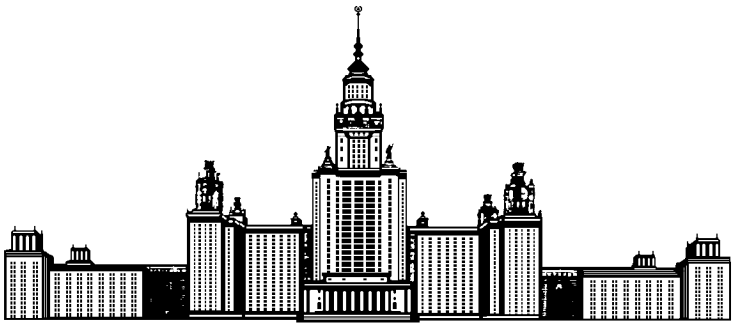
Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования



Магистерская программа
«Логические и комбинаторные методы анализа данных»

Магистерская диссертация

«Оценивание устойчивости и полноты тематических моделей мультидисциплинарных текстовых коллекций»

Работу выполнила
Лукашкина Юлия Николаевна

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

Москва, 2017

Содержание

Аннотация

Данная работа посвящена вероятностному тематическому моделированию — статистическому подходу к определению тем в текстовых коллекциях документов.

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.

В настоящей работе изучается зависимость устойчивости и полноты тематических моделей от их разреженности. Полнота модели определяется как способность модели находить максимально возможное число хорошо интерпретируемых тем. Рассматривается вопрос о возможности сокращения числа запусков обучения модели для поиска всех интерпретируемых тем вплоть до одного путём подбора регуляризаторов.

1 Введение

Данная работа посвящена вероятностному тематическому моделированию — статистическому подходу к определению тем в текстовых коллекциях документов. Под темой подразумевается дискретное распределение на множестве терминов.

Область применения тематического моделирования весьма широка и включает в себя определение трендов в новостных потоках, поиск научной информации, категоризацию документов, рекомендательные сервисы.

Построение тематической модели сводится к решению некорректно поставленной задачи неотрицательного матричного разложения. Некорректность этой задачи заключается в том, что 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.

Требования устойчивости и интерпретируемости моделей имеют большое значение в задачах компьютерной лингвистики, однако в научной литературе проблема устойчивости изучалась недостаточно, за исключением работ [1, 2, 3, 5], а проблема определения, оценивания и повышения полноты моделей даже не ставилась.

В данной работе полнота модели определяется как способность модели находить максимально возможное число хорошо интерпретируемых тем. Рассматривается вопрос о возможности сокращения числа запусков обучения модели для поиска всех интерпретируемых тем вплоть до одного путём подбора регуляризаторов. Помимо исследования описанных вопросов изучается зависимость устойчивости и полноты тематических моделей от их разреженности. Проводятся вычислительные эксперименты на полусинтетических данных. Полусинтетические данные позволяют оценивать качество восстановления «истинных» тем.

Работа имеет следующую структуру.

В разделе 2 вводятся основные понятия и подход аддитивной регуляризации тематических моделей.

Раздел 3 содержит постановку проблем полноты и устойчивости. Проводится обзор существующих подходов к определению устойчивости тематических моделей.

В разделе 4 приводится описание экспериментов и вводятся используемые функционалы качества.

Раздел 5 включает в себя описание проведённых экспериментов и их результаты.

В разделе 6 описаны основные результаты, полученные в данной работе.

2 Постановка задачи тематического моделирования

2.1 Вероятностный латентный семантический анализ (PLSA)

Рассмотрим D — коллекцию текстовых документов и W — множество всех употребляемых в них терминов (слов или словосочетаний). Каждый документ $d \in D$ представляется последовательностью терминов $w \in W$, в нём встречающихся, и их частотами n_{dw}.

Предполагается, что существует конечное число скрытых переменных T — тем. Коллекция документов — это множество троек (d, w, t), выбранных случайно и независимо из дискретного распределения $p(d, w, t)$, заданного на конечном множестве $D \times W \times T$.

Для выявления тем порядок терминов в документах не важен, это предположение называют *гипотезой «мешка слов»*. Порядок документов в коллекции также не влияет на темы.

Гипотеза условной независимости — появление слов, относящихся к теме t, описывается общим для всех коллекции распределением $p(w|t)$ и не зависит от документа d.

$$p(w|d,t) = p(w|t).$$

Вероятностная тематическая модель описывает процесс порождения документов: сначала выбирается тема t из распределения $\theta_{td} = p(t|d)$, затем выбирается слово из распределения $\phi_{wt} = p(w|t)$. Каждый документ d описывается распределением:

$$p(w|d) = \sum_t p(w|t)p(t|d) = \sum_t \phi_{wt}\theta_{td}.$$

Построение тематической модели — это поиск её параметров ϕ_{wt}, θ_{td} и числа тем $|T|$ по заданной текстовой коллекции документов.

В модели PLSA (вероятностного латентного семантического анализа) задача поиска параметров ставится как задача максимизации логарифма правдоподобия коллекции документов D:

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

при ограничениях нормировки и неотрицательности

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

Поставленную задачу можно решать с помощью метода простых итераций (EM-алгоритма). Сначала случайно выбираются начальные приближения ϕ_{wt} и θ_{td}, затем итеративно повторяются E и M шаги.

На E-шаге по текущему приближению параметров рассчитывается апостериорное распределение скрытых переменных и n_{dwt} — число вхождений каждого слова w в каждый документ d, связанных с темой t:

$$p(t|d,w) = \frac{p(t|d)p(w|t)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{st}\theta_{sd}};$$
$$n_{dwt} = n_{dw} p(t|d,w).$$

На M-шаге параметры ϕ_{wt} и θ_{td} вычисляются как частотные оценки соответствующих условных вероятностей.

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_d n_{dwt}, \qquad n_t = \sum_{d,w} n_{dwt};$$
$$\theta_{td} = \frac{n_{dt}}{n_d}, \qquad n_{dt} = \sum_w n_{dwt}, \qquad n_d = \sum_{w,t} n_{dwt}. \tag{2}$$

2.2 Аддитивная регуляризация тематических моделей (ARTM)

В *аддитивной регуляризации тематических моделей* (ARTM) предлагается оптимизировать сумму логарифма правдоподобия $\mathcal{L}$ (1) и регуляризатора $R(\Phi, \Theta)$, зависящего от параметров модели. Регуляризатор $R(\Phi, \Theta)$ может быть суммой нескольких функционалов R_i, взятых с весами, называемыми коэффициентами регуляризации τ_i.

$$\begin{cases} R(\Phi,\Theta) = \sum\limits_i \tau_i R_i(\Phi,\Theta), \quad \log \mathcal{L}(\Phi,\Theta) + R(\Phi,\Theta) \to \max\limits_{\Phi,\Theta}, \\ \phi_{wt} \geqslant 0, \quad \sum\limits_{w \in W} \phi_{wt} = 1, \qquad \theta_{td} \geqslant 0, \quad \sum\limits_{t \in T} \theta_{td} = 1. \end{cases}$$

В [14] показано, что алгоритм многокритериальной оптимизации вероятностной тематической модели, независимо от числа критериев R_i, может быть получен из обычных EM-подобных алгоритмов PLSA или LDA заменой формулы M-шага (2).

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

где

$$\underset{t\in T}{\text{norm}}(z_t) = \frac{\max\{z_t, 0\}}{\sum\limits_{s\in T}\max\{z_s, 0\}} \text{ — оператор нормировки вектора.}$$

Регуляризаторы позволяют формализовать знания предметной области и получать модели с требуемыми свойствами.

3 Проблема устойчивости и полноты

Построение ВТМ можно рассматривать как задачу стохастического матричного разложения: исходную матрицу вероятностей слов в документах $F = ||p(w|d)||$ требуется представить в виде произведения двух матриц меньших размеров — матрицы вероятностей слов в темах $\Phi = ||\phi_{wt}||_{W\times T}$ и матрицы вероятностей тем в документах $\Theta = ||\theta_{td}||_{T\times D}$. Матрицы Φ и Θ — *стохастические*, их столбцы неотрицательны и нормированы:

$$F \approx \Phi\Theta.$$

Эта задача — некорректно поставленная, множество её решений в общем случае бесконечно и имеет вид $(\Phi S)(S^{-1}\Theta)$, где S — произвольная невырожденная матрица, такая, что матрицы $\Phi' = \Phi S$, $\Theta' = S^{-1}\Theta$ также являются стохастическими.

Для решения данной задачи могут быть использованы различные модели: PLSA, LDA, ARTM. Их обучение производится различными методами (EM-алгоритм, вариационный вывод, сэмплирование Гиббса), каждый из которых зависит от некоторого стартового начального приближения, выбираемого, как правило, случайным образом. Таким образом, в модель вносится элемент стохастики, приводящий к тому, что от запуска к запуску результаты моделирования получаются различными.

Прежде всего рассмотрим существующие подходы к оцениванию устойчивости тематических моделей.

3.1 Существующие подходы

Проблема неустойчивости тематических моделей рассматривалась в работах [1, 2, 3, 5].

В одной из первых работ [2], посвященных этому вопросу, вводится понятие «устойчивости» темы. Это темы, которые появляются в модели с фиксированными

параметрами при разных начальных приближениях. В своих экспериментах авторы рассматривают две модели с одинаковым набором параметров и разными случайными инициализациями. После этого они производят сопоставление тем полученных моделей, в качестве метрики близости используется симметрическое KL расстояние:

$$\mathrm{KL}_{sym}(j_1, j_2) = \frac{1}{2}\sum_{k=1}^{|W|} \phi_k^{'(j_1)} \log_2 \frac{\phi_k^{'(j_1)}}{\phi_k^{''(j_2)}} + \frac{1}{2}\sum_{k=1}^{|W|} \phi_k^{''(j_1)} \log_2 \frac{\phi_k^{''(j_1)}}{\phi_k^{'(j_2)}}.$$

Авторы сделали вывод, что на практике решения из различных начальных приближений отличаются, но многие темы появляются во всех запусках, то есть являются устойчивыми.

В работе [5] рассматривается аналогичный подход, но предлагается новая метрика сравнения тем — нормализованная KL-близость:

$$\mathrm{NKLS}(\phi_1, \phi_2) = 1 - \frac{\mathrm{KL}(\phi_1, \phi_2)}{\max\limits_{\phi_1', \phi_2''} \mathrm{KL}(\phi_1', \phi_2'')}.$$

Здесь и далее, *топ-словами* называются некоторое число наиболее вероятных слов в каждой теме. В работе рассмотрены три категории близости тем: совпадение топ-слов и их вероятностей, совпадение только топ-слов и отсутствие любого совпадения.

Авторы статьи [1] рассматривают темы как ранжированные списки. Предлагается метрика, основанная на расстоянии Жаккара, посчитанная по различному числу топ-слов темы.

$$\mathrm{AJ}(\phi_i, \phi_j) = \frac{1}{t}\sum_{d=1}^{t} \gamma_d(\phi_i, \phi_j), \qquad \gamma_d(\phi_i, \phi_j) = \frac{\left|\mathrm{top}(\phi_i, d) \cap \mathrm{top}(\phi_j, d)\right|}{\left|\mathrm{top}(\phi_i, d) \cup \mathrm{top}(\phi_j, d)\right|},$$

где $\mathrm{top}(\phi_i, d)$ — d топ-слов темы ϕ_i. И в этой, и в предшествующих работах для поиска наиболее подходящего соответствия между наборами тем разных моделей используется венгерский алгоритм [9].

В [3] предлагается рассматривать не матрицу слова-темы, как в описанных выше работах, а матрицу тем-документов Θ. Мерой близости в работе является KL дивергенция:

$$\mathrm{KL}(j_1, j_2) = \sum_{k=1}^{|W|} \theta_k^{'(j_1)} \log_2 \frac{\theta_k^{'(j_1)}}{\theta_k^{''(j_2)}}.$$

3.2 Постановка задачи

В рамках исследования проблем полноты и устойчивости тематических моделей в данной работе рассматриваются следующие вопросы.

Проблема устойчивости: в разных запусках получаем разные наборы тем.

- Какими величинами можно количественно оценить устойчивость?
- Насколько хорошо регуляризация способствует повышению устойчивости?

Проблема полноты: в каждом запуске модель находит какие-то новые темы.

- Действительно ли темы новые или это выпуклые комбинации предыдущих?
- Можно ли найти все темы? Сколько моделей для этого нужно построить?
- Как найти полный набор тем за один запуск?

Для решения этих проблем предлагается следующий эмпирический подход.
По заданной коллекции D строится множество моделей

$$(\Phi_j, \Theta_j),\ j = 1, \ldots, J.$$

Из множества X всех тем, образованного вектор-столбцами матриц Φ_j, удаляются дублирующие, линейно зависимые и плохо интерпретируемые темы. Остаётся базисное подмножество $V \subset X$ существенно различных и хорошо интерпретируемых тем. Новые модели строятся до тех пор, пока подмножество V не перестанет пополняться.

Полнота тематической модели (Φ, Θ) оценивается как доля вошедших в неё базисных тем из множества V.

Устойчивость метода обучения тематической модели определяется как мера различия найденных тем в различных запусках.

Количественные оценки устойчивости и полноты предлагается применять для выбора структурных параметров тематической модели: числа тем, набора регуляризаторов и коэффициентов регуляризации.

В данной работе для поиска зависимых тем предлагается использовать алгоритмы построения приближенной выпуклой оболочки множества векторов.

3.3 Определения и обозначения

ε-базисное множество Столбцы матриц Φ_j образуют конечное множество векторов $X = \{x_1, \ldots, x_l\},\ x_i \in \mathbb{R}^{|W|}$, лежащих в единичном симплексе размерности $k = |W| - 1$, который описывает множество всевозможных дискретных вероятностных распределений над словарём терминов W.

$$\Delta = \Big\{\phi = (\phi_w)_{w \in W} : \sum_{w \in W} \phi_w = 1,\ \phi_w \geqslant 0\Big\},$$

Множество V — *ε-базисное* для $X \subset \Delta$, если $V = \{v_1, \ldots, v_m\} \subset \Delta$ такое, что любой вектор $x \in X$ имеет ε-приближенное представление в выпуклой оболочке векторов из V:

$$\min_{v \in \mathrm{conv}V} \rho(x, v) \leqslant \varepsilon, \quad \mathrm{conv}V = \Big\{ v = \sum_{i=1}^{m} \alpha_i v_i \;\Big|\; v_i \in V,\ \sum_{i=1}^{m} \alpha_i = 1,\ \alpha_i \geqslant 0 \Big\}, \tag{3}$$

где $\rho(x, y)$ — функция расстояния.

линейная ε-независимость векторов Введём множество $V' = \{v'_1, \ldots, v'_m\} \subset \Delta$ *линейно ε-независимых векторов*:

$$\forall v'_j \in V' \qquad \min_{u \in \mathrm{conv}(V' \setminus \{v'_j\})} \rho(v'_j, u) > \varepsilon,$$

где

$$\mathrm{conv}(V' \setminus \{v'_j\}) = \Big\{ v = \sum_{i=1}^{m-1} \alpha_i v'_i \;\Big|\; v'_i \in V' \setminus \{v'_j\},\ \sum_{i=1}^{m-1} \alpha_i = 1, \alpha_i \geqslant 0 \Big\}.$$

Требуется определить, какие векторы из другого множества

$$V'' = \{v''_1, \ldots, v''_k\} \subset \Delta$$

являются линейно ε-независимыми с выпуклой оболочкой множества V'.

3.4 Жадный алгоритм построения ε-базисного множества

Алгоритм 3.1. Жадный алгоритм построения ε-базисного множества

Входные данные: множество векторов-тем X, ε;

Выходные данные: базисное подмножество $V \subset X$;

```
1 инициализировать V = X;
2 повторять
3     цикл i = 1, ..., |V| выполнять
4         если ∃j : min_{v ∈ convV \ v_j} ρ(v_j, v) ⩾ ε тогда
5             V = V \ v_j ;
6 до тех пор, пока V базисное множество;
```

Опишем алгоритм построения множества V по множеству X (3.1). Предлагается инициализировать $V = X$, а затем итеративно удалять линейно зависимые векторы, до тех пор, пока все векторы не станут линейно независимыми. Данный алгоритм работает довольно долго, поскольку для каждой темы из набора требуется многократно определить её зависимость от всех остальных.

Для ускорения процесса построения ε-базисного множества предлагаются следующие эвристики:

- для каждого вектора из X хранить проекцию на V, что позволит ускорить определение линейной зависимости векторов;
- на шаге 5 итеративно удалять не по одному вектору, линейно выражающемуся через остальные, а сразу группу подобных векторов, находящихся на достаточном расстоянии друг от друга.

3.5 Алгоритм определения линейной ε-независимости

Множество векторов V' образует столбцы матрицы $\Phi' = ||\phi'_{wt}||_{n\times m}$, векторы V'' — столбцы матрицы $\Phi'' = ||\phi''_{wt}||_{n\times k}$. Требуется найти матрицу коэффициентов линейных комбинаций $A = ||\alpha_{ts}||_{m\times k}$, такую, что

$$\Phi'' \approx \Phi' A. \tag{4}$$

Существуют различные варианты решения поставленной задачи. Можно решать k задач многомерной минимизации расстояния $\rho(u, v)$. Для каждого фиксированного вектора v''_s требуется найти вектор коэффициентов $\hat{\alpha_s} = (\alpha_{ts})_{m\times 1}$, являющийся решением оптимизационной задачи:

$$\rho(v''_s, \Phi'\hat{\alpha_s}) \to \min_{\hat{\alpha_s}},$$

со следующими ограничениями:

$$\alpha_{ts} \geqslant 0, \quad \sum_{t=1}^{m} \alpha_{ts} = 1.$$

Другая идея решения исходной задачи (4) — переформулировать её в терминах тематической модели.

Заметим, что на матрицу A накладываются такие же условия, как и на матрицу в BTM, т.е. она должна быть стохастической. В данной постановке матрица

документы-слова Φ'' — это матрица новых векторов V'', матрица слова-темы Φ'— матрица старых линейно ε-независимых векторов V'', матрица темы-документы — это матрица коэффициентов линейных комбинаций A.

Если в качестве функции ρ выбрать дивергенцию Кульбака–Лейблера, то получим следующую оптимизационную задачу:

$$\mathrm{KL}(\Phi''||\Phi' A) \to \min_A,$$

для решения которой можно воспользоваться EM-алгоритмом с фиксированной матрицей Φ.

4 Описание экспериментов

Предположим, что имеется тематическая модель (Φ_0, Θ_0) с различными и интерпретируемыми темами, построенная по некоторой текстовой коллекции.

На синтетической коллекции $F = \Phi_0\Theta_0$ производится многократное построение новых разложений $\Phi_i \times \Theta_i$ из разных случайных начальных приближений, возможно с числом тем, отличным от оригинального (которое известно), и рассматриваются следующие проблемы:

1. Устойчивость: степень различности Φ_i в разных запусках.
2. Полнота:
   - возможность получения истинных тем в процессе серии запусков;
   - возможность поиска всех истинных тем в ходе обучения единственной модели с подходящим набором регуляризаторов.
3. Влияние разреженности матриц (Φ_0, Θ_0) на результаты экспериментов.
4. Влияние регуляризаторов на результаты экспериментов.

Общее число тем в восстановленных моделях $\Phi_i \times \Theta_i$ может достигать нескольких тысяч (например, если рассмотреть десяток моделей по 100 тем). Однако среди них будет много дублирующих и линейно зависимых тем, которые не представляют особого интереса, затрудняют восприятие и оценивание построенных моделей. Достаточно рассмотреть только подмножество существенно различных тем, которые описывают всю коллекцию — ε-базисное множество (3). Столбцы полученных матриц

Φ_i образуют множество векторов X. Для построения ε-базисного множества $V \subset X$ воспользуемся алгоритмом, описанным в разделе 3.4. Таким образом вопрос о восстановлении исходных тем можно рассматривать только для построенного ε-базисного множества, являющимся полным базисным набором.

4.1 Сопоставление исходных и восстановленных тем

Рассмотрим несколько функционалов, которые будем использовать для оценки близости восстановленных тем и исходных.

$\Phi_0 = ||\phi_{wt}||_{W \times T}$ — исходные темы, $\hat{\Phi} = ||\hat{\phi}_{wt}||_{W \times \hat{T}}$ — восстановленные темы. Очевидно, что применение какой-либо перестановки к столбцам матрицы Φ_0 меняет порядок тем, не изменяя сами темы. Для каждой темы из построенного набора $\hat{\Phi}$ нужно найти ближайшую тему из исходных — эта задача решается венгерским алгоритмом, где матрица стоимости — матрица расстояний

$$D(\Phi_0, \hat{\Phi}) = ||d_{ij}||_{T \times \hat{T}} = ||\rho(\phi_t, \hat{\phi}_s)||_{T \times \hat{T}}.$$

Матрицы Φ_0 и $\hat{\Phi}$ рассматриваются по столбцам: $\Phi = (\phi_1, \ldots, \phi_T)$. Рассмотрим различные соответствия между столбцами восстановленной матрицы и исходной:

$$\pi : \hat{\phi}_s \to \phi_t.$$

Будем называть соответствие π *допустимым*, если

$$\forall s,\ \rho(\hat{\phi}_s, \pi(\hat{\phi}_s)) < \varepsilon,$$

и выполняется следующее условие:

$$\forall \hat{\phi}_t \neq \hat{\phi}_s \to \pi(\hat{\phi}_t) \neq \pi(\hat{\phi}_s).$$

Зафиксируем значение расстояния ε, при котором можно считать, что темы близкие (подробнее раздел 5.3), будем рассматривать только допустимые соответствия π:

$$D_1(\Phi_0, \hat{\Phi}, \rho, \varepsilon) = \left| \left\{ \hat{\phi}_s \middle|\ \hat{\phi}_s \in \hat{\Phi}, \exists \phi_{s'} \in \Phi_0 : \phi_{s'} = \pi(\hat{\phi}_s) \right\} \right|, \tag{5}$$

— число построенных тем ($\hat{\Phi}$), которые восстанавливают исходные темы (Φ_0).

В общем случае нельзя гарантировать точного совпадения полученных и исходных тем (при заданном пороге близости тем ε). Восстановленные темы могут являться линейной комбинацией исходных. Например, если число восстанавливаемых

тем значительно меньше исходных, скорее всего восстанавливаемые темы будут более общими по смыслу и объединять несколько исходных. В таком случае интересно рассмотреть, как близко восстановленные темы проектируются на исходные и сколько в среднем исходных тем нужно скомбинировать, чтобы получить одну восстановленную. Для этого вводится следующий функционал:

$$D_2(\Phi_0, \hat{\Phi}, \rho, \varepsilon) = \left| \left\{ \hat{\phi}_s \middle|\ \hat{\phi}_s \in \hat{\Phi}, \exists \phi \in \text{conv}(\Phi_0) : \rho(\hat{\phi}_j, \phi) < \varepsilon \right\} \right|, \tag{6}$$

— число построенных тем ($\hat{\Phi}$), которые ε-приближают линейную комбинацию исходных тем (Φ_0).

Любой вектор $\phi \in \text{conv}(\Phi_0)$ можно представить следующим образом:

$$\phi = \sum_{t=1}^{|T|} \alpha_t \phi_t^0, \quad \phi_t^0 \in \Phi_0, \ \sum_{t=1}^{|T|} \alpha_t = 1, \ \alpha_t \geqslant 0.$$

Обозначим $C_{\text{conv}(\Phi_0)}(\phi) = \{\alpha_t |\ \alpha_t > 0\}$ — множество коэффициентов, которые участвуют в разложении этого вектора ϕ через выпуклую оболочку Φ_0. Пусть ϕ_i — ε-проекция какого-либо $\hat{\phi} \in \hat{\Phi}$, тогда:

$$C_2(\Phi_0, \hat{\Phi}, \rho, \varepsilon) = \frac{1}{D_2(\Phi_0, \hat{\Phi}, \rho, \varepsilon)} \sum_{i=1}^{D_2(\Phi_0, \hat{\Phi}, \rho, \varepsilon)} \left| C_{\text{conv}(\Phi_0)}(\phi_i) \right|, \tag{7}$$

— среднее число исходных тем (Φ_0), которые нужно скомбинировать, чтобы получить одну восстановленную ($\hat{\Phi}$).

Чем меньше значение функционала (7), тем ближе проекция восстановленной темы к какой-либо исходной. В вырожденном случае, когда исходные темы ε-точно восстанавливаются исходными $C_2(\Phi_0, \hat{\Phi}, \rho, \varepsilon) = 1$.

Значения функционалов (5) и (6) используются, чтобы измерять полноту модели. Эти функционалы количественно определяют, сколько восстановленных тем совпадают с исходными, и как хорошо они проектируются на исходные соотвественно. Анализ значений функционала (7) позволяет посмотреть на структуру этих проекций.

4.2 Устойчивость

Устойчивость метода обучения рассчитывается на основе различности матриц $\hat{\Phi}_i$, полученных в результате обучения моделей с разными начальными приближениями.

Предлагается попарно сравнивать полученные матрицы с фиксированной первой $\hat{\Phi}_1$ и количественно определять устойчивость с помощью функционала (5):

$$S(\{\hat{\Phi}_i\}_{i=1}^m, \rho, \varepsilon, m) = \frac{1}{(m-1)|T|} \sum_{i=2}^{m} D_1(\hat{\Phi}_1, \hat{\Phi}_i, \rho, \varepsilon), \tag{8}$$

где $|T|$ — число тем в $\hat{\Phi}_i$ (одинаковое для серии рассматриваемых матриц), а m — число запусков обучения моделей с разичными начальными приближениями.

5 Эксперименты

5.1 Данные

Эксперименты проводились на текстах мультидисциплинарной коллекции ПостНаука [1]. Коллекция состоит из статей посвященных различным областям науки (физика, программирование, социология, биология и другие). Средняя длина документа составляет 180 уникальных терминов и 330 терминов всего. Суммарное колличество документов ≈ 2500 документов, объём исходного словаря ≈ 20000 слов.

Коллекция прошла следующую предобработку: приведение текста к нижнему регистру, стемминг, удаление редко и часто встречающихся слов (стоп-слов), выделение коллокаций. Последний шаг направлен на выявление специфической терминологии. Это операция особенно важна для научных междисциплинарных текстов по причине наличия в них терминов, состоящих из двух и более слов, которые в униграммном представлении теряют свою семантику.

5.2 Создание синтетических коллекций

Для построения синтетических данных с известным полным набором тем производятся следующие действия. Строится тематическая модель исходной коллекции, после чего перемножаются получившиеся матрицы параметров Φ_0 и Θ_0. Домножение полученного произведения на длины документов приводит к получению реалистичных полусинтетических данных. Такая операция производится для моделей с различными степенями разреженности матриц параметров, что было достигнуто использованием разнообразных стратегий регуляризации моделей в процессе их обучения. Полученная совокупность данных используется в последующих экспериментах.

[1] `https://postnauka.ru/`

№	Z_{Φ_0}	Z_{Θ_0}	P	$\|W\|$	R_1	R_2	R_3
1	0.87	0.00	1046	17489	–	–	–
2	0.87	0.65	1036	17489	–	+	–
3	0.87	0.85	1026	17489	–	+	–
4	0.87	0.91	1024	17489	–	+	–
5	0.87	0.95	1021	17488	–	+	–
6	0.87	0.97	1009	17488	–	+	–
7	0.87	0.98	912	17485	–	+	–
8	0.99	0.80	1472	4705	+	+	–
9	0.99	0.87	1454	4628	+	+	–
10	0.99	0.93	1378	4495	+	+	–
11	0.99	0.96	1477	4190	+	+	–
12	0.92	0.89	1079	17489	+	+	+
13	0.96	0.82	1221	17448	+	+	+
14	0.97	0.78	1227	16234	+	+	+

Таблица 1: Характеристики синтетических данных

Этот способ порождения коллекции (предложен в [15]) отличается от большинства исследований, где матрицы (Φ_0, Θ_0) генерируются из распределений Дирихле.

В таблице 1 показаны некоторые характеристики получившихся синтетических наборов данных: степень разреженности[2] матриц Φ и Θ — столбцы 3 и 4 соответственно, значение перплексии P, размер получившегося словаря слов $|W|$.

Все полусинтетические наборы были построены с фиксированным числом тем $|T| = 100$. Видно, что PLSA и без регуляризаторов достаточно хорошо разреживает матрицу Φ. Также в таблице указаны стратегии регуляризации, используемые для построения моделей, где R_1, R_2 — регуляризаторы разреживания матриц Φ и Θ соответственно, а R_3 — регуляризатор декорреляции тем. Подробнее эти регуляризаторы описаны в работе [12]. Построенные наборы являются представительным множеством полусинтетических данных.

Стоит отметить, что регуляризация моделей, и, в частности, принудительное разреживание матрицы Φ может целиком занулить некоторые строки этой матрицы, что приводит к сокращению размера словаря.

[2] $\mathrm{Z_M}$ — доля нулевых элементов в матрице M

5.3 Оценивание близости тем

Процедура автоматического определения схожести тем предполагает наличие, помимо функции расстояния, порога, относительно которого темы разделяются на схожие и различные. В качестве функции расстояния $\rho(u, v), u \in \mathbb{R}^n, v \in \mathbb{R}^n$ используются следующие величины:

- Расстояние Хеллингера (Hellinger)

$$\mathrm{H}(u, v) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n} (\sqrt{u_i} - \sqrt{v_i})^2}$$

- Дивергенция Кульбака-Лейблера (Kullback–Leibler)

$$\mathrm{KL}(u||v) = \sum_{i=1}^{n} u_i \ln \frac{u_i}{v_i}$$

- Косинусное расстояние

$$\cos(u, v) = 1 - \frac{\sum\limits_{i=1}^{n} u_i v_i}{\sqrt{\sum\limits_{i=1}^{n} u_i^2} \sqrt{\sum\limits_{i=1}^{n} v_i^2}}$$

- Расстояние Жаккара (Jaccard)[3]

$$\mathrm{J}(u, v, k) = \frac{|\mathrm{top}(u, k) \cap \mathrm{top}(v, k)|}{|\mathrm{top}(u, k) \cup \mathrm{top}(v, k)|}$$

Для каждой из описанных метрик подбор порога производится по следующей схеме. Множество значений функции (отрезок) разбивается на сегменты длины 0.1. Затем для каждого полученного сегмента случайным образом выбирается множество пар тем, расстояние между которыми находится в пределах данного сегмента. Для каждой темы предъявляются наиболее вероятные слова и документы, в которых данная тема имеет наибольшую вероятность. На основании этих данных была проведена процедура ручной оценки близости каждой из рассматриваемых пар тем, был выявлен эмпирический оптимальный порог для всех последующих экспериментов.

В таблице 2 приведён пример нескольких пар тем с различными расстояниями между ними. Результаты асессорской работы показали, что метрикой, наилучшим

[3]top(u, k) — слова, соответствующие k наибольшим вероятностям распределения u

образом характеризующей близость тем является расстояние Хеллингера. Во всех остальных проведённых экспериментах использовалось именно оно. Стоит отметить, что у тем, расстояние между которыми < 0.1 топ-слова совпадали в точности до порядка следования. А темы, расстояние между которыми < 0.25, имели различия в 1-2 топ-словах. Оптимальное значение порога $= 0.25$, то есть темы, расстояние между которыми меньше выбранного порога, считались близкими.

5.4 Эксперименты с серией моделей

Построение ϵ-базисного множества В даной серии экспериментов производится построение полного набора независимых тем.

Полученные ранее синтетические данные используются для построения серии моделей с фиксированными параметрами и различными начальными приближениями. В данном эксперименте под *итерацией* понимается обучение очередной модели. Для выделения полного набора тем воспользуемся алгоритмом построения ε-базисного множества V 3.1.

На графиках 1 и 2 показана зависимость числа тем в множестве $|V|$ от итерации для моделей с числом тем равным 20 и 100 соотвественно. На каждой итерации к V добавляются столбцы новой модели и производится процедура фильтрации ε линейно-зависимых тем. Негладкость графика обусловлена тем, что мощность множества V может как увеличиться, так и уменьшиться (например, если было отброшено несколько линейно-зависимых тем и добавлена одна новая опорная тема).

Число тем в полном наборе может превышать число исходных. Это может быть связано с выбором заниженного порога различности тем ε, либо с тем, что, вообще говоря, исходное разложение может быть не единственным.

Из графиков видно, что стабилизация мощности базисного множества модели с 20 темами происходит при достижении 20-ой итерации, в то время как для модели с 100 темами — при 10-ой.

Процедура построения полного базисного множества проводится для моделей с различной степенью разреженности, которые строятся на каждом из синтетических наборов. Таблица 3 содержит показатели разреженности построенных матриц Φ и Θ, число итераций для построения множества V и значения функционала (7), который характеризует среднее число элементов базисного множества, которые участвуют в линейных комбинациях приближающих весь построенный набор тем. Для наглядности в этой таблице приведены не все возможные комбинации построенных

Расстояние ≈ 0.12				Расстояние ≈ 0.23			
япония	0.335	япония	0.345	технология	0.093	технология	0.083
японец	0.136	японец	0.139	производство	0.045	производство	0.041
страна	0.128	страна	0.135	изобретение	0.033	изобретение	0.031
корея	0.076	корея	0.078	машина	0.029	инновация	0.027
безработица	0.050	безработица	0.052	инновация	0.029	машина	0.026
работа	0.041	работа	0.042	создание	0.028	создание	0.025
аляска	0.037	аляска	0.038	ПР[4]	0.027	ПР	0.025
общество	0.031	общество	0.032	использование	0.026	использование	0.023
ЮК[5]	0.030	ЮК	0.031	материал	0.024	компания	0.021
процент	0.024	процент	0.025	компания	0.024	разработка	0.020
калифорния	0.018	калифорния	0.018	устройство	0.023	устройство	0.020
европеец	0.013	европеец	0.013	разработка	0.023	материал	0.020
юг	0.012	юг	0.012	энергия	0.020	страна	0.020

Расстояние ≈ 0.29				Расстояние ≈ 0.46			
чёрный_дыра	0.038	чёрный_дыра	0.052	слово	0.066	слово	0.027
вселенная	0.031	гравитация	0.019	словарь	0.027	предложение	0.034
объект	0.015	объект	0.018	русский_язык	0.021	глагол	0.025
гравитация	0.015	вселенная	0.018	язык	0.019	текст	0.023
пространство	0.014	пространство	0.015	предложение	0.017	язык	0.022
теория	0.013	теория	0.012	глагол	0.016	русский_язык	0.021
материя	0.010	масса	0.012	гласный	0.010	значение	0.016
масса	0.009	ОТО[6]	0.008	жест	0.010	синтаксис	0.014
волна	0.008	струна	0.008	синтаксис	0.009	существительное	0.011
физика	0.007	горизонт	0.008	значение	0.008	форма	0.011
ОТО	0.006	поверхность	0.008	диалект	0.008	перевод	0.011
горизонт	0.006	теория_струна	0.008	лингвист	0.008	падеж	0.011
струна	0.006	материя	0.007	существительное	0.008	грамматика	0.010

Таблица 2: Топ-слова нескольких пар тем и расстояние между ними.

ОТО — общий_теория_относительность, ЮК — южная корея, ПР — промышленный_революция

Z_Φ	Z_Θ	$N_{\|V\|}$	C_2
0.81	0.00	22	1.63
0.96	0.00	12	1.42
0.96	0.38	9	1.34
0.96	0.61	7	1.35
0.96	0.93	5	1.12
0.96	0.95	6	1.09

Таблица 3: Зависимость разреженности восстановленных моделей от $N_{|V|}$, где $N_{|V|}$ — номер итерации, при которой стабилизируется базисное множество V

моделей и синтетических наборов данных, а лишь те, показатели разреженности для которых существенно различаются.

Увеличение степени разреженности восстановленной модели приводит к уменьшению числа итераций, требуемых для стабилизации множества V. Также увеличение разреженности приводит к уменьшению числа новых тем, выявляемых на каждой итерации. Низкое значение функционала C_2 (среднее значение ≈ 1.2) показывает, что каждая тема, которая не вошла в базисное множество, выражается через небольшое число элементов данного множества.

Рис. 1: Зависимость $|V|$ от номера итерации, модель с 20 темами, синтетический набор данных №11, $\varepsilon = 0.25$

Восстановление исходных тем Рассмотрим, как число тем в построенных моделях и их разреженность влияют на качество восстановления исходных данных.

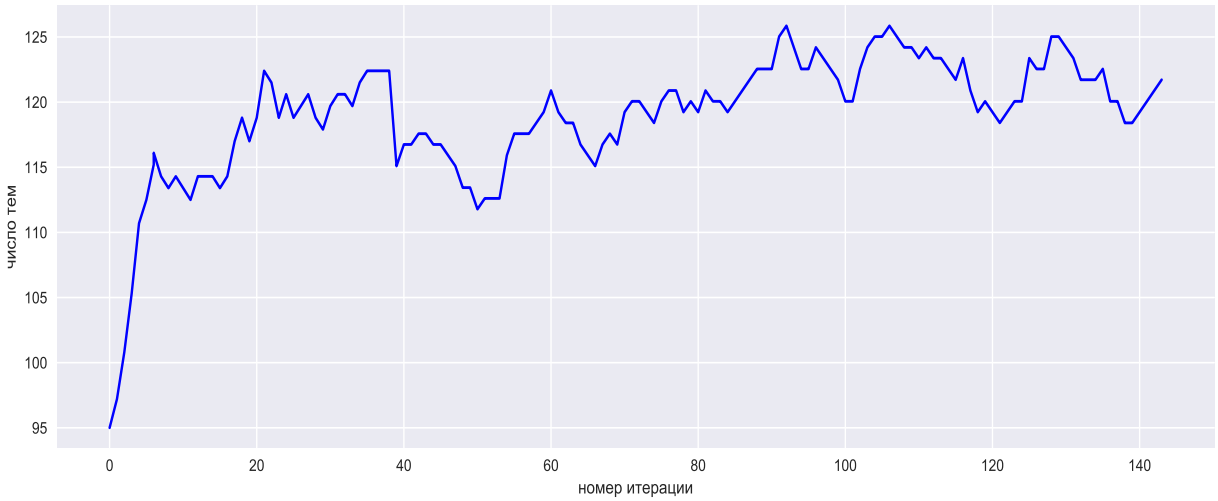


Рис. 2: Зависимость $|V|$ от номера итерации, модель с 100 темами, синтетический набор данных №11, $\varepsilon = 0.25$

Для изучения качества приближения исходных тем полным набором построенных рассмотрим функционалы, которые были введены в 4.1.

В таблице 4 приведены результаты эксперимента для различных моделей, наилучшим образом приближающих исходные данные среди аналогичных моделей с фиксированным числом тем и стратегией регуляризации.

Полный набор, полученный на моделях с числом тем существенно меньшим, чем их исходное количество (20 тем против 100) не может обеспечить точное приближение истинного набора тем. Ни одна из использованных стратегий регуляризации (разреживание матрицы Φ, разреживание матрицы Θ, декорреляция матрицы Φ) или их комбинация не позволила получить иного результата. Тем не менее, подобные полные наборы хорошо приближают линейную комбинацию нескольких исходных тем. Этот результат кажется закономерным, поскольку темы в моделях с небольшим числом тем получаются довольно общими. Это приводит к тому, что каждая их них описывает несколько истинных тем.

Для случая моделей с числом тем $= 100$ и более ситуация выглядит иначе. Высокое значение функционала D_1 позволяет сделать вывод о том, что такие модели хорошо приближают истинные темы. Но линейные комбинации исходных тем приближаются ими ещё лучше (функционал D_2).

Если сравнить значения функционала C_2 для моделей с 20 и 100 темами можно сделать вывод, что описание исходных тем небольшими моделями требует линейных комбинаций большего числа тем, чем при описании более крупными моделями.

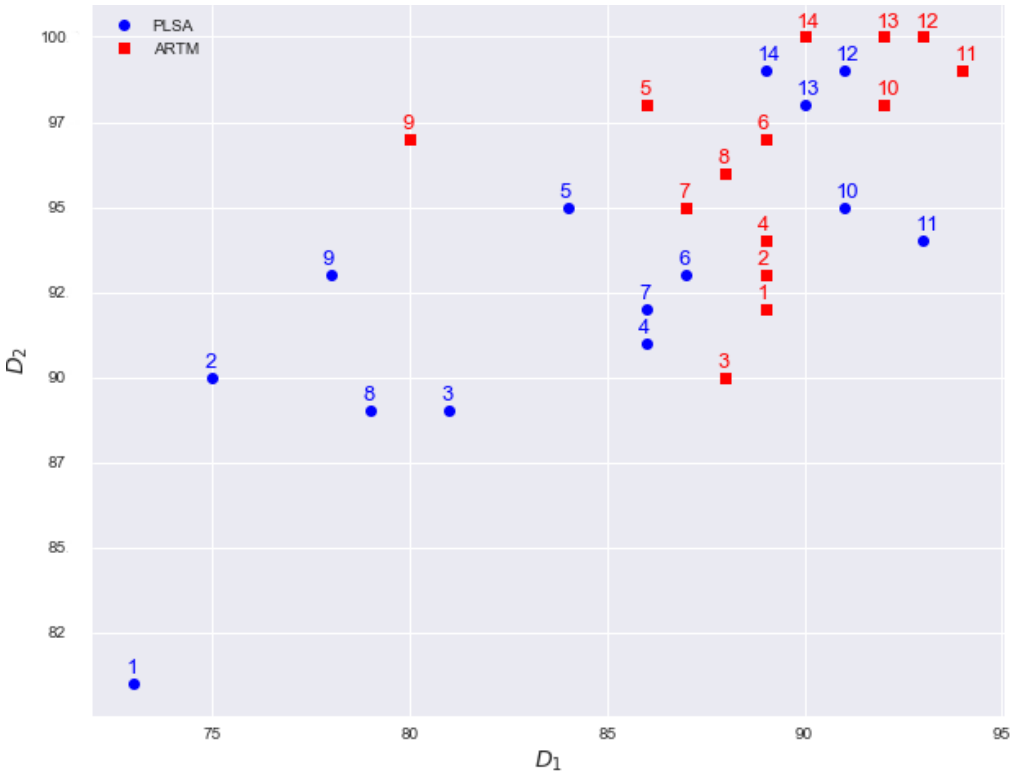


Рис. 3: Показатели качества приближения исходных данных полным набором тем для наилучших моделей; $T = 100$; $\varepsilon = 0.1$.

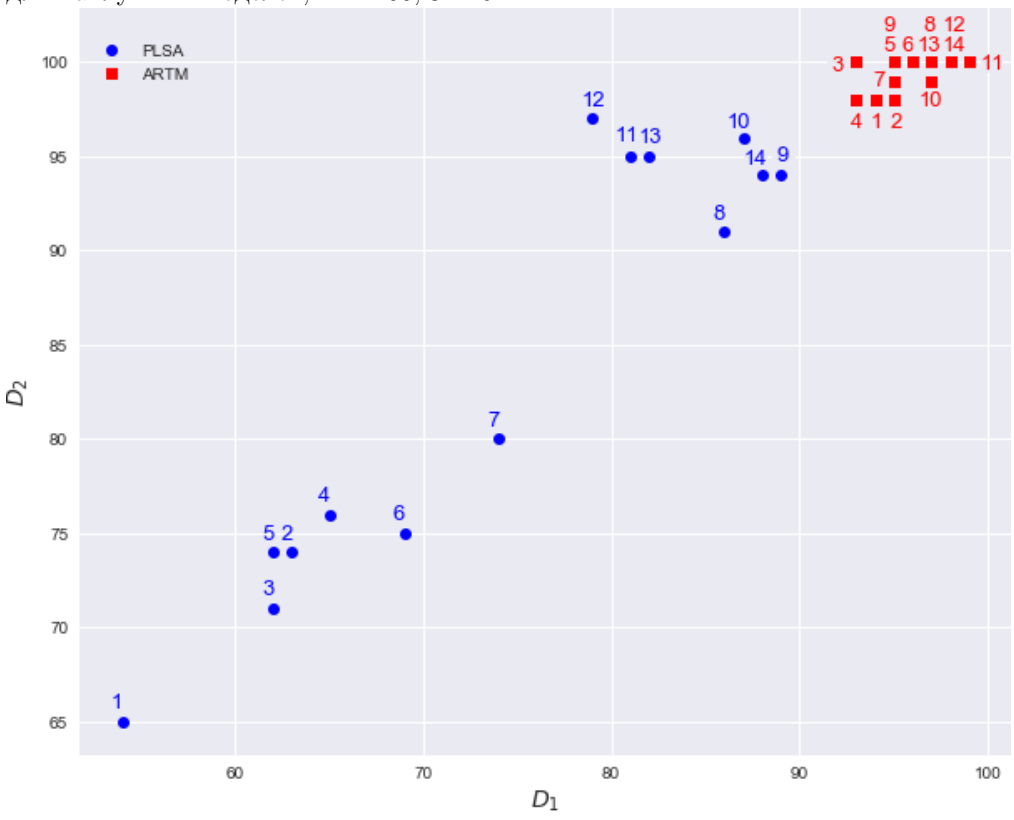


Рис. 4: Показатели качества приближения исходных данных полным набором тем для наилучших моделей; $T = 150$; $\varepsilon = 0.1$

T	R_1	R_2	R_3	D_1	D_2	C_2
20	–	–	–	0	34	7.54
20	–	+	–	0	77	5.67
20	+	+	–	0	89	5.91
20	+	+	+	0	91	5.43
100	–	–	–	84	95	2.45
100	–	+	–	83	95	1.30
100	+	+	–	86	99	1.29
100	+	+	+	87	98	1.34
150	–	–	–	81	95	2.12
150	–	+	–	98	100	1.09
150	+	+	–	95	100	1.16
150	+	+	+	99	100	1.15

Таблица 4: Показатели качества приближения исходных данных полным набором тем для наилучших моделей

Особенный интерес представляет влияние разреженности исходных синтетических данных на качество восстановления тем. На основании результатов эксперимента (см. таблицу 5) можно заключить, что более сильная разреженность данных приводит к более высокому качеству восстановления тем даже при использовании простой модели PLSA. В то же время при уменьшении степени разреженности приводит к усилению роли регуляризаторов. При этом в обоих случаях регуляризация позволяет достигнуть более высоких результатов.

На графиках 3, 4 отдельно отображены показатели качества восстановления исходных тем D_1, D_2 для моделей из таблицы 5. Каждая точка графика соответствует одной построенной модели, номер точки — номер исходного полусинтетического набора данных. Для всех наборов данных показатели D_1, D_2 для ARTM моделей выше, чем для PLSA (с одинаковым числом тем). PLSA модели с числом тем $= 150$ восстанавливают исходные темы хуже, чем PLSA модели с числом тем $=\ 100$. Однако ARTM модели с числом тем $= 150$ восстанавливают исходные данные качественнее, чем ARTM модели с 100 темами и PLSA модели. Значения функционала D_2 выше соответствующих значений D_1 для любой модели.

№	Z_{Φ_0}	Z_{Θ_0}	M	Z_Φ	Z_Θ	D_1	D_2	№	Z_{Φ_0}	Z_{Θ_0}	M	Z_Φ	Z_Θ	D_1	D_2
1	0.87	0.00	PLSA	0.72	0.00	73	81	1	0.87	0.00	PLSA	0.65	0.00	54	65
1	0.87	0.00	ARTM	0.81	0.00	89	92	1	0.87	0.00	ARTM	0.87	0.04	94	98
2	0.87	0.65	PLSA	0.80	0.00	75	90	2	0.87	0.65	PLSA	0.80	0.00	63	74
2	0.87	0.65	ARTM	0.87	0.65	89	93	2	0.87	0.65	ARTM	0.88	0.63	95	98
3	0.87	0.85	PLSA	0.86	0.00	81	89	3	0.87	0.85	PLSA	0.88	0.00	62	71
3	0.87	0.85	ARTM	0.87	0.85	88	90	3	0.87	0.85	ARTM	0.88	0.86	93	100
4	0.87	0.91	PLSA	0.88	0.00	86	91	4	0.87	0.91	PLSA	0.90	0.00	65	76
4	0.87	0.91	ARTM	0.88	0.89	89	94	4	0.87	0.91	ARTM	0.88	0.91	93	98
5	0.87	0.95	PLSA	0.88	0.00	84	95	5	0.87	0.95	PLSA	0.90	0.00	62	74
5	0.87	0.95	ARTM	0.88	0.94	86	98	5	0.87	0.95	ARTM	0.88	0.94	95	100
6	0.87	0.97	PLSA	0.88	0.00	87	93	6	0.87	0.97	PLSA	0.91	0.00	69	75
6	0.87	0.97	ARTM	0.88	0.97	89	97	6	0.87	0.97	ARTM	0.88	0.98	96	100
7	0.87	0.98	PLSA	0.88	0.00	86	92	7	0.87	0.98	PLSA	0.91	0.00	74	80
7	0.87	0.98	ARTM	0.88	0.98	87	95	7	0.87	0.98	ARTM	0.88	0.98	95	99
8	0.99	0.80	PLSA	0.89	0.00	79	89	8	0.99	0.80	PLSA	0.90	0.00	86	91
8	0.99	0.80	ARTM	0.97	0.80	88	96	8	0.99	0.80	ARTM	0.98	0.82	97	100
9	0.99	0.87	PLSA	0.91	0.00	78	93	9	0.99	0.87	PLSA	0.92	0.00	89	94
9	0.99	0.87	ARTM	0.97	0.81	80	97	9	0.99	0.87	ARTM	0.98	0.85	95	100
10	0.99	0.93	PLSA	0.95	0.00	91	95	10	0.99	0.93	PLSA	0.95	0.00	87	96
10	0.99	0.93	ARTM	0.97	0.85	92	98	10	0.99	0.93	ARTM	0.98	0.89	97	99
11	0.99	0.96	PLSA	0.96	0.00	93	94	11	0.99	0.96	PLSA	0.97	0.00	81	95
11	0.99	0.96	ARTM	0.98	0.90	94	99	11	0.99	0.96	ARTM	0.98	0.92	99	100
12	0.92	0.89	PLSA	0.89	0.00	91	99	12	0.92	0.89	PLSA	0.90	0.00	79	97
12	0.92	0.89	ARTM	0.91	0.75	93	100	12	0.92	0.89	ARTM	0.92	0.88	98	100
13	0.96	0.82	PLSA	0.88	0.00	90	98	13	0.96	0.82	PLSA	0.90	0.00	82	95
13	0.96	0.82	ARTM	0.96	0.81	92	100	13	0.96	0.82	ARTM	0.96	0.80	97	100
14	0.97	0.78	PLSA	0.89	0.00	89	99	14	0.97	0.78	PLSA	0.90	0.00	88	94
14	0.97	0.78	ARTM	0.95	0.75	90	100	14	0.97	0.78	ARTM	0.97	0.79	98	100

Таблица 5: Показатели качества приближения исходных данных полным набором тем для наилучших моделей; слева $T = 100$, справа $T = 150$; $\varepsilon = 0.1$

На основании сделанных выводов можно заключить, что оптимальная стратегия заключается в использовании модели с большим числом тем и аккуратным подбором траектории регуляризации.

Сравнение качества приближения истинного набора тем всеми восстановленными темами и только теми из них, которые вошли в полный набор показывает, что тем полного набора достаточно для описания истинных тем с высокой точностью.

$V,\ D_1$	0	0	0	83	86	87	98	95	99
$V,\ D_2$	77	89	91	95	99	98	100	100	100
$X,\ D_1$	0	0	0	83	86	87	98	95	99
$X,\ D_2$	78	89	93	95	100	98	100	100	100

Таблица 6: Сравнение показателей качества приближения исходных данных полным набором тем и всеми возможными темами.

Данный вывод подтверждается результатами, описанными в таблице 6. Один столбец — одна построенная модель, по строкам — показатели D_1, D_2 для полного набора тем (V) и для всех построенных тем (X). Соответствующие показатели по строкам практически не отличаются.

Устойчивость Рассматривается зависимость устойчивости метода обучения от разреженности получаемых моделей. Таблица 7 содержит значения функционала (8) для PLSA моделей с различным числом фиксированных тем и несколькими значениями порога близости ε. Также в этой таблице приводятся значения средних разреженностей матриц Φ полученных моделей (последние столбцы). Таблица 8 содержит аналогичные показатели для моделей, к которым были применены различные стратегии регуляризации, позволившие увеличить устойчивость этих моделей.

На основании полученных результатов можно сделать вывод о том, что более сильная разреженность исходных синтетических данных приводит к большей устойчивости даже простой модели PLSA (при фиксированном пороге схожести тем). При этом качественный подбор регуляризаторов позволяет получить более устойчивые модели. Их устойчивость в среднем одинаковая на всех наборах данных.

5.5 Эксперименты с одной моделью

В предыдущих разделах был рассмотрен вопрос о приближении исходных тем полным набором восстановленных, который строится по серии моделей из разных запусков. Интересно рассмотреть возможность поиска всех истинных тем в ходе обучения единственной модели с подходящим набором регуляризаторов.

По результатам предыдущих экспериментов рассматривались только те модели, число тем в которых превышало число исходных. Тщательный подбор коэффициентов регуляризации, которые позволяют приблизить структуру разреженности исходных данных, позволил достичь результатов, сопоставимых с аналогичными резуль-

№	100 тем $\varepsilon = 0.1$	150 тем $\varepsilon = 0.1$	100 тем $\varepsilon = 0.25$	150 тем $\varepsilon = 0.25$	100 тем Z_Φ	150 тем Z_Φ
1	40	38	59	65	0.72	0.65
2	42	38	56	57	0.80	0.80
3	43	37	54	56	0.86	0.88
4	41	39	58	60	0.88	0.90
5	55	55	62	59	0.88	0.90
6	58	44	66	63	0.88	0.91
7	57	43	66	60	0.88	0.91
8	68	59	73	66	0.89	0.90
9	72	58	74	64	0.91	0.92
10	69	88	72	93	0.95	0.95
11	71	58	73	65	0.96	0.97
12	67	63	71	69	0.89	0.90
13	61	58	74	71	0.88	0.90
14	60	55	72	70	0.89	0.90

Таблица 7: Показатели устойчивости S (8) для PLSA моделей

№	100 тем $\varepsilon = 0.1$	150 тем $\varepsilon = 0.1$	100 тем $\varepsilon = 0.25$	150 тем $\varepsilon = 0.25$	100 тем Z_Φ	100 тем Z_Θ	150 тем Z_Φ	150 тем Z_Θ
1	59	68	70	79	0.81	0.00	0.87	0.04
2	63	65	68	71	0.87	0.65	0.88	0.63
3	67	69	73	76	0.87	0.85	0.88	0.86
4	66	70	69	72	0.88	0.89	0.88	0.91
5	69	73	71	76	0.88	0.94	0.88	0.94
6	65	74	74	80	0.88	0.97	0.88	0.98
7	68	71	73	79	0.88	0.98	0.88	0.98
8	76	74	82	79	0.97	0.80	0.98	0.82
9	75	80	80	86	0.97	0.81	0.98	0.85
10	77	79	80	84	0.97	0.85	0.98	0.89
11	82	84	87	87	0.98	0.90	0.98	0.92
12	73	79	82	86	0.91	0.75	0.92	0.88
13	76	82	84	90	0.96	0.81	0.96	0.80
14	74	78	81	88	0.95	0.75	0.97	0.79

Таблица 8: Показатели устойчивости S (8) для ARTM моделей

№	Z_Φ	Z_Θ	D_1'	D_1''
1	0.72	0.00	73	61
3	0.88	0.86	93	79
4	0.88	0.89	89	74
6	0.88	0.97	89	78
11	0.98	0.92	99	98
12	0.92	0.88	98	92

Таблица 9: Сравнение показателей качества приближения исходных данных полным набором тем D_1' и одной моделью D_1''

татами по построению серии моделей (см. таблицу 9, D_1' — значения функционала D_1 полного набора тем, D_1'' — для одной построенной модели). Такое качество восстановления тем с помощью единственной модели возможно только при достаточной степени разреженности построенных моделей: низкий уровень разреженности приводит к существенному ухудшению результатов одной модели.

Значительную роль в регуляризации матрицы Φ в данном эксперименте сыграл регуляризатор декорреляции тем.

Была выявлена следующая закономерность: чем сильнее разрежена модель, тем меньше разница в качестве восстановления исходных тем с помощью единственной модели и серии моделей.

5.6 Используемая система для экспериментов

Все вычисления производились на компьютере Toshiba Satellite. Использовался процессор 2 ГГц Intel Core i5, 4 Гб оперативной памяти 1600 МГц, операционная система Ubuntu 14.04.2 LTS.

6 Заключение

Было проведено исследование по влиянию разреженности матричного разложения на устойчивость и полноту модели. Оно показало, что приемлимые значения как устойчивости, так и полноты можно добиться только при высокой степени разреженности построенных моделей.

На защиту в данной работе выносятся следующие результаты:

- Предложена формализация понятий устойчивости и полноты вероятностных тематических моделей.
- В экспериментах на синтетических данных показано, что нахождение полного набора тем возможно при большом числе запусков моделирования.
- Предложена стратегия регуляризации, позволяющая сокращать число запусков вплоть до одного.

Список литературы

[1] Greene, D., O'Callaghan, D., Cunningham P.: How many topics? stability analysis for topic models. In: Springer Berlin Heidelberg. pp. 498–513. MIT Press (2014)

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

[3] De Waal, A., Barnard, E.: Evaluating topic models with stability. In: 19th Annual Symposium of the Pattern Recognition Association of South Africa. (2008)

[4] Chemudugunta, C., Smyth, P., Steyvers, M.: Modeling general and specific aspects of documents with a probabilistic topic model. In: Advances in Neural Information Processing Systems. vol. 19, pp. 241–248. MIT Press (2007)

[5] Koltcov, S., Koltsova, O., Nikolenko, S.I.: Latent dirichlet allocation: Stability and applications to studies of user-generated content. In: Proceedings of the 2014 ACM conference on Web science, pp. 161–165 (2014)

[6] Koltcov S., Koltsova, O., Nikolenko, S.I.: Stable topic modeling for web science: granulated LDA. In: Proceedings of the 8th ACM Conference on Web Science, pp. 342–343 (2016)

[7] Steyvers M., Griffiths T.: Finding scientific topics. In: Proceedings of the National Academy of Sciences. vol. 101, no. Suppl. 1., pp. 5228—5235 (2004)

[8] Gaussier E., Goutte C.: Relation between PLSA and NMF and impliations. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conferene on Research and development in information retrieval, pp. 601–602 (2005)

[9] Kuhn H. W.: The Hungarian method for the assignment problem. In: Naval research logistics quarterly, pp. 83–97 (1955)

[10] Tikhonov, A.N., Arsenin, V.Y.: Solution of ill-posed problems. W. H. Winston, Washington, DC (1977)

[11] Vorontsov, K.V., Potapenko, A.A.: Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In: AIST'2014, Springer CCIS vol. 436, pp. 29–46 (2014)

[12] Vorontsov, K.V., Potapenko, A.A.: Additive regularization of topic models. Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications 101(1), 303–323 (2015)

[13] Vorontsov, K., Frei, O., Apishev, M., Romov, P., Suvorova, M., Yanina, A.: Non-bayesian additive regularization for multimodal topic modeling of large collections. In: Proc. of TM '15, pp. 29–37, ACM, New York, NY, USA (2015)

[14] Vorontsov, K.: Additive regularization for topic models of text collections. Doklady Mathematics 89(3), 301–304 (2014)

[15] Плавин А.В., Потапенко А.А., Воронцов К.В.: Энтропийный регуляризатор отбора тем в вероятностных тематических моделях. Математические методы распознавания образов, 228–229 (2015)