Построение полного набора тем вероятностных тематических моделей

Сухарева А.В., Воронцов К.В.

Интерпретируемость, линейное увеличение сложности с ростом данных, масштабируемость сделали тематическое моделирование одним из наиболее популярных инструментов статистического анализа текстов. Однако есть и ряд недостатков, вызванных зависимостью решения от инициализации. Известно, что построение тематической модели сводится к решению некорректно поставленной задачи неотрицательного матричного разложения. Множество её решений в общем случае бесконечно. Всякий раз модель находит локальный экстремум. Многократное обучение модели по одной и той же коллекции может приводить к обнаружению всё новых и новых тем. На практике часто требуется определить все темы корпуса. Для решения этой задачи в статье предложен и исследован новый алгоритм нахождения полного набора тем, который основан на построении выпуклой оболочки. Экспериментально показано, что за конечное число моделей можно построить базис тем. Правдоподобие базиса тем выше, чем одной модели с аналогичным числом тем. Сравнение базисов моделей LDA (latent Dirichlet allocation) и ARTM (additive regularization for topic modeling) позволяет сделать вывод, что темы наборов совпадают с высокой точностью.

Ключевые слова: вероятностное тематическое моделирование, устойчивость тематических моделей, полный набор тем тематических моделей, латентное размещение Дирихле, LDA, регуляризация, ARTM, BigARTM.

1. Введение

Вероятностные тематические модели являются статистическими алгоритмами, предназначенными для обнаружения абстрактных тем, которые содержатся в большом и неструктурированном наборе документов. Тематические модели наиболее известны как инструмент для интеллек-

туального анализа текста. Они также имеют приложения в биоинформатике, анализе изображений и социальных сетей [1].

Построение тематической модели сводится к решению некорректно поставленной задачи неотрицательного матричного разложения. Множество её решений в общем случае бесконечно. Это приводит к неустойчивости вычислительных методов и зависимости решения от случайного начального приближения. Многократное построение модели по одной и той же коллекции может приводить к нахождению всё новых и новых тем. Несмотря на важность требования устойчивости в задачах компьютерной лингвистики и информационного поиска, проблема до сих пор относительно мало изучена. В литературе определено понятие устойчивости [2], предлагаются меры устойчивости [3, 4, 5]. Также в работах рассматриваются модели, повышающие устойчивость [6, 7].

В данной статье ставится проблема полноты: можно ли найти все темы корпуса, сколько запусков моделирования необходимо для нахождения полного набора тем, возможно ли сократить число запусков с помощью регуляризации. Описанная постановка в литературе не рассматривалась. Актуальность проблемы обусловлена большим числом прикладных задач анализа текстов, в которых требуется как можно полнее определить тематический состав коллекции документов.

Статья организована следующим образом. В разделе 2 мы вводим основные обозначения и терминологию, кратко описываем тематические модели pLSA, LDA и модели подхода ARTM. В разделе 3 определяется понятие базиса тем множества матриц тематических моделей. Далее, предлагается алгоритм для построения полного набора тем на основе выпуклой оболочки. Эмпирические результаты построения полного набора тем обсуждаются в разделе 4. Наконец, в заключении представлены наши выводы.

2. Вероятностные тематические модели

Введем следующие обозначения: $D = \{d_1, \ldots, d_{|D|}\}$ — коллекция текстовых документов, $W = \{w_1, \ldots, w_{|W|}\}$ — словарь термов, встретившихся в них, $T = \{t_1, \ldots, t_{|T|}\}$ — конечное множество тем. В качестве термов могут использоваться слова, n-граммы, коллокации, именованные сущности и т.д. Число тем является гиперпараметром и задается заранее ($|T| \ll |D|$). Каждое вхождение терма $w \in W$ в документ $d \in D$ связано с некоторой темой $t \in T$. Термы и документы являются наблюдаемыми переменными, темы — латентными (скрытыми). Требуется

определить, к каким темам относится каждый документ и какие термы образуют каждую тему.

2.1. Модель pLSA

Пусть коллекция документов представляет собой последовательность пар «документ-слово» (d, w). Подход pLSA (probabilistic latent semantic analysis) [8] моделирует вероятность $p(w|d)$ появления термов w в документах d как смесь условно независимых мультиномиальных распределений. Компоненты смеси можно рассматривать как представления «тем» t. Наблюдаемые пары (d, w) встречаются независимо, что соответствует представлению документов в виде «мешка слов». Появление терма w зависит только от темы t и не зависит от документа d. Каждое слово генерируется из одной темы. В этом случае тематическая модель появления слов в документах выглядит следующим образом:

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

где w — терм, t — тема, d — документ коллекции.

При построении тематической модели требуется оценить по известной коллекции D параметры модели $\varphi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$. Ставится задача максимизации логарифма правдоподобия:

$$L(\Phi, \Theta) = \sum_{d\in D}\sum_{w\in W} n_{dw} \log p(d, w) = \sum_{d\in D}\sum_{w\in W} n_{dw} \log \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \tag{1}$$

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

где $\Phi = (\phi_{wt})_{W\times T}$, $\Theta = (\theta_{td})_{T\times D}$, n_{dw} — число вхождений терма w в документ d.

Задача решается с помощью EM-алгоритма. Вероятностные тематические модели находят локальный максимум логарифма правдоподобия (1). Известно, что pLSA не моделирует процесс выбора документов, распределение $p(d)$, это частично генеративная модель. Чтобы полностью реализовать генеративный процесс на уровне документа и усовершенствовать его для тем, было предложено скрытое распределение Дирихле (latent Dirichlet allocation, LDA) [9].

Следует отметить, что модель pLSA легко переобучается из-за необходимости настраивать большое число параметров. Решением данной проблемы может стать небольшая модификация E-шага

EM-алгоритма (Tempered EM) [8], которая также позволяет ускорить обучение модели.

Другая трудность заключается в том, что алгоритм не разделяет тематические и фоновые компоненты смеси тем, не учитывает модальности данных (авторы, теги, картинки, ссылки и т.д.). Кроме того, модель pLSA не позволяет управлять разреженностью. Действительно, если в начале работы алгоритма $\phi_{wt} = 0$ или $\theta_{td} = 0$, то и после завершения работы алгоритма значения этих параметров останется равным 0. Эти проблемы могут быть решены с помощью регуляризации модели pLSA, представленной в общем подходе ARTM, который будет описан ниже.

2.2. Модель LDA

Латентное размещение Дирихле (latent Dirichlet allocation, LDA) [9] — широко известная генеративная вероятностная тематическая модель для дискретных данных, в частности текстов. Подобный подход схож с pLSA с той разницей, что в LDA накладываются дополнительные ограничения на вид распределений $\phi_t \sim Dir(\beta)$ и $\theta_d \sim Dir(\alpha)$. Это приводит к различным модификациям M-шага. Самая простая и популярная из них следующая:

$$\phi_{wt} \propto n_{wt} + \beta_w, \quad \theta_{td} \propto n_{td} + \alpha_t,$$

что совпадает с регуляризатором сглаживания (4) в подходе ARTM. В результате получается более эффективная модель, что подтверждается успешным применением в задачах классификации и коллаборативной фильтрации.

2.3. Подход ARTM

Распределения ϕ_{wt} и θ_{td}, полученные моделью pLSA, слабо разрежены, что нежелательно на практике. Хотелось бы, чтобы в модели каждый документ d представлялся небольшим числом тем t с большими вероятностями $p(t|d)$ и каждая тема t состояла из небольшого числа слов с высокими вероятностями $p(w|t)$. Такая стратегия повышает интерпретируемость модели, обеспечивает более компактное представление данных. Однако ничто в моделировании pLSA не поощряет такой механизм обучения. В последнее время предпринималось немало попыток построить более разреженные распределения. Рассмотрим один из предложенных подходов — аддитивную регуляризацию тематических моделей (additive regularization for topic modeling, ARTM) [10].

Основная идея ARTM заключается в том, чтобы обеспечить гибкий способ добавления некоторой дополнительной информации о задаче к оптимизируемой функции правдоподобия (1). Делается это с помощью взвешенной суммы критериев регуляризации R_i:

$$R(\Phi,\Theta)=\sum_i \tau_i R_i(\Phi,\Theta), \quad L(\Phi,\Theta)+R(\Phi,\Theta)\to\max_{\Phi,\Theta}, \tag{2}$$

где $\tau_i \geq 0$ — коэффициенты регуляризации, регулируют силу действия регуляризатора и настраиваются экспериментально.

Для обучения тематической модели применяется EM-алгоритм. На E-шаге, как и в модели pLSA, оценивается по формуле Байеса вероятность $p(t|d,w)$ и выражается n_{tdw} — число троек, в которых терм w документа d связан с темой t:

$$n_{tdw}=n_{dw}p(t|d,w), \quad p(t|d,w)=\frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}.$$

Введем оператор norm, который преобразует произвольный вектор $(x_i)_{i\in I}$ в вектор вероятностей дискретного распределения:

$$\underset{i\in I}{\operatorname{norm}}(x_i)=\frac{\max(0,x_i)}{\sum\limits_{j\in I}\max(0,x_j)}, \text{ для всех } i\in I,$$

если $x_i \leq 0$ для всех $i \in I$, то $\operatorname{norm}(x)=\mathbf{0}$.

Для решения сформулированной задачи многокритериальной оптимизации (2) достаточно модифицировать M-шаг:

$$\phi_{wt}=\underset{w\in W}{\operatorname{norm}}\left(n_{wt}+\phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \quad \theta_{td}=\underset{t\in T}{\operatorname{norm}}\left(n_{td}+\theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \tag{3}$$

$$n_{wt}=\sum_{d\in D} n_{tdw}, \quad n_{td}=\sum_{w\in d} n_{tdw},$$

где $R(\Phi,\Theta)$ — непрерывно дифференцируемая функция.

Различные регуляризаторы позволяют не только разреживать распределения, но и повышать согласованность, различность тем, отбирать темы, документы и термы, сглаживать распределения, если это необходимо, выполнять тематическую сегментацию, строить иерархические модели и т.д. Подход ARTM дает возможность комбинировать самостоятельные модели в одну общую модель путем преобразования M-шага.

В данной статье рассматривается минимальный набор регуляризаторов, который покрывает требования большинства задач тематического моделирования: регуляризаторы сглаживания, разреживания и декоррелирования.

Разреживающий и сглаживающий регуляризаторы предполагают, что модельные распределения ϕ_t и θ_d близки по дивергенции Кульбака-Лейблера к некоторым заданным распределениям $\beta = (\beta_w)_{w \in W}$ и $\alpha = (\alpha_t)_{t \in T}$.

- Разреживающий регуляризатор:

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

  По формулам (3) выражение для M-шага записывается следующим образом:

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

- Сглаживающий регуляризатор:

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

  Применение формул (3) дает тоже выражение для M-шага, что и самая простая модификация модели LDA:

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}}\big(n_{wt} + \beta_0 \beta_w\big), \quad \theta_{td} = \underset{t \in T}{\operatorname{norm}}\big(n_{td} + \alpha_0 \alpha_t\big), \tag{4}$$

  если в качестве векторов гиперпараметров взять дискретные распределения α и β, умноженные на коэффициенты регуляризации: $(\beta_0 \beta_w)_{w \in W}$, $(\alpha_0 \alpha_t)_{t \in T}$.

- Декорреляция тем как минимизация ковариаций между столбцами ϕ_t и ϕ_s матрицы Φ:

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

  где $\tau \geq 0$ — коэффициент регуляризации.

Формулы (3) М-шага в данном случае принимают вид:

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

Декоррелирующий регуляризатор стремится уменьшить пересечение между распределениями слов по темам ϕ_t, это повышает различность тем, что положительно влияет на интерпретируемость модели [11].

3. Полный набор тем

Построение вероятностной тематической модели является некорректно поставленной задачей стохастического матричного разложения. Множество ее решений в общем случае бесконечно:

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

где S — произвольная невырожденная матрица, при условии, что матрицы $(\Phi S), (S^{-1}\Theta)$ стохастические.

Каждый раз модель находит локальный экстремум, поэтому нельзя гарантировать, что были найдены все темы корпуса. Однако в этом пространстве можно построить базис векторов тем тематических моделей ϕ_t, состоящий из линейно независимых, хорошо интерпретируемых тем.

3.1. Алгоритм построения полного набора тем

Из неустойчивости тем следует проблема полноты набора тем, найденного тематической моделью. Возникают следующие вопросы:

- действительно ли темы новые или это комбинации предыдущих;
- можно ли найти все темы корпуса, сколько моделей для этого нужно построить.

Для ответов на эти вопросы рассмотрим алгоритм построения базиса тем моделей, отличающихся инициализацией. Все описанные выше модели — pLSA, ARTM и LDA — находят темы в пространстве распределений над словами. Каждое такое распределение можно рассматривать как точку в единичном $(|W| - 1)$-симплексе слов $\triangle$.

Определение. *Множество* $V = \{v_1, \ldots, v_m\} \subset \triangle$ *— базис тем множества матриц тематических моделей* $\Phi_1, \ldots, \Phi_n \subset \triangle$, *если* $\forall \phi \in \Phi_j$ *выполняется:*

$$\min_{v \in convV} \rho(\phi, v) \leq \varepsilon, \tag{5}$$

$$convV = \Big\{ v = \sum_{i=1}^{m} \alpha_i v_i \big| v_i \in V, \quad \sum_i \alpha_i = 1, \quad \alpha_i \geq 0 \Big\},$$

где $\rho(\phi, v)$ *— функция расстояния.*

Из определения следует, что базис V состоит из векторов тем $\phi \in \Phi_j, \quad j = \overline{1, n}$, для которых $\min\limits_{v \in convV} \rho(\phi, v) > \varepsilon$, именно они линейно независимы. Кроме того, решение каждой отдельной тематической модели локально оптимально, а значит его можно улучшить с помощью замены тем на близкие им в случае повышения правдоподобия (log-likelihood). На этом и основан предлагаемый жадный алгоритм для нахождения базиса тем.

Алгоритм состоит из чередования двух этапов. На первом этапе происходит замена тем с целью расширения имеющейся системы векторов тем и максимизации правдоподобия. Алгоритм итеративный, на каждой итерации для тем набора находятся схожие с некоторым порогом γ и выполняется замена, если она улучшает правдоподобие всего набора тем. На втором этапе происходит поиск темы для добавления в набор:

- включаются линейно независимые темы;
- исключаются выбросы (выбросами считались темы, которые имеют более двух коэффициентов $\alpha_i > 0.15$ в (5), что соответствовало несогласованным темам, зависимым нелинейно от тем полного набора);
- выбирается тема, максимально повышающая правдоподобие набора тем.

Таким образом, мы дополняем линейно независимую систему векторов до базиса. Описанные шаги повторяются до сходимости. Так как алгоритм итеративный, то после того, как алгоритм сошелся, нужно выполнить процедуру замены близких тем, используя обученные модели в обратном порядке.

Для решения оптимизационной задачи (5) использовался алгоритм SLSQP (Sequential Least SQuares Programming) [12], который реализован во многих популярных математических пакетах, в том числе и SciPy.

В результате экспериментов было установлено, что все темы корпуса можно найти за конечное число итераций алгоритма построения выпуклой оболочки тем тематических моделей. Регуляризатор декорреляции, применяемый при построении моделей ARTM, способствует нахождению более мелких тем, поэтому в полном наборе содержится больше тем, чем в базисе моделей LDA. На рис. 1, рис. 2 видно, что правдоподобие полного набора тем выше, чем одной модели с аналогичным числом тем. Замечено, что число тем в полном наборе зависит от степени разреженности матриц Ф. Базис более разреженных моделей содержит больше тем.

4. Эмпирические результаты

В этом разделе мы приводим результаты работы алгоритмов на реальных данных коллекции ПостНаука[1]. Коллекция ПостНаука — небольшой корпус текстов интернет-журнала ПостНаука, состоящий из научно-популярных статей о современной фундаментальной науке и учёных, которые её создают.

В экспериментах, описанных ниже, использовались теги к документам коллекции ПостНаука, размеченные экспертами. Всего было 930 тегов, которые обозначали общие и частные понятия, например, биология, клетка, эволюционная биология, ген, эукариоты, квантовая физика, Россия, человек и т.д. Каждый документ в среднем описывался 6 тегами. Данные были случайно разделены на обучающую и контрольную выборки в отношении 80% к 20%.

4.1. Моделирование документов

Для построения полного набора рассматривались модели LDA Gibbs sampling и ARTM с 20 темами. LDA[2] с параметрами $\eta = 0.01$, $\alpha = \frac{1}{20}$ обучалась 3500 итераций. При построении моделей ARTM использовалась следующая стратегия регуляризации: сначала включался декоррелирующий регуляризатор до сходимости модели по перплексии, затем он

[1] https://postnauka.ru/

[2] https://github.com/lda-project/lda

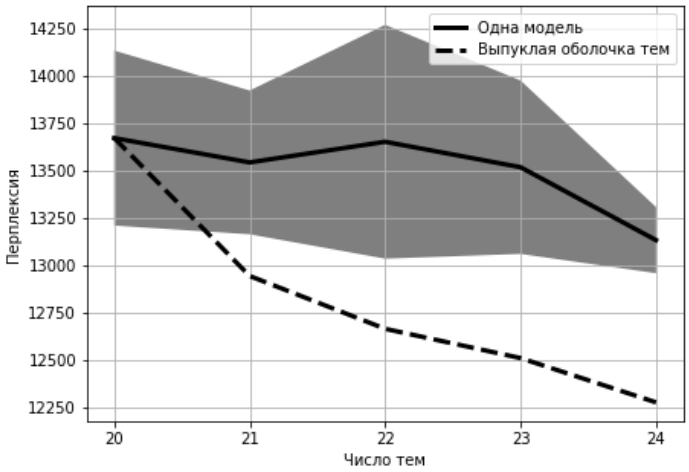

Рис. 1. Сравнение перплексии одной модели и полного набора тем моделей ARTM на тегах коллекции ПостНаука, параметр $\delta = 1.2, \gamma = 0.43$.

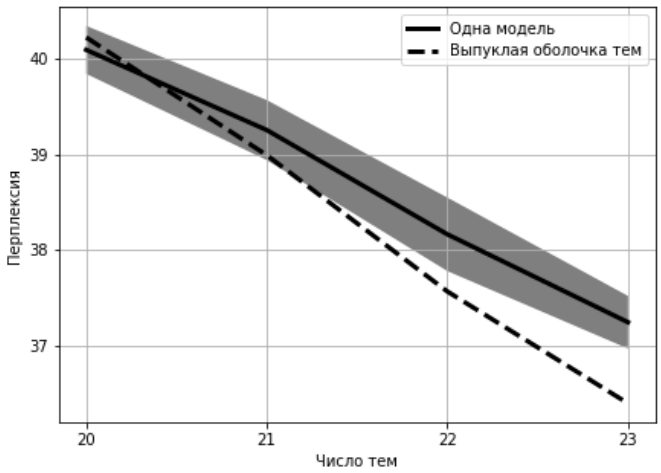

Рис. 2. Сравнение перплексии одной модели и полного набора тем моделей LDA (с использованием Gibbs Sampling) на тегах коллекции ПостНаука, параметр $\delta = 0.64, \gamma = 0.43$.

отключался, и применялись регуляризаторы разреживания матрицы Θ и сглаживания матрицы Φ. Во всех случаях матрицы Θ, Φ — случайные стохастические матрицы. Модели ARTM строились с помощью библиотеки BigARTM[3].

4.2. Полнота и интерпретируемость тематических моделей

Тематические модели находят локально экстремальное решение, которое не является полным. При этом на практике часто возникает задача найти все темы корпуса. Для решения данной задачи в статье был предложен алгоритм нахождения базиса тем.

В табл. 3, табл. 4 приводятся некоторые темы, найденные с помощью алгоритма построения полного набора тем. Алгоритм применялся для моделей LDA и ARTM. Процесс накопления тем базиса проиллюстрирован на рис. 1 и рис. 2. На графиках видно, что правдоподобие базиса тем выше, чем правдоподобие одной модели. Кроме того, было проведено сравнение одной модели и базиса тем по дополнительным критериям качества тематических моделей (см. табл. 1, табл. 2): критерию Р. Аруна [13], когерентности (автоматической меры интерпретируемости) [14] и критерию Цао Хуана [15]. На обучении полный набор тем моделей LDA лучше по всем критериям, чем одна модель. На контроле ухудшение только по когерентности. Базис моделей ARTM имеет более высокое правдоподобие на обучении и контроле, чем одна модель.

Полный набор тем ARTM позволяет более подробно описать корпус, чем базис моделей LDA. Например, тема о генетике в полном наборе моделей LDA содержит термы биология, ген, днк, генетика, а у моделей ARTM — молекулярная биология, клеточная биология, генетика, генная инженерия. Однако не все темы ARTM согласованные, даже в топе содержатся слова, не относящиеся к данной теме, например, русская философия, бактерии, народная культура, Русь. Еще один недостаток базиса ARTM заключается в том, что могут быть темы, которая не выделились в отдельные, а были размыты по нескольким темам базиса. Эти недостатки можно исправить с помощью дальнейшей регуляризации.

В экспериментах было построено два полных набора тем моделей LDA. В первом случае модели следовали в прямом порядке, во втором — в обратном. Темы базисов соотносились между собой с помощью венгерского алгоритма. Темы совпали с высокой точностью. Однако одна тема в первом базисе была разделена на две во втором: астрономия,

[3]https://github.com/bigartm/bigartm

астрофизика, космология, гравитация на астрономию, космос и физику, астрофизику. Базис моделей ARTM тоже разделяет эти две темы.

На рис. 3 и рис. 4 изображены плоские проекции[4] базисов ARTM и LDA. Видно, что в базисе моделей LDA все темы крупные и разделились на следующие кластеры: биология, физика, химия, лингвистика, экономика, психология, математика, история, философия, культура, социология. В полном наборе тем ARTM содержатся как крупные темы, так и мелкие, например, русская философия, судопроизводство и искусствоведение.

Критерии	На обучении		На контроле	
	Одна модель	Базис тем	Одна модель	Базис тем
Перплексия	13132.6	**12227.45**	12835.1	**12534.36**
Критерий Аруна	**52.15**	56.90	**9.95**	12.96
Когерентность	**-7.89**	-9.01	-13.50	**-12.75**
Критерий Хуана	**0.01**	0.02	**0.01**	0.02

Таблица 1. Сравнение результатов работы одной модели ARTM с 24 темами (бралось среднее значение по пяти моделям) и полного набора тем ARTM. Модели набора содержали 20 тем, в результате работы алгоритма было отобрано 24 темы. Жирным выделены оптимальные значения в сравнении.

Критерии	На обучении		На контроле	
	Одна модель	Базис тем	Одна модель	Базис тем
Перплексия	37.36	**36.39**	44.91	**43.55**
Критерий Аруна	100.93	**95.18**	31.84	**28.88**
Когерентность	-3.01	**-2.91**	**-5.70**	-6.44
Критерий Хуана	0.07	**0.06**	0.07	**0.06**

Таблица 2. Сравнение результатов работы одной модели LDA с 23 темами (бралось среднее значение по пяти моделям) и полного набора тем LDA с применением Gibbs sampling.

[4]https://github.com/bmabey/pyLDAvis

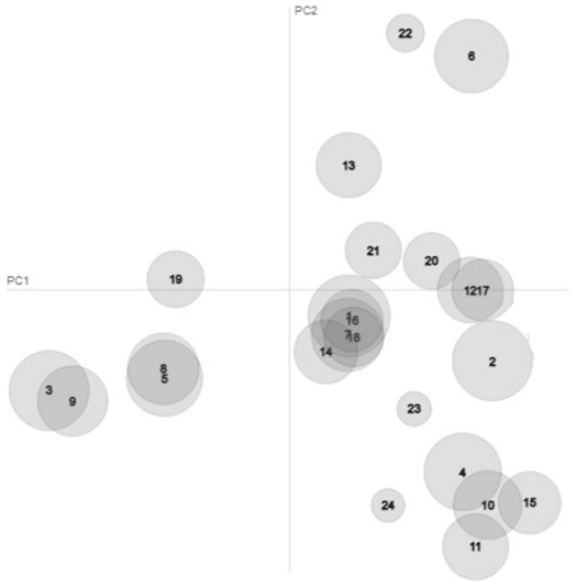

Рис. 3. Плоская проекция полного набора тем моделей ARTM, построенного на тегах коллекции ПостНаука.

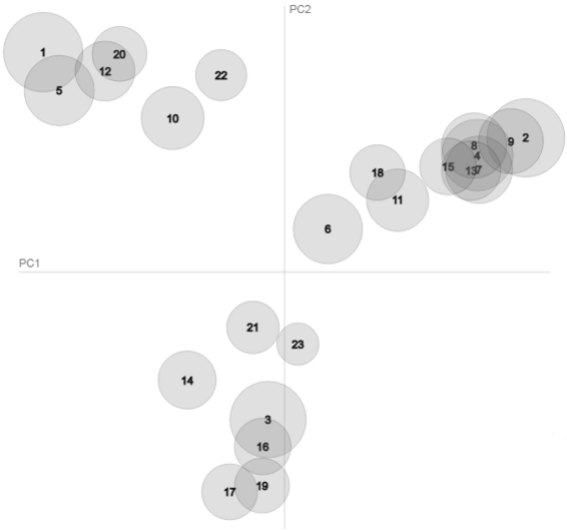

Рис. 4. Плоская проекция полного набора тем моделей LDA, построенного на тегах коллекции ПостНаука.

Базис тем, полученный в порядке $\{\Phi_i\}_{i=N}^{0}$	Базис тем, полученный в порядке $\{\Phi_i\}_{i=0}^{N}$
история, культура, религия, христианство, востоковедение, китай, ислам	история, культура, религия, христианство, востоковедение, ислам, археология
культура, общество, культурология, массовая культура, философия, кино, искусство	культура, культурология, массовая культура, философия, кино, общество, искусство
биология, ген, днк, геном, клетка, генетика, микробиология	биология, ген, днк, геном, генетика, клетка, белки
физика, квантовая физика, технологии, атом, квантовая механика, сверхпроводимость, квантовые технологи	физика, квантовая физика, технологии, оптика, квантовые технологии, квантовая механика, сверхпроводимость
медицина, биология, генетика, биомедицина, онкология, клетка, технологии	медицина, биология, генетика, биомедицина, онкология, клетка, стволовые клетки

Таблица 3. Сравнение двух полных наборов тем моделей LDA, построенных на тегах коллекции ПостНаука. В первом случае модели следовали в обратном порядке, во втором — в прямом.

5. Заключение

Данная работа посвящена изучению вопроса полноты тематических моделей. Проведен сравнительный анализ моделей pLSA, LDA и моделей подхода ARTM по критериям, связанным с исследуемой проблемой. Во всех моделях только часть тем оказалась устойчивой. Показано, что декоррелирующий регуляризатор ухудшает устойчивость модели и способствует нахождению более мелких тем.

Неустойчивость тематических моделей является известным фактом, однако связанная с ней проблема полноты до сих пор в литературе не изучалась. Для решения этой задачи в статье предложен новый алгоритм нахождения полного набора тем, основанный на построении выпуклой оболочки векторов тем тематических моделей, отличающихся только инициализацией матрицы Ф. Алгоритм состоит из двух процедур — замены близких тем и добавления новой темы — которые выполняются

Базис тем, полученный в порядке $\{\Phi_i\}_{i=0}^{N}$	Базис тем, полученный в порядке $\{\Phi_i\}_{i=N}^{0}$
история, культура, религия, христианство, востоковедение, ислам, археология	история науки, история, римское право, право, религия, история права, социология права
культура, культурология, массовая культура, философия, кино, общество, искусство	культура, массовая культура, культурология, кино, искусство, кинематограф, фрейд зигмунд
биология, ген, днк, геном, генетика, клетка, белки	молекулярная биология, клеточная биология, биология, генетика, генная инженерия, ген, днк
физика, квантовая физика, технологии, оптика, квантовые технологии, квантовая механика, сверхпроводимость	русская философия, бактерии, народная культура, коренные народы, мертон роберт, русь, доказательная медицина
медицина, биология, генетика, биомедицина, онкология, клетка, стволовые клетки	медицина, стволовые клетки, биология, биомедицина, молекулярная биология, индуцированные плюрипотентные стволовые клетки, регенеративная медицина

Таблица 4. Сравнение полного набора тем моделей LDA и полного набора тем моделей ARTM, построенных на тегах коллекции ПостНаука.

по очереди до сходимости алгоритма. Таким образом, происходит приближенное дополнение линейно независимой системы до базиса. По построению в базис добавляются только те новые темы, в δ-окрестности которых нет выпуклых комбинаций тем базиса.

Замечено, что число тем и скорость сходимости алгоритма зависит от степени разреженности матриц Ф. Базис более разреженных моделей содержит больше тем. Самый маленький базис у моделей LDA. Он содержит наиболее крупные и общие темы, что можно наблюдать на плоской проекции базиса LDA.

Правдоподобие базиса моделей выше, чем одной модели с аналогичным числом тем, на обучении и контроле. Кроме того, было проведено сравнение одной модели и базиса тем по дополнительным критериям качества тематических моделей: критерию Р. Аруна, когерентности и кри-

терию Цао Хуана. На обучении полный набор тем моделей LDA лучше по всем критериям, чем одна модель. На контроле ухудшение только по когерентности.

В экспериментах также было проведено сравнение полных наборов тем моделей LDA и ARTM. В базисах LDA, построенных в прямом и обратном порядке следования моделей, темы совпали с высокой точностью. Однако одна тема в первом базисе была разделена на две во втором: астрономия, астрофизика, космология, гравитация на астрономию, космос и физику, астрофизику. Базис моделей ARTM также разделил эти две темы.

Список литературы

[1] Blei D., Carin L., Dunson D, "Probabilistic Topic Models: A focus on graphical model design and applications to document and image analysis", *IEEE signal processing magazine*, **27**:6 (2010), 55.

[2] Steyvers M., Griffiths T., "Probabilistic topic models", *Handbook of latent semantic analysis*, **427**:7 (2007), 424–440.

[3] De Waal A., Barnard E., "Evaluating topic models with stability", 2008.

[4] Koltcov S., Koltsova O., Nikolenko S., "Latent dirichlet allocation: stability and applications to studies of user-generated content", *Proceedings of the 2014 ACM conference on Web science*, «ACM», 2014, 161–165.

[5] Greene D., O'Callaghan D., Cunningham P., "How many topics? stability analysis for topic models", *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, «Springer», Berlin, Heidelberg, 2014, 498–513.

[6] Balagopalan A., "Improving topic reproducibility in topic models.", 2012.

[7] Lancichinetti A. et al., "High-reproducibility and high-accuracy method for automated topic classification", *Physical Review X.*, **5**:1 (2015), 011007.

[8] Hofmann T., "Probabilistic latent semantic indexing", *ACM SIGIR Forum*, **51**:2 (2017), 211–218.

[9] Blei D.M., Ng A.Y., Jordan M.I., "Latent dirichlet allocation", *Journal of machine Learning research*, **3**:Jan (2003), 993–1022.

[10] Vorontsov K., Potapenko A., "Additive regularization of topic models", *Machine Learning*, **101**:1–3 (2015), 303–323.

[11] Tan Y., Ou Z., "Topic-weak-correlated latent dirichlet allocation", *2010 7th International Symposium on Chinese Spoken Language Processing*, «IEEE», 2010, 224–228.

[12] Kraft D., "A software package for sequential quadratic programming", *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[13] Arun R. et al., “On finding the natural number of topics with latent dirichlet allocation: Some observations”, *Pacific-Asia conference on knowledge discovery and data mining*, «Springer», Berlin, Heidelberg, 2010, 391–402.

[14] Mimno D. et al., “Optimizing semantic coherence in topic models”, *Proceedings of the conference on empirical methods in natural language processing*, «Association for Computational Linguistics», 2011, 262–272.

[15] Cao J. et al., “A density-based method for adaptive LDA model selection”, *Neurocomputing*, **72**:7–9 (2009), 1775–1781.

Building a complete set of topics of probabilistic topic models

Sukhareva A.V., Vorontsov K.V.

Interpretability, linear increase in complexity with data growth, scalability made topic modeling one of the most popular tools for statistical text analysis. However, there are a number of disadvantages caused by the dependence of the solution on the initialization. It is known that the building of a topic model is reduced to solving an ill-posed problem of the non-negative matrix factorization. The set of its solutions in the general case is infinite. Every time the model finds a local extremum. Repeated training of the model for the same collection can lead to detection of more and more new topics. In practice, it is often necessary to define all the topics of the corpus. To solve this problem, the article proposed and investigated a new algorithm for finding the complete set of topics based on the construction of a convex hull. It was shown experimentally that it is possible to construct a basis for the finite number of models. The likelihood of the basis is higher than for a single model with a similar number of topics. Compare of the basis of LDA models (latent Dirichlet allocation) and ARTM models (additive regularization for topic modeling) suggests that the topics of the sets coincide with high accuracy.

Keywords: probabilistic topic modeling, stability of topic models, complete set of topics of topic models, latent Dirichlet allocation, LDA, regularization, ARTM, BigARTM.