающий регуляризатор отбора тем. Выработаны рекомендации по настраиванию коэффициентов регуляризации.

Предложен алгоритм построения тематических представлений PWE, которые позволяют определять семантическую близость слов и при этом являются интерпретируемыми. Продемонстрировано применение подхода аддитивной регуляризации для повышения разреженности представлений слов, а также для построения единого векторного пространства для сущностей различных модальностей (слова, время, категории). Подход обобщен на случай сегментированного текста. 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 (Bojanowski и др., 2017). Другое направление недавних исследований связано с обучением контексто-зависимых представлений слов. Модель ELMo (Peters и др., 2018) 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.

Публикации автора по теме диссертации

Список публикаций в изданиях, рекомендованных ВАК:

1. *Vorontsov K. V., Potapenko A. A.* Additive Regularization of Topic Models // *Machine Learning Journal*. — 2015. — Vol. 101. — Pp. 303–323.
2. *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.
3. *Potapenko A., Popov A., Vorontsov K.* Interpretable Probabilistic Embeddings: 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.

4. *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.
5. *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.
6. *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.
7. *Воронцов К. В., Потапенко А. А.* Регуляризация, робастность и разреженность вероятностных тематических моделей // *Компьютерные исследования и моделирование*. — 2012. — Т. 4, № 4. — С. 693–706.

Список публикаций в других изданиях:

8. *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.
9. *Потапенко А. А.* Разреживание вероятностных тематических моделей // Математические методы распознавания образов: 16-ая Всеросс. конф.: Докл. — МАКС Пресс, 2013. — P. 89.
10. *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.
11. *Воронцов К. В., Потапенко А. А.* Робастные разреженные вероятностные тематические модели // Интеллектуализация обработки информации (ИОИ-2012): Докл. — Торус Пресс, 2012. — Pp. 605–608.
12. 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. — ACL, 2018. — Pp. 200–210.

Подписано в печать « » декабря 2018 г.

Формат 60x90/16. Тираж 100 экз.

ООО «Армсинг»

ИНН/КПП: 7705461204 / 770201001

Адрес: г. Москва, Проспект Мира, д. 47, стр. 1.

Тел.: +7(499) 653-63-83.