# Обзор вероятностных тематических моделей

К. В. Воронцов (vokov@forecsys.ru)

20 июля 2017 г.

**Аннотация**

Тематическое моделирование — это область статистического анализа текстов, активно развивающаяся последние 15 лет. Вероятностная тематическая модель выявляет тематику коллекции текстовых документов, представляя каждую тему частотным словарём слов. Задача тематического моделирования сводится к матричному разложению и имеет бесконечное множество решений. Для доопределения решения вводятся дополнительные критерии — регуляризаторы. Данный обзор описывает в терминах регуляризации основные типы тематических моделей: мультимодальные, мультиязычные, иерархические, темпоральные, сегментирующие, дистрибутивные, графовые. Аддитивная регуляризация тематических моделей (ARTM) позволяет комбинировать регуляризаторы и строить тематические модели с заданными свойствами. Модульная технология моделирования и параллельный онлайновый EM-алгоритм положены в основу библиотеки тематического моделирования с открытым кодом BigARTM.

## 1 Введение

Тематическое моделирование — это одно из современных направлений статистического анализа текстов, активно развивающееся с конца 90-х годов. *Вероятностная тематическая модель* (probabilistic topic model) выявляет тематику коллекции документов, представляя каждую тему дискретным распределением вероятностей слов, а каждый документ — дискретным распределением вероятностей тем.

Тематическое моделирование похоже на кластеризацию документов. Отличие в том, что при кластеризации документ целиком относится к одному кластеру, тогда как тематическая модель осуществляет «мягкую кластеризацию» (soft clustering), относя документ к нескольким кластерам-темам с некоторыми вероятностями. Тематические модели называют также моделями мягкой би-кластеризации, поскольку каждое слово также распределяется по нескольким темам.

Последнее свойство позволяет обходить проблемы синонимии и полисемии слов. Синонимы, взаимозаменяемые в схожих контекстах, группируются в одних и тех же темах. Многозначные слова и омонимы, наоборот, распределяют свои вероятности по нескольким семантически не связанным темам. Например, значение слова «ядро» может быть понято из того, какая тема доминирует в контексте данного слова — математика, физика, биология или военная история.

Вместо отдельных слов модель может строиться на словосочетаниях или терминах. Тема образуется семантически связанными, часто совместно встречающимися

терминами. Такое определение «темы» допускает точную математическую формализацию, но может отличаться от принятых в лингвистике или литературоведении.

Сжатое описание документа в виде вектора вероятностей тем содержит важнейшую информацию о семантике документа и может использоваться для решения многих задач текстовой аналитики. Тематические модели применяются для выявления трендов в новостных потоках, патентных базах, архивах научных публикаций [152, 121], многоязычного информационного поиска [131, 130], поиска тематических сообществ в социальных сетях [154, 123, 97, 27], классификации и категоризации документов [106, 155], тематической сегментации текстов [139], анализа изображений и видеопотоков [49, 66, 42, 122], тегирования веб-страниц [58], обнаружения текстового спама [10], в рекомендательных системах [146, 134, 62, 149, 148], в биоинформатике для анализа нуклеотидных [59] и аминокислотных последовательностей [111, 57], в задачах популяционной генетики [99]. Многие другие разновидности и приложения тематических моделей упоминаются в обзорах [35, 22].

Построение тематической модели по коллекции документов является некорректно поставленной оптимизационной задачей, которая может иметь бесконечное множество решений. Согласно теории регуляризации А. Н. Тихонова [11], решение такой задачи возможно доопределить и сделать устойчивым, добавив к основному критерию *регуляризатор* — дополнительный критерий, учитывающий какие-либо специфические особенности прикладной задачи или знания предметной области. В сложных приложениях дополнительных критериев может быть несколько.

*Аддитивная регуляризация тематических моделей* (additive regularization of topic models, ARTM) — это многокритериальный подход, в котором оптимизируется взвешенная сумма критериев [3, 126]. Большинство известных тематических моделей либо изначально формулируются в терминах регуляризации, либо допускают такую переформулировку. ARTM позволяет отделять регуляризаторы от одних моделей и использовать в других. Это приводит к модульной технологии моделирования. Собрав библиотеку часто используемых регуляризаторов, можно затем их комбинировать, чтобы строить тематические модели с требуемыми свойствами. Оптимизация любых моделей и их комбинаций производится одним и тем же обобщённым EM-алгоритмом. Эти идеи реализованы в библиотеке с открытым кодом `BigARTM`, доступной по адресу `http://bigartm.org`.

ARTM не является ещё одной тематической моделью или ещё одним численным методом. Это общий подход к построению и комбинированию моделей.

В литературе по тематическому моделированию в настоящее время доминируют методы байесовского обучения. Из-за сложности математического аппарата в статьях часто опускаются важные для понимания детали. Иногда авторы ограничиваются упрощённым описанием модели в виде порождающего процесса (generative story) или схематичного представления (plate notation). Последующий переход к алгоритму и его программной реализации остаётся неоднозначным и неочевидным.

Цель данного теоретического обзора — показать разнообразие задач и подходов тематического моделирования, сосредоточившись на первом и очень важном этапе моделирования — как от исходных требований и предположений перейти к формальной постановке оптимизационной задачи. Дальнейшие шаги в ARTM настолько проще байесовского вывода, что изложение удаётся сильно сократить, даже не скрывая математических выкладок. Сопоставимый по охвату и обстоятельности обзор байесовских тематических моделей занял бы сотни страниц.

Вводные разделы 2–4 являются базовыми. В разделах 5–12 в терминах регуляризации описываются разновидности тематических моделей. Эти разделы практически не связаны друг с другом, их можно читать в произвольном порядке или использовать как путеводитель по ссылкам на литературу. Раздел 13 посвящён оцениванию качества. В разделе 14 обсуждается применение тематического моделирования для разведочного информационного поиска. В разделе 15 — краткое заключение.

# 2 Основы тематического моделирования

В этом разделе будут введены основные понятия и постановка задачи тематического моделирования. Будет показан самый простой из известных способов её решения.

**Предварительная обработка текста.** Перед построением тематических моделей текст естественного языка обычно подвергается серии преобразований.

*Лемматизация* — это приведение каждого слова в документе к его нормальной форме. В русском языке нормальными формами считаются: для существительных — именительный падеж, единственное число; для прилагательных — именительный падеж, единственное число, мужской род; для глаголов, причастий, деепричастий — глагол в инфинитиве. Хорошими лемматизаторами для русского языка считаются последние версии `mystem` и `pymorphy`.

*Стемминг* — это отбрасывание окончаний и других изменяемых частей слов. Он подходит для английского языка, для русского предпочтительна лемматизация.

*Стоп-слова* — это частые слова, встречающиеся в текстах любой тематики. Они бесполезны для тематического моделирования и могут быть отброшены. К ним относятся предлоги, союзы, числительные, местоимения, некоторые глаголы, прилагательные и наречия. Число таких слов обычно варьируется в пределах нескольких сотен. Их отбрасывание почти не влияет на объём словаря, но может приводить к заметному сокращению длины некоторых текстов.

*Редкие слова* также рекомендуется отбрасывать, поскольку они не могут повлиять на тематику коллекции. Отбрасывание редких слов, а также строк, не являющихся словами естественного языка (например, чисел), помогает во много раз сокращать объём словаря, снижая затраты времени и памяти на построение моделей.

*Ключевые фразы* — это словосочетания, характерные для предметной области. Их использование вместо отдельных слов или наряду с ними улучшает интерпретируемость тем. Для их выделения можно использовать тезаурусы [8] или методы автоматического выделения терминов (automatic term extraction, ATE), не требующие привлечения экспертов [40, 67, 108].

*Именованные сущности* — это названия объектов реального мира, относящихся к определённым категориям: персоны, организации, геолокации, события, даты, и т. д. Для распознавания именованных сущностей (named entities recognition, NER) используются различные методы машинного обучения [83, 60, 89].

Пусть $D$ — множество (коллекция) текстовых документов, $W$ — множество (словарь) всех употребляемых в них термов. *Термами* могут быть слова, нормальные формы слов, словосочетания или термины, в зависимости от того, какие виды предварительной обработки текстов были выполнены. Каждый документ $d \in D$ представляет собой последовательность $n_d$ термов $w_1, \ldots, w_{n_d}$ из словаря $W$.

**Гипотеза о существовании тем:** каждое вхождение терма $w$ в документ $d$ связано с некоторой темой $t$ из заданного конечного множества $T$. Коллекция документов представляет собой последовательность троек $\Omega_n = \big\{(w_i, d_i, t_i) \;\big|\; i = 1, \ldots, n\big\}$. Термы $w_i$ и документы $d_i$ являются наблюдаемыми переменными, темы $t_i$ не известны и являются *латентными* (скрытыми) переменными.

**Гипотеза «мешка слов»:** порядок термов в документах не важен для выявления тематики, то есть тематику документа можно узнать даже после произвольной перестановки термов, хотя для человека такой текст потеряет смысл. Это предположение называют гипотезой «мешка слов» (bag of words). Порядок документов в коллекции также не имеет значения — это предположение называют гипотезой «мешка документов». Гипотеза «мешка слов» позволяет перейти к компактному представлению документа как *мультимножества* — подмножества $d \subset W$, в котором каждый элемент $w \in d$ повторён $n_{dw}$ раз.

**Гипотеза о вероятностном порождении данных:** множество $\Omega = D \times W \times T$ является конечным *вероятностным пространством* с неизвестной функцией вероятности $p(d, w, t)$. Из этого вероятностного пространства, независимо друг от друга, порождаются тройки: $(d_i, w_i, t_i) \sim p(d, w, t)$. Предположение о независимости является вероятностным уточнением гипотезы «мешка слов». Благодаря этому предположению реализовавшуюся выборку $\Omega_n$ элементов из $\Omega$ можно рассматривать как новое вероятностное пространство с $n$ равновероятными элементарными исходами. В пространстве $\Omega_n$ легко находить вероятности различных событий, и они будут совпадать с частотными оценками вероятностей тех же событий в пространстве $\Omega$.

**Гипотеза условной независимости:** появление слов в документе $d$ по теме $t$ зависит от темы, но не зависит от документа $d$, и описывается общим для всех документов распределением $p(w\,|\,t)$:

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

**Вероятностная тематическая модель порождения текста.** Согласно формуле полной вероятности и гипотезе условной независимости, распределение термов в документе $p(w\,|\,d)$ описывается *вероятностной смесью* распределений термов в темах $\varphi_{wt} = p(w\,|\,t)$ с весами $\theta_{td} = p(t\,|\,d)$:

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

Вероятностная модель (2) описывает процесс порождения коллекции по известным распределениям $p(w\,|\,t)$ и $p(t\,|\,d)$. Этот процесс показан в алгоритме 1 и на рис. 1.

**Построение тематической модели** — это обратная задача: по заданной коллекции $D$ требуется найти параметры $\varphi_{wt}$ и $\theta_{td}$, при которых тематическая модель (2) хорошо приближает частотные оценки условных вероятностей $\hat{p}(w\,|\,d) = \frac{n_{dw}}{n_d}$.

Распределение вида $p(t\,|\,x)$ будем называть *тематикой* объекта $x$. Можно говорить о тематике документа $p(t\,|\,d)$, терма $p(t\,|\,w)$, терма в документе $p(t\,|\,d, w)$.

**Алгоритм 1.** Вероятностный процесс порождения коллекции документов.

**Вход**: распределения $p(w\,|\,t)$, $p(t\,|\,d)$;
**Выход**: выборка пар $(d_i, w_i)$, $i = 1, \ldots, n$;

**для всех** $d \in D$
  задать длину $n_d$ документа $d$;
  **для всех** $i = 1, \ldots, n_d$
    $d_i := d$;
    выбрать случайную тему $t_i$ из распределения $p(t\,|\,d_i)$;
    выбрать случайный терм $w_i$ из распределения $p(w\,|\,t_i)$;

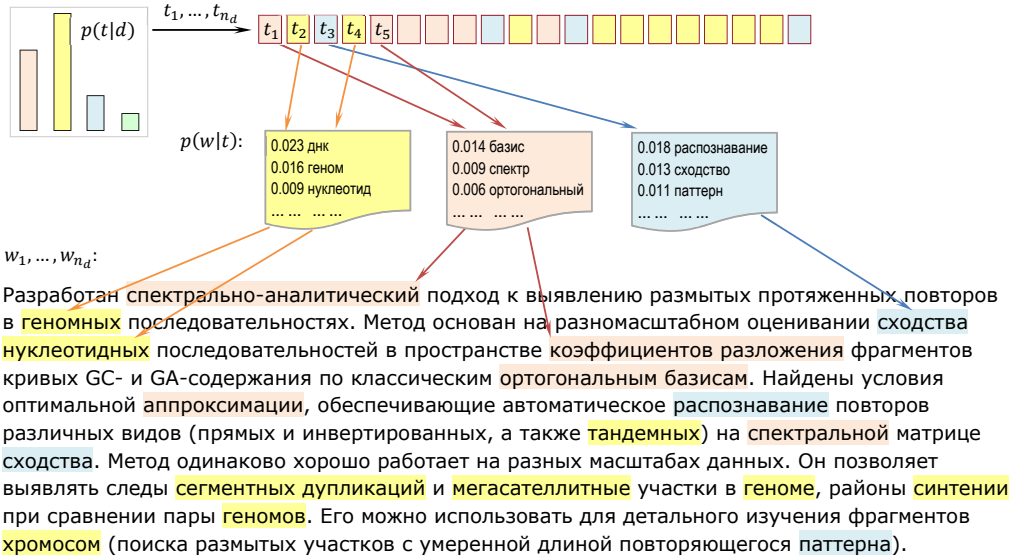



Рис. 1: Процесс порождения текстовой коллекции вероятностной тематической моделью (2): в каждой позиции $i$ документа $d_i$ сначала порождается тема $t_i \sim p(t\,|\,d_i)$, затем терм $w_i \sim p(w\,|\,t_i)$.

Целью тематического моделирования является определение тематики документов и связанных с ними объектов. Также требуется находить распределения $\varphi_{wt} = p(w\,|\,t)$, описывающие семантику каждой темы $t$ словами естественного языка.

**Частотные оценки условных вероятностей.** В пространстве $\Omega_n$ вероятности, выражающиеся через переменные $d$ и $w$, совпадают с частотами соответствующих наблюдаемых событий:

$$p(d,w) = \frac{n_{dw}}{n}, \qquad p(d) = \frac{n_d}{n}, \qquad p(w) = \frac{n_w}{n}, \qquad p(w\,|\,d) = \frac{n_{dw}}{n_d}; \tag{3}$$

$n_{dw}$ — число вхождений терма $w$ в документ $d$;
$n_d = \sum_w n_{dw}$ — длина документа $d$ в термах;
$n_w = \sum_d n_{dw}$ — число вхождений терма $w$ во все документы коллекции;
$n = \sum_d \sum_w n_{dw}$ — длина коллекции в термах.

Вероятности, связанные со скрытой переменной $t$, тоже определяются как частоты:

$$p(t) = \frac{n_t}{n}, \qquad p(w\,|\,t) = \frac{n_{wt}}{n_t}, \qquad p(t\,|\,d) = \frac{n_{td}}{n_d}, \qquad p(t\,|\,d,w) = \frac{n_{tdw}}{n_{dw}}; \tag{4}$$

$n_{tdw}$ — число троек, в которых терм $w$ документа $d$ связан с темой $t$;
$n_{td} = \sum_w n_{tdw}$ — число троек, в которых терм документа $d$ связан с темой $t$;
$n_{wt} = \sum_d n_{tdw}$ — число троек, в которых терм $w$ связан с темой $t$;
$n_t = \sum_d \sum_w n_{tdw}$ — число троек, связанных с темой $t$.

В отличие от (3), эти частотные оценки не могут быть вычислены непосредственно по исходным данным, так как темы $t_i$ неизвестны.

Согласно закону больших чисел, при $n \to \infty$ частотные оценки, определяемые формулами (3)–(4), стремятся к соответствующим вероятностям в пространстве $\Omega$.

**EM-алгоритм.** Заметим, что все оценки (4) выражаются через $n_{tdw} = n_{dw} p(t\,|\,d,w)$. Зная условные распределения $p(t\,|\,d,w)$, можно оценить искомые параметры тематической модели $\varphi_{wt} = p(w\,|\,t)$ и $\theta_{td} = p(t\,|\,d)$. И, наоборот, зная параметры модели, можно выразить условные вероятности $p(t\,|\,d,w)$ по формуле Байеса:

$$p(t\,|\,d,w) = \frac{p(t,w\,|\,d)}{p(w\,|\,d)} = \frac{p(w\,|\,t)p(t\,|\,d)}{p(w\,|\,d)} = \frac{\varphi_{wt}\theta_{td}}{\sum\limits_{s\in T} \varphi_{ws}\theta_{sd}}.$$

Таким образом, получаем систему нелинейных уравнений относительно параметров модели $\varphi_{wt}$, $\theta_{td}$ и вспомогательных переменных $p_{tdw}$, $n_{wt}$, $n_{td}$:

$$p_{tdw} = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}}; \tag{5}$$

$$\varphi_{wt} = \frac{n_{wt}}{\sum_{w'} n_{w't}}; \qquad n_{wt} = \sum_{d\in D} n_{dw} p_{tdw}; \tag{6}$$

$$\theta_{td} = \frac{n_{td}}{\sum_{t'} n_{t'd}}; \qquad n_{td} = \sum_{w\in d} n_{dw} p_{tdw}. \tag{7}$$

Для её решения удобно применять метод простых итераций: сначала выбираются начальные приближения параметров $\varphi_{wt}$ и $\theta_{td}$, по ним вычисляются вспомогательные переменные $p_{tdw}$, которые позволяют найти следующее приближение параметров $\varphi_{wt}$ и $\theta_{td}$. Вычисления по формулам (5)–(7) продолжаются в цикле до сходимости.

Этот итерационный процесс является частным случаем EM-алгоритма, предназначенного для определения параметров вероятностных моделей со скрытыми переменными [37]. В терминах EM-алгоритма вычисление условных распределений скрытых переменных по формуле (5) называется E-шагом (expectation), а вычисление параметров модели по формулам (6)–(7) — M-шагом (maximization).

Далее мы выведем EM-алгоритм из общей оптимизационной постановки задачи. Сейчас мы пришли к нему самым простым и интуитивным путём, но этот путь не даёт ответов на вопросы, сходится ли алгоритм к решению системы уравнений, и почему эта система описывает тематическую модель, приближающую $\hat{p}(w\,|\,d)$.

**Алгоритм 2.** Рациональный EM-алгоритм для тематической модели (2).

**Вход**: коллекция $D$, число тем $|T|$, начальные приближения $\{\varphi_{wt}\}$, $\{\theta_{td}\}$;
**Выход**: параметры модели $\{\varphi_{wt}\}$ и $\{\theta_{td}\}$;

**повторять**
  обнулить $n_{wt}$, $n_{td}$, $n_t$ для всех $d \in D,\ w \in W,\ t \in T$;
  **для всех** $d \in D,\ w \in d$
    $n_{tdw} := n_{dw}\varphi_{wt}\theta_{td} \,/\, \sum_\tau \varphi_{w\tau}\theta_{\tau d}$;
    увеличить $n_{wt}$, $n_{td}$, $n_t$ на $n_{tdw}$ для всех $t \in T$;
  $\varphi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := n_{td}/n_d$ для всех $d \in D,\ t \in T$;
**пока** $\{\varphi_{wt}\}$ и $\{\theta_{td}\}$ не сойдутся;

**Рациональный EM-алгоритм.** Вычисление переменных $n_{wt}$, $n_{td}$, $n_t$ на M-шаге требует однократного прохода всей коллекции в цикле по всем документам $d \in D$ и всем термам $w \in d$. Внутри этого цикла переменные $p_{tdw}$ можно вычислять только в тот момент, когда они нужны. От этого результат алгоритма не изменяется, E-шаг встраивается внутрь M-шага без дополнительных вычислительных затрат, отпадает необходимость хранения трёхмерной матрицы $p_{tdw}$. Этот вариант реализации EM-алгоритма будем называть *рациональным*; он показан в алгоритме 2.

# 3 Оптимизация и регуляризация

В этом разделе будет описан общий оптимизационный подход к тематическому моделированию. Разработанный здесь формализм будет использоваться на протяжении всего дальнейшего изложения для построения различных тематических моделей.

**Принцип максимума правдоподобия** используется в математической статистике для оценивания неизвестных параметров вероятностных моделей по наблюдаемым данным. Согласно этому принципу, выбираются такие значения параметров, при которых наблюдаемая выборка наиболее правдоподобна.

*Функция правдоподобия* определяется как зависимость вероятности выборки от параметров модели. Благодаря предположению о независимости наблюдений, она равна произведению вероятностей слов в документах:

$$p\big((d_i, w_i)_{i=1}^n; \Phi, \Theta\big) = \prod_{i=1}^n p(d_i, w_i) = \prod_{d\in D}\prod_{w\in d} p(w\,|\,d)^{n_{dw}} \underbrace{p(d)^{n_{dw}}}_{\text{const}} \to \max_{\Phi,\Theta}.$$

Прологарифмировав правдоподобие, перейдём от произведения к сумме и отбросим слагаемые, не зависящие от параметров модели. Получим задачу максимизации log-правдоподобия при ограничениях неотрицательности и нормировки:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \ \to \max_{\Phi,\Theta}; \tag{8}$$

$$\sum_{w\in W} \varphi_{wt} = 1; \quad \varphi_{wt} \geqslant 0; \qquad \sum_{t\in T} \theta_{td} = 1; \quad \theta_{td} \geqslant 0. \tag{9}$$

**Низкоранговое матричное разложение.** Обычно число тем $|T|$ много меньше $|D|$ и $|W|$, и задача сводится к поиску приближённого представления заданной матрицы частот термов в документах $P = \big(\hat{p}(w\,|\,d)\big)_{W\times D}$ в виде произведения $P = \Phi\Theta$ двух неизвестных матриц меньшего размера — *матрицы термов тем* $\Phi = (\varphi_{wt})_{W\times T}$ и *матрицы тем документов* $\Theta = (\theta_{td})_{T\times D}$. Все три матрицы $P, \Phi, \Theta$ являются *стохастическими*, то есть имеют неотрицательные нормированные столбцы $p_d, \varphi_t, \theta_d$, представляющие дискретные распределения.

**Регуляризация.** Задача называется *корректно поставленной* по Адамару, если её решение существует, единственно и устойчиво.

Задача стохастического матричного разложения является некорректно поставленной, так как множество её решений в общем случае бесконечно. Если $\Phi\Theta$ — решение, то $(\Phi S)(S^{-1}\Theta)$ также является решением для всех невырожденных матриц $S$, при условии, что матрицы $\Phi S$ и $S^{-1}\Theta$ — стохастические.

Существует общий подход к решению некорректно поставленных обратных задач, называемый *регуляризацией* [11]. Когда оптимизационная задача недоопределена, к основному критерию добавляют дополнительный критерий — регуляризатор, учитывающий специфику решаемой задачи и знания предметной области. В практических задачах автоматической обработки текстов дополнительных критериев и ограничений на решение может быть много.

**Многокритериальная оптимизация.** *Аддитивная регуляризация тематических моделей* (ARTM) [3] основана на максимизации линейной комбинации логарифма правдоподобия и нескольких *регуляризаторов* $R_i(\Phi,\Theta)$, $i = 1,\dots,k$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta); \qquad (10)$$

при ограничениях (9), где $\tau_i$ — неотрицательные *коэффициенты регуляризации*. Преобразование вектора критериев в один скалярный критерий — это приём, широко используемый в многокритериальной оптимизации и называемый *скаляризацией*.

Задача (10), (9) относится к классу невыпуклых задач математического программирования. Для неё возможно найти лишь локальный экстремум, качество которого зависит от начального приближения. На практике поиск глобального экстремума не столь важен, как адекватная формализация дополнительных критериев и поиск компромисса между этими критериями.

**Необходимые условия максимума.** Введём оператор norm, который преобразует произвольный заданный вектор $(x_i)_{i\in I}$ в вектор вероятностей $(p_i)_{i\in I}$ дискретного распределения путём обнуления отрицательных элементов и нормировки:

$$p_i = \operatorname*{norm}_{i\in I}(x_i) = \frac{(x_i)_+}{\sum\limits_{j\in I}(x_j)_+}, \text{ для всех } i \in I,$$

где $(x)_+ = \max\{0, x\}$ — операция положительной срезки. Если $x_i \leqslant 0$ для всех $i \in I$, то результатом оператора norm является нулевой вектор.

**Теорема 1.** *Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Тогда точка $(\Phi, \Theta)$ локального экстремума задачи (10), (9) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}$*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{11}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{12}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}; \tag{13}$$

*для всех тем $t$ и документов $d$, не вырожденных в смысле следующих определений:*

*а) модель темы $t$ вырождена, если $n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} \leqslant 0$ для всех термов $w \in W$;*

*б) модель документа $d$ вырождена, если $n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \leqslant 0$ для всех тем $t \in T$.*

Вырожденность возникает в тех редких на практике случаях, когда регуляризатор $R$ оказывает чрезмерное разреживающее воздействие на параметры модели. Вырожденные темы и документы исключаются из модели. Сокращение числа тем может быть желательным побочным эффектом регуляризации. Вырожденность документа может означать, что модель не в состоянии его описать, например, если он слишком короткий или не соответствует тематике коллекции.

**Доказательство.** Запишем необходимые условия Каруша–Куна–Таккера локального экстремума задачи (10), (9):

$$\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w\,|\,d)} + \frac{\partial R}{\partial \varphi_{wt}} = \lambda_t - \lambda_{wt}; \quad \lambda_{wt} \geqslant 0; \quad \lambda_{wt}\varphi_{wt} = 0; \tag{14}$$

$$\sum_{w\in W} n_{dw}\frac{\varphi_{wt}}{p(w\,|\,d)} + \frac{\partial R}{\partial \theta_{td}} = \mu_d - \mu_{td}; \quad \mu_{td} \geqslant 0; \quad \mu_{td}\theta_{td} = 0; \tag{15}$$

где множители Лагранжа $\lambda_t$, $\mu_d$ соответствуют ограничениям нормировки, $\lambda_{wt}$, $\mu_{td}$ — ограничениям неотрицательности.

Умножим обе части равенства (14) на $\varphi_{wt}$, обе части равенства (15) на $\theta_{td}$, и выделим вспомогательные переменные $p_{tdw}$, $n_{wt}$ и $n_{td}$:

$$\varphi_{wt}\lambda_t = \sum_{d\in D} n_{dw}\frac{\varphi_{wt}\theta_{td}}{p(w\,|\,d)} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} = n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}};$$

$$\theta_{td}\mu_d = \sum_{w\in W} n_{dw}\frac{\varphi_{wt}\theta_{td}}{p(w\,|\,d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} = n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}.$$

Предположение $\lambda_t \leqslant 0$ противоречит условию невырожденности темы $t$. Поэтому рассмотрим только случай, когда $\lambda_t > 0$. Тогда либо $\varphi_{wt} = 0$, либо обе части равенства положительны. Объединим эти два случая в одну формулу:

$$\varphi_{wt}\lambda_t = \left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right)_+. \tag{16}$$

Аналогично, предположение $\mu_d \leqslant 0$ противоречит условию невырожденности документа $d$. Поэтому рассмотрим только случай, когда $\mu_d > 0$. Тогда либо $\theta_{td} = 0$, либо обе части равенства положительны. Объединим эти два случая в одну формулу:

$$\theta_{td}\mu_d = \Big(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big)_+. \tag{17}$$

Суммируем левую и правую части равенства (16) по $w$, равенства (17) по $t$:

$$\lambda_t = \sum_{w \in W}\Big(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\Big)_+; \tag{18}$$

$$\mu_d = \sum_{t \in T}\Big(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big)_+. \tag{19}$$

Подставляя $\lambda_t$ из (18) в (16), получим (12).
Подставляя $\mu_d$ из (19) в (17), получим (13).
Теорема доказана.

**Модель вероятностного латентного семантического анализа** (probabilistic latent semantic analysis, PLSA) — это первая вероятностная тематическая модель, предложенная Томасом Хофманном в 1999 году [48]. В ARTM она соответствует нулевому регуляризатору, $R(\Phi, \Theta) = 0$. В этом случае система (11)–(13) совпадает с системой (5)–(7), которую мы получили ранее из элементарных соображений.

Переменные $p_{tdw}$ вычисляются по формуле Байеса (5) при любой регуляризации.

**EM-алгоритм с регуляризацией** является применением метода простых итераций для решения системы (11)–(13). Сначала выбираются начальные приближения $\varphi_{wt}, \theta_{td}$, затем в цикле до сходимости чередуются *E-шаг* (11) и *M-шаг* (12)–(13).

Известно, что EM-алгоритм без регуляризации сходится в слабом смысле: на каждой итерации правдоподобие увеличивается [37]. Аналогичные условия слабой сходимости для ARTM получены И. А. Ирхиным[1]. Разновидности EM-алгоритма для тематического моделирования рассматриваются в [18, 5].

**Онлайновый EM-алгоритм** считается наиболее быстрым и хорошо распараллеливается [47, 20]. Основная его идея в том, что на больших коллекциях матрица Φ обычно сходится после обработки относительно небольшой доли документов. Даже одного прохода по коллекции бывает достаточно для построения модели. Это позволяет применять онлайновые алгоритмы для анализа потоковых данных.

Алгоритм 3 показывает организацию вычислительного процесса для коллекции $D$, разбитой на пакеты документов $D_b$, $b = 1, \ldots, B$. Обработка каждого пакета выполняется функцией `ProcessBatch`. Для каждого документа $d$ из пакета $D_b$ производятся итерации вектора $\theta_d$ со встроенным E-шагом при фиксированной матрице Φ. На последней итерации документа обновляются счётчики $\tilde{n}_{wt}$ текущего пакета. В моменты синхронизации происходит объединение обновлений $\tilde{n}_{wt}$, накопленных

[1] Илья Ирхин. Сходимость численных методов вероятностного тематического моделирования. 2016. ФИВТ МФТИ.
`http://www.MachineLearning.ru/wiki/images/0/05/Irkhin2016msc.pdf`

**Алгоритм 3.** Онлайновый EM-алгоритм для ARTM

**Вход**: коллекция $D_b$, весовые коэффициенты $k_{\texttt{decay}}$ и $k_{\texttt{apply}}$;
**Выход**: матрица $\Phi$;

1 инициализировать $\varphi_{wt}$ для всех $w \in W,\ t \in T$;
2 $n_{wt} := 0,\ \tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;
3 **для всех** пакетов $D_b$, $b = 1, \ldots, B$
4 $\quad (\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \texttt{ProcessBatch}\,(D_b, \Phi)$;
5 $\quad$ **если** пора выполнить синхронизацию, **то**
6 $\quad\quad n_{wt} := k_{\texttt{decay}} n_{wt} + k_{\texttt{apply}} \tilde{n}_{wt}$ для всех $w \in W,\ t \in T$;
7 $\quad\quad \varphi_{wt} := \underset{w \in W}{\operatorname{norm}}\big(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\big)$ для всех $w \in W,\ t \in T$;
8 $\quad\quad \tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;

9 **функция** $\texttt{ProcessBatch}$ (пакет $D_b$, матрица $\Phi$) $\mapsto$ матрица $(\tilde{n}_{wt})$;
10 $\tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;
11 **для всех** $d \in D_b$
12 $\quad$ инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
13 $\quad$ **повторять**
14 $\quad\quad p_{tdw} := \underset{t \in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big)$ для всех $w \in d,\ t \in T$;
15 $\quad\quad \theta_{td} := \underset{t \in T}{\operatorname{norm}}\big(\sum\limits_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\big)$ для всех $t \in T$;
16 $\quad$ **пока** $\theta_d$ не сойдётся;
17 $\quad \tilde{n}_{wt} := \tilde{n}_{wt} + n_{dw} p_{tdw}$ для всех $w \in d,\ t \in T$;

в результате обработки нескольких пакетов, в матрице $\Phi$, см. шаги 5–8. Коэффициенты $k_{\texttt{decay}}$ и $k_{\texttt{apply}}$ позволяют управлять темпом забывания предыдущих пакетов.

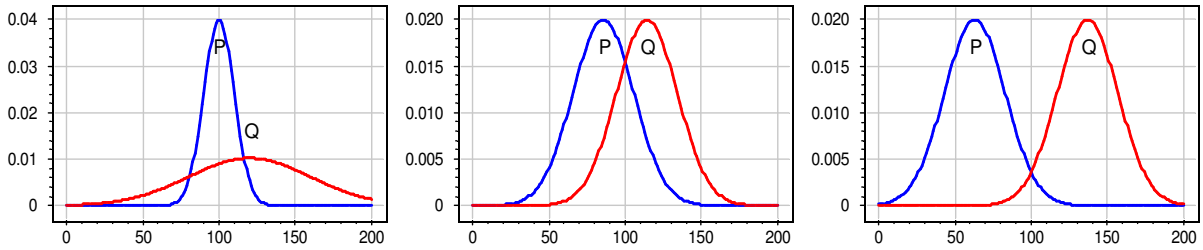
В онлайновом алгоритме можно хранить матрицу $\Phi$ в оперативной памяти, а матрицу $\Theta$ вообще не хранить. Тематическую модель документа можно получать «на лету» и сразу использовать. Коэффициенты регуляризации задаются в момент создания модели, но потом могут быть в любой момент изменены, в том числе в ходе EM-итераций. Детали параллельной реализации оффлайнового и онлайнового EM-алгоритма в библиотеке `BigARTM` описаны в [43].

**Дивергенция Кульбака–Лейблера** (*KL-дивергенция*, относительная энтропия) далее будет одним из важнейших инструментов конструирования регуляризаторов. Это несимметричная функция расстояния между дискретными распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$, с совпадающими носителями, $\{i\colon p_i > 0\} = \{i\colon q_i > 0\}$:

$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} = H(P,Q) - H(P),$$

где $H(P) = -\sum_i p_i \ln p_i$ — *энтропия* распределения $P$, $H(P,Q) = -\sum_i p_i \ln q_i$ — *кросс-энтропия* распределений $P$ и $Q$.

KL-дивергенция неотрицательна и равна нулю тогда и только тогда, когда распределения совпадают, $p_i \equiv q_i$.




$\mathrm{KL}(P\|Q) = 0.44$ $\mathrm{KL}(Q\|P) = 2.97$ | $\mathrm{KL}(P\|Q) = 0.44$ $\mathrm{KL}(Q\|P) = 0.44$ | $\mathrm{KL}(P\|Q) = 2.97$ $\mathrm{KL}(Q\|P) = 2.97$

Рис. 2: Дивергенция Кульбака–Лейблера $\mathrm{KL}(P\|Q)$ является несимметричной мерой вложенности распределения $P = (p_i)_{i=1}^n$ в распределение $Q = (q_i)_{i=1}^n$. Вложенность $P$ в $Q$ приблизительно одинакова на левом и среднем графиках, вложенность $Q$ в $P$ — на левом и правом графиках.

Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то распределение $P$ сильнее вложено в $Q$, чем $Q$ в $P$, см. рис. 2. Таким образом, KL является мерой вложенности двух распределений.

Если $P$ — эмпирическое распределение, а $Q(\alpha)$ — параметрическая модель, то минимизация KL-дивергенции эквивалентна минимизации кросс-энтропии и максимизации правдоподобия:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_{\alpha} \quad \Longleftrightarrow \quad \sum_{i=1}^{n} p_i \ln q_i(\alpha) \to \max_{\alpha}.$$

Максимизация правдоподобия (8) эквивалентна минимизации взвешенной суммы KL-дивергенций между эмпирическими распределениями $\hat{p}(w\,|\,d) = \frac{n_{dw}}{n_d}$ и модельными $p(w\,|\,d)$, по всем документам $d$ из $D$:

$$\sum_{d\in D} n_d \,\mathrm{KL}_w\Big(\frac{n_{dw}}{n_d} \,\Big\|\, \sum_{t\in T} \varphi_{wt}\theta_{td}\Big) \;\to\; \min_{\Phi,\Theta},$$

где весом документа $d$ является его длина $n_d$. Если веса $n_d$ убрать, то все документы будут искусственно приведены к одинаковой длине. Такая модификация функционала качества может быть полезна при моделировании коллекций, содержащих документы одинаковой важности, но существенно разной длины.

# 4 Байесовская регуляризация и модель LDA

**Байесовская регуляризация.** До сих пор мы предполагали, что данные порождаются вероятностной моделью с параметрами $(\Phi, \Theta)$, которые не известны и не случайны. В байесовском подходе предполагается, что параметры также случайны и подчиняются некоторому *априорному* распределению $p(\Phi, \Theta; \gamma)$ с неслучайным *гиперпараметром* $\gamma$. В этом случае максимизация совместного правдоподобия данных и модели приводит к принципу *максимума апостериорной вероятности* (maximum a posteriori probability, MAP):

$$p(D, \Phi, \Theta; \gamma) = p(D\,|\,\Phi, \Theta)\, p(\Phi, \Theta; \gamma) = p(\Phi, \Theta; \gamma) \prod_{i=1}^{n} p(d_i, w_i\,|\,\Phi, \Theta) \;\to\; \max_{\Phi,\Theta,\gamma}.$$

После логарифмирования получаем модификацию задачи (8), в которой логарифм априорного распределения является регуляризатором:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + \underbrace{\ln p(\Phi,\Theta;\gamma)}_{R(\Phi,\Theta)} \;\to\; \max_{\Phi,\Theta,\gamma}. \tag{20}$$

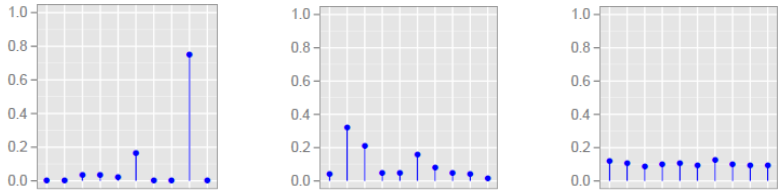
В байесовском подходе применяется также принцип максимизации неполного правдоподобия, в котором по случайным параметрам $(\Phi,\Theta)$ производится интегрирование и оптимизируются гиперпараметры $\gamma$. Считается, что этот приём снижает размерность задачи и риск переобучения. Действительно, размерность вектора $\gamma$, как правило, много меньше размеров матриц $\Phi,\Theta$ и не зависит от объёма коллекции. Однако для решения прикладных задач всё равно нужны именно эти матрицы. Формулы для них выводятся громоздкими приближёнными методами, но в итоге мало отличаются от MAP-оценок [18].

В байесовском подходе оцениваются не сами параметры $\Phi,\Theta$, а их апостериорное распределение $p(\Phi,\Theta\,|\,D;\gamma)$. Для задач тематического моделирования в этом нет особого смысла. На практике полученное распределение используется исключительно для того, чтобы вернуться к точечным оценкам математического ожидания. Другие оценки используются крайне редко, даже точечные оценки медианы или моды.

Техники приближённого байесовского вывода (вариационный вывод [120], сэмплирование Гиббса [116], распространение ожидания) не позволяют легко комбинировать модели и добавлять регуляризаторы, не имеющие вероятностной интерпретации. Для каждой новой модели приходится заново выполнять математические выкладки и программную реализацию. В прикладных проектах сроки, стоимость и риски таких разработок становятся непреодолимым барьером. Поэтому на практике пользуются простой устаревшей моделью LDA, а байесовское тематическое моделирование редко выходит за рамки академических исследований. Тем не менее, в литературе по тематическому моделированию байесовский подход доминирует.

Многокритериальный не-байесовский подход ARTM — это попытка изменить ситуацию. Байесовские тематические модели в большинстве случаев удаётся переформулировать в терминах регуляризации, записав постановку задачи в виде (20). С этого момента регуляризатор отделяется от модели и может быть использован в других моделях. Это приводит к модульной технологии тематического моделирования, которая реализована и развивается в проекте `BigARTM`.

**Латентное размещение Дирихле.** Дэвид Блэй, Эндрю Ын и Майкл Джордан предложили модель LDA (latent Dirichlet allocation) [26] для решения проблемы переобучения в PLSA, которая предсказывала вероятности слов $p(w\,|\,d)$ на новых документах заметно хуже, чем на обучающей коллекции. Позже выяснилось, что на больших коллекциях обе модели почти не переобучаются, а их правдоподобия отличаются незначительно [73, 143, 69]. Различия проявляются только на низкочастотных термах, которые не важны для образования тем. В робастных вариантах PLSA и LDA такие термы игнорируются, что резко снижает как переобучение, так и различие в правдоподобии моделей [98]. Сам вопрос о переобучении поставлен не вполне корректно. Во-первых, тематические модели строятся не ради предсказания слов в документах, а для выявления латентной кластерной структуры коллекции. Во-вторых, переобучение зависит не столько от самой модели, сколько от того, как мы догово-




$\mathrm{Dir}(\varphi; 0.1)$ $\mathrm{Dir}(\varphi; 1)$ (равномерное) $\mathrm{Dir}(\varphi; 100)$

Рис. 3: Пример трёх неотрицательных нормированных векторов $\varphi_t \in \mathbb{R}^{10}$, порождённых соответственно тремя симметричными распределениями Дирихле с параметрами 0.1, 1, 100.

римся измерять её качество. Для измерения обычно используется перплексия, которая сильно штрафует заниженные вероятности низкочастотных термов. Тем не менее, LDA до сих пор считается моделью №1 в тематическом моделировании, а про PLSA вспоминают всё реже.

Модель LDA основана на предположении, что столбцы $\theta_d$ и $\varphi_t$ являются случайными векторами, которые порождаются распределениями Дирихле с параметрами $\alpha \in \mathbb{R}^{|T|}$ и $\beta \in \mathbb{R}^{|W|}$ соответственно:

$$\mathrm{Dir}(\theta_d; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \qquad \alpha_t > 0, \quad \alpha_0 = \sum_t \alpha_t, \qquad \theta_{td} > 0, \quad \sum_t \theta_{td} = 1;$$

$$\mathrm{Dir}(\varphi_t; \beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \varphi_{wt}^{\beta_w - 1}, \quad \beta_w > 0, \quad \beta_0 = \sum_w \beta_w, \quad \varphi_{wt} > 0, \quad \sum_w \varphi_{wt} = 1;$$

где $\Gamma(z)$ — гамма-функция. Параметры распределения Дирихле связаны с математическим ожиданием порождаемых случайных векторов: $\mathsf{E}\theta_{td} = \frac{\alpha_t}{\alpha_0}$, $\mathsf{E}\varphi_{wt} = \frac{\beta_w}{\beta_0}$.

Распределения Дирихле способны порождать как разреженные, так и плотные векторы дискретных распределений, рис. 3. Чем меньше $\beta_w$, тем более разрежена соответствующая $w$ компонента $\varphi_{wt}$ в порождаемых векторах $\varphi_t$. Если вектор параметров состоит из равных значений $\beta_w$, то распределение Дирихле называется *симметричным*. При $\beta_w \equiv 1$ симметричное распределение Дирихле совпадает с равномерным распределением на единичном симплексе.

Тематическая модель порождения данных является двухуровневой: сначала из распределения Дирихле порождаются вектор-столбцы $\varphi_t$, которые задают темы. Затем из полученных распределений $p(w\,|\,t) = \varphi_{wt}$ порождаются слова, которые образуют монотематичные части документов $d$, описываемые эмпирическими распределениями $\hat{p}(w\,|\,t, d)$. Таким образом, двухуровневая модель порождения текста способна описывать кластерные структуры в текстовых коллекциях. Векторы распределений $p(w\,|\,t)$ интерпретируются как центроиды кластеров, а распределения $\hat{p}(w\,|\,t, d)$ являются точками этих кластеров.

Более убедительных лингвистических обоснований распределение Дирихле не имеет. Его широкое распространение в тематическом моделировании объясняется скорее математическим удобством и популярностью байесовского обучения. Распределение Дирихле является сопряжённым к мультиномиальному распределению, что существенно упрощает байесовский вывод. Благодаря этому свойству оно оказывается «на особом положении» в байесовском тематическом моделировании, и большинство моделей строятся с использованием распределений Дирихле.

Согласно (20), модели LDA соответствует регуляризатор, с точностью до константы равный логарифму априорного распределения Дирихле:

$$R(\Phi,\Theta) = \ln\prod_{t\in T}\mathrm{Dir}(\varphi_t;\beta)\prod_{d\in D}\mathrm{Dir}(\theta_d;\alpha) + \mathrm{const} = \\ = \sum_{t\in T}\sum_{w\in W}(\beta_w - 1)\ln\varphi_{wt} + \sum_{d\in D}\sum_{t\in T}(\alpha_t - 1)\ln\theta_{td}. \qquad (21)$$

Применение уравнений (12)–(13) к этому регуляризатору даёт формулы М-шага:

$$\varphi_{wt} = \underset{w\in W}{\mathrm{norm}}\big(n_{wt} + \beta_w - 1\big); \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}\big(n_{td} + \alpha_t - 1\big).$$

При $\beta_w = 1,\ \alpha_t = 1$ априорное распределение Дирихле совпадает с равномерным распределением на симплексе, формулы М-шага переходят в несмещённые частотные оценки условных вероятностей, а модель LDA переходит в PLSA [44].

При $\beta_w > 1,\ \alpha_t > 1$ регуляризатор имеет сглаживающий эффект: он делает большие вероятности ещё больше, при этом малые вероятности за счёт нормировки становятся меньше, однако никогда не достигают нуля.

При $0 < \beta_w < 1,\ 0 < \alpha_t < 1$ регуляризатор имеет разреживающий эффект и способен обнулять малые вероятности.

**Не-вероятностная интерпретация модели LDA.** Регуляризатор (21) можно эквивалентным образом записать через KL-дивергенции:

$$R(\Phi,\Theta) = |W|\sum_{t\in T}\mathrm{KL}_w\big(\tfrac{1}{|W|}\,\big\|\,\varphi_{wt}\big) - \beta_0\sum_{t\in T}\mathrm{KL}_w\big(\tfrac{\beta_w}{\beta_0}\,\big\|\,\varphi_{wt}\big) + \\ + |T|\sum_{d\in D}\mathrm{KL}_t\big(\tfrac{1}{|T|}\,\big\|\,\theta_{td}\big) - \alpha_0\sum_{d\in D}\mathrm{KL}_t\big(\tfrac{\alpha_t}{\alpha_0}\,\big\|\,\theta_{td}\big).$$

Отсюда следует, что модель LDA оказывает сглаживающие и разреживающие воздействия на матрицы $\Phi, \Theta$. Все столбцы матрицы $\Phi$ должны быть близки к одному и тому же распределению $\frac{\beta_w}{\beta_0}$, причём параметр $\beta_0$ становится коэффициентом регуляризации. Аналогично, все столбцы матрицы $\Theta$ должны быть близки к распределению $\frac{\alpha_t}{\alpha_0}$, и этим требованием управляет коэффициент регуляризации $\alpha_0$. Кроме этих сглаживающих воздействий имеются слабые неуправляемые разреживающие воздействия: столбцы обеих матриц должны быть далеки от равномерного распределения. Дальше всего от равномерного распределения находятся вырожденные распределения, в которых единичная вероятность сконцентрирована в единственном элементе. Поэтому разреживание приводит к обнулению малых вероятностей в матрицах $\Phi, \Theta$.

# 5 Интерпретируемость тем

Отказ от априорных распределений Дирихле позволяет обобщить модель LDA: снять ограничения на знаки гиперпараметров в (21) и свободнее обращаться со сглаживанием и разреживанием для улучшения интерпретируемости тематических моделей.

*Гипотеза разреженности* является одним из естественных необходимых условий интерпретируемости. Предполагается, что каждая тема характеризуется небольшим

числом термов, и каждый документ относится к небольшому числу тем. В таком случае значительная часть вероятностей $\varphi_{wt}$ и $\theta_{td}$ должны быть равны нулю.

Многократные попытки разреживания модели LDA приводили к чрезмерно сложным конструкциям [110, 39, 133, 61, 32] из-за внутреннего противоречия между требованиями разреженности и ограничениями строгой положительности параметров в распределении Дирихле. Проблема решается неожиданно просто, если оставить кросс-энтропийный регуляризатор (21) и разрешить гиперпараметрам $\alpha_t, \beta_w$ принимать любые значения, включая отрицательные. По всей видимости, впервые она была предложена в динамической модели PLSA для обработки видеопотоков [122], где документами являлись короткие видеофрагменты, термами — признаки на изображениях, темами — появление определённого объекта в течение определённого времени, например, проезд автомобиля. Сильно разреженные распределения потребовались для описания тем с кратким «временем жизни».

**Сглаживание и разреживание.** По аналогии с (21) введём обобщённый регуляризатор сглаживания и разреживания:

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \varphi_{wt} + \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td}.$$

Подставив этот регуляризатор в (12)–(13), получим формулы M-шага:

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\big(n_{wt} + \beta_{wt}\big); \qquad \theta_{td} = \underset{t \in T}{\operatorname{norm}}\big(n_{td} + \alpha_{td}\big).$$

Положительное значение параметра $\alpha_{td}$ или $\beta_{wt}$ соответствует сглаживанию, отрицательное — разреживанию.

**Частичное обучение.** В процессе создания, использования или оценивания тематической модели эксперты, пользователи или асессоры могут отмечать в темах релевантные или нерелевантные термы и документы. Размеченные данные позволяют фиксировать интерпретации тем и повышают устойчивость модели. Разметка может затрагивать лишь часть документов и тем, поэтому её использование относится к задачам *частичного обучения* (semi-supervised learning).

Пусть для каждой темы $t \in T$ заданы четыре подмножества:
$W_t^+$ — «белый список» релевантных термов;
$W_t^-$ — «чёрный список» нерелевантных термов;
$D_t^+$ — «белый список» релевантных документов;
$D_t^-$ — «чёрный список» нерелевантных документов.

*Частичное обучение по релевантности* является частным случаем регуляризатора сглаживания и разреживания при

$$\beta_{wt} = \beta_+[w \in W_t^+] - \beta_-[w \in W_t^-],$$
$$\alpha_{td} = \alpha_+[d \in D_t^+] - \alpha_-[d \in D_t^-],$$

где $\beta_\pm$ и $\alpha_\pm$ — коэффициенты регуляризации.

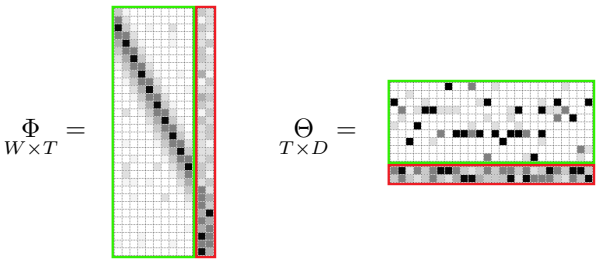


Рис. 4: Структура разреженности матриц Φ и Θ с предметными и фоновыми темами.

**Предметные и фоновые темы.** Чтобы модель была интерпретируемой, каждая тема должна иметь *семантическое ядро* — множество слов, характеризующих определённую предметную область и редко употребляемых в других темах. Для этого матрицы Φ и Θ должны иметь структуру разреженности, аналогичную показанной на рис. 4. Множество тем разбивается на два подмножества, $T = S \sqcup B$.

*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 [30, 98], в которых использовалось только одно фоновое распределение.

**Сфокусированный тематический поиск.** Частичное обучение тем можно рассматривать как разновидность тематического информационного поиска. В качестве запроса задаётся *семантическое ядро* одной или нескольких тем. Это может быть любой фрагмент текста, «белый список» термов (seed words) или *z-метки* — темы, приписанные отдельным словам или фрагментам в документах [15]. Тематическая поисковая система должна не только найти и ранжировать релевантные документы, но и разложить поисковую выдачу по темам. В типичных приложениях релевантный контент составляет ничтожно малую долю коллекции. Тем не менее, именно этот контент должен быть тщательно систематизирован. Образно говоря, требуется «классифицировать иголки в стоге сена» [27]. Темы становятся элементом графического интерфейса пользователя, инструментом навигации и понимания текстовой коллекции. Отсюда важность требования интерпретируемости каждой темы.

Частичное обучение использовалось для поиска и кластеризации новостей [52], поиска в социальных медиа информации, связанной с болезнями, симптомами и методами лечения [92, 93], с преступностью и экстремизмом [70, 109], с национальностями и межнациональными отношениями [27, 56, 91].

В модели ATAM (ailment topic aspects model) в качестве сглаживающего распределения $\beta_{wt}$ использовалась большая коллекция медицинских статей [93].

В моделях SSLDA (semi-supervised LDA) и ISLDA (interval semi-supervised LDA) для поиска этнорелевантных тем использовалось сглаживание по словарю из нескольких сотен этнонимов [27]. В модели SSLDA для каждой этнорелевантной темы задаётся свой словарь этнонимов, связанных с одним определённым этносом.

В модели ISLDA множество тем разбивается на интервалы, и для всех тем каждого интервала задаётся общий словарь этнонимов. Преимущество этих моделей в том, что интерпретация каждой темы известна заранее. Недостатки в том, что трудно предугадывать число тем для каждой этничности и строить полиэтничные темы для выявления межэтнических конфликтов. Альтернативный подход заключается в том, чтобы задать число этно-тем и применить к ним общее сглаживание по словарю этнонимов. Тематическая модель сама определит, как разделить их по этничностям [16, 17]. Недостаток этого подхода в том, что интерпретируемость найденных тем приходится проверять вручную.

**Декоррелирование.** Тематическая модель не должна содержать дублирующихся или похожих тем. Чем различнее темы, тем информативнее модель. Для повышения различности тем будем минимизировать сумму попарных скалярных произведений $\langle \varphi_t, \varphi_s \rangle = \sum_w \varphi_{wt}\varphi_{ws}$ между столбцами матрицы $\Phi$. Получим регуляризатор:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \varphi_{wt}\varphi_{ws}.$$

Формула М-шага, согласно (12), имеет вид

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

Этот регуляризатор контрастирует строки матрицы $\Phi$. В каждой строке, независимо от остальных, вероятности $\varphi_{wt}$ наиболее значимых тем терма $w$ увеличиваются, вероятности остальных тем уменьшаются и могут обращаться в нуль. Разреживание — это сопутствующий эффект декоррелирования. В [118] был замечен ещё один полезный эффект: слова общей лексики группируются в отдельные темы. Эксперименты с комбинированием регуляризаторов сглаживания, разреживания и декоррелирования в ARTM подтверждают это наблюдение [6, 128, 127].

Декоррелирование впервые было предложено в модели TWC-LDA (topic-weak-correlated LDA) в рамках байесовского подхода [118]. Соответствующее априорное распределение не является сопряжённым к мультиномиальному, поэтому байесовский вывод сталкивается с техническими трудностями. В ARTM расчётные формулы выводятся в одну строку.

**Комбинация регуляризаторов** сглаживания фоновых тем, разреживания предметных тем в матрице $\Theta$ и декоррелирования столбцов матрицы $\Phi$ использовалась уже во многих работах для улучшения интерпретируемости тем [6, 127, 128, 129, 12]. Подобрав коэффициенты регуляризации, можно одновременно значительно улучшить разреженность, контрастность, чистоту и когерентность тем при незначительной потере правдоподобия модели [128]. Были выработаны основные рекомендации: декоррелирование и сглаживание включать сразу, разреживание — после 10–20 итераций, когда образуется тенденция к сходимости параметров модели.

Та же комбинация регуляризаторов была использована для тематического разведочного поиска в [12]. Оказалось, что она существенно улучшает качество поиска, хотя никакие критерии качества поиска непосредственно не оптимизировались.

# 6 Определение числа тем

*Регуляризатор отбора тем* предложен в [127] для удаления незначимых тем из тематической модели. Он основан на идее кросс-энтропийного разреживания распределения $p(t)$, которое легко выражается через параметры тематической модели:

$$R(\Theta) = \tau n \sum_{t \in T} \frac{1}{|T|} \ln p(t), \qquad p(t) = \sum_d p(d)\theta_{td}.$$

Подставим этот регуляризатор в формулу М-шага (13):

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

Заменим $\theta_{td}$ в правой части равенства несмещённой оценкой $\frac{n_{td}}{n_d}$:

$$\theta_{td} = \underset{t \in T}{\text{norm}}\Big(n_{td}\Big(1 - \tau \frac{n}{n_t |T|}\Big)\Big).$$

Этот регуляризатор разреживает целиком строки матрицы $\Theta$. Если значение счётчика $n_t$ в знаменателе достаточно мало, то все элементы $t$-й строки оказываются равными нулю, и тема $t$ полностью исключается из модели. При использовании данного регуляризатора сначала устанавливается заведомо избыточное число тем $|T|$. В ходе итераций число нулевых строк матрицы $\Theta$ постепенное увеличивается.

Отбор тем в ARTM намного проще непараметрических байесовских моделей — иерархического процесса Дирихле (hierarchical Dirichlet process, HDP) [119] или процесса китайского ресторана (Chinese restaurant process, CRP) [24].

В обоих подходах, ARTM и HDP, имеется управляющий параметр, выбирая который, можно получать модели с числом тем, различающимся на порядки (в ARTM это коэффициент регуляризации $\tau$, в HDP — гиперпараметр $\gamma$). Поэтому про оба подхода нельзя сказать, что они определяют оптимальное число тем.

В [129] были проведены эксперименты на полусинтетических данных, представляющих собой смесь двух распределений $p(w\,|\,d)$ — реальной коллекции, для которой истинное число тем неизвестно, и синтетической коллекции с заданным числом тем. Оказалось, что HDP и ARTM способны определять истинное число тем на синтетических и полусинтетических данных. При этом ARTM определяет его более точно и устойчиво. Однако чем ближе полусинтетические данные к реальным, тем менее чётко различим момент, когда модель достигает истинного числа тем. На реальных данных он неразличим вовсе, причём для обоих подходов. Отсюда можно сделать вывод, что в реальных текстовых коллекциях никакого «истинного числа тем» просто не существует. Чем больше коллекция, тем более мелкие семантические различия в темах возможно уловить. Эти соображения подтверждаются опытом построения иерархических тематических моделей и рубрикаторов. Темы можно дробить на более мелкие подтемы вплоть до порога статистической значимости. Выбор этого порога также является эвристикой, и от него зависит итоговое число тем.

В ходе экспериментов [129] также выяснилось, что регуляризатор отбора тем имеет полезный сопутствующий эффект: он удаляет из модели дублирующие, расщеплённые и линейно зависимые темы.

По скорости вычислений `BigARTM` с регуляризатором отбора тем оказался в 100 раз быстрее свободно доступной реализации HDP.

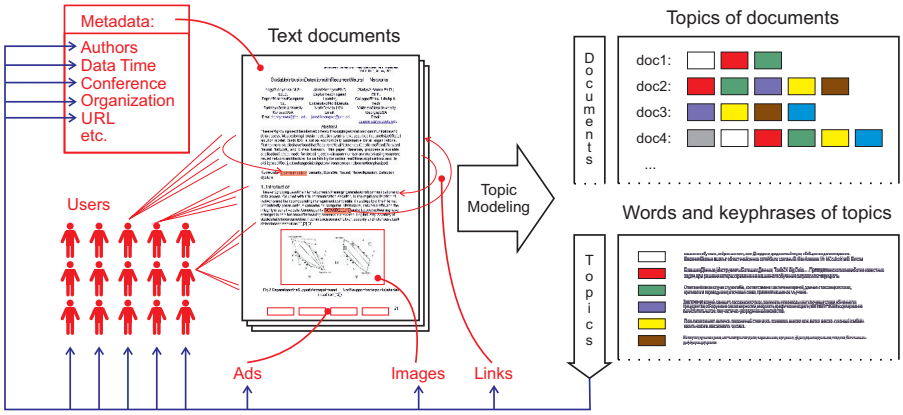



Рис. 5: Обычная тематическая модель определяет распределение тем в каждом документе $p(t\,|\,d)$ и распределение термов в каждой теме $p(w\,|\,t)$. Мультимодальная модель распространяет семантику тем на элементы всех остальных модальностей, в том числе нетекстовые.

# 7 Модальности

*Мультимодальная тематическая модель* описывает документы, содержащие метаданные наряду с основным текстом. Метаданные помогают более точно определять тематику документов, и, наоборот, тематическая модель может использоваться для выявления семантики метаданных или предсказания пропущенных метаданных.

Каждый тип метаданных образует отдельную *модальность* со своим словарём. Слова естественного языка, словосочетания [132, 141], теги [58], именованные сущности [85] — это примеры текстовых модальностей. Для анализа коротких текстов с опечатками используют модальность буквенных $n$-грамм, что позволяет улучшать качество информационного поиска [50]. Примерами нетекстовых модальностей являются (рис. 5): авторы [105], моменты времени [121, 152, 122], классы, жанры или категории [106, 155], цитируемые или цитирующие документы [38] или авторы [55], пользователи электронных библиотек, социальных сетей или рекомендательных систем [62, 113, 134, 148, 149], графические элементы изображений [25, 49, 66], рекламные объявления на веб-страницах [96].

Все перечисленные случаи, несмотря на разнообразие интерпретаций, описываются единым формализмом модальностей в ARTM. Каждый документ рассматривается как универсальный контейнер, содержащий токены различных модальностей, включая обычные слова.

Пусть $M$ — множество модальностей. Каждая модальность имеет свой словарь токенов $W_m$, $m \in M$. Эти множества попарно не пересекаются. Их объединение будем обозначать через $W$. Модальность токена $w \in W$ будем обозначать через $m(w)$.

Тематическая модель модальности $m$ аналогична модели (2):

$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t)\,p(t\,|\,d) = \sum_{t\in T} \varphi_{wt}\theta_{td}, \qquad w \in W_m, \quad d \in D. \tag{22}$$

Каждой модальности $m$ соответствует стохастическая матрица $\Phi_m = \big(\varphi_{wt}\big)_{W_m\times T}$. Совокупность матриц $\Phi_m$, если их записать в столбец, образует $W\times T$-матрицу $\Phi$. Распределение тем в каждом документе является общим для всех модальностей.

Мультимодальная модель строится путём максимизации взвешенной суммы логарифмов правдоподобия модальностей и регуляризаторов. Веса $\tau_m$ позволяют сба-

лансировать модальности по их важности и с учётом их частотности в документах:

$$\sum_{m\in M}\tau_m\sum_{d\in D}\sum_{w\in W_m} n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}; \tag{23}$$

$$\sum_{w\in W_m}\varphi_{wt}=1; \quad \varphi_{wt}\geqslant 0; \qquad \sum_{t\in T}\theta_{td}=1; \quad \theta_{td}\geqslant 0. \tag{24}$$

**Теорема 2.** *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка $(\Phi,\Theta)$ локального экстремума задачи (23)–(24) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}$ для всех невырожденных тем $t$ и документов $d$:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{25}$$

$$\varphi_{wt} = \underset{w\in W_m}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D}\tau_{m(w)}n_{dw}p_{tdw}; \tag{26}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\right); \qquad n_{td} = \sum_{m\in M}\sum_{w\in W^m}\tau_m n_{dw}p_{tdw}. \tag{27}$$

Теорема 1 является частным случаем теоремы 2 в случае, когда модальность только одна, $|M|=1$ и $\tau_m=1$. Таким образом, переход от одной модальности к произвольному числу модальностей сводится к двум поправкам:

1) матрица $\Phi$ разбивается на блоки $\Phi_m$, которые нормируются по-отдельности;

2) исходные данные $n_{dw}$ домножаются на веса модальностей $\tau_{m(w)}$.

В проекте `BigARTM` реализована возможность комбинировать любое число модальностей с любыми регуляризаторами [16].

**Модальность языков.** Мультиязычные текстовые коллекции используются для кросс-язычного информационного поиска, когда по запросу на одном языке требуется найти семантически близкие документы на другом языке. Для связывания языков используются параллельные тексты или двуязычные словари. Первые мультиязычные тематические модели появились почти одновременно [36, 80, 90] и представляли собой мультимодальную модель, в которой модальностями являются языки, и каждая связка параллельных текстов объединяется в один документ. Оказалось, что связывания документов достаточно для синхронизации тем в двух языках и кросс-язычного поиска. Попытки более точного и трудоёмкого выравнивания по предложениям или по словам практически не улучшают качество поиска. Обстоятельный обзор мультиязычных тематических моделей можно найти в [130].

Для использования двуязычного словаря в [7] был предложен регуляризатор сглаживания. Он формализует предположение, что если слово $u$ в языке $k$ является переводом слова $w$ из языка $\ell$, то тематики этих слов $p(t\,|\,u)$ и $p(t\,|\,w)$ должны быть близки в смысле KL-дивергенции:

$$R(\Phi) = \sum_{w,u}\sum_{t\in T} n_{ut}\ln\varphi_{wt}.$$

Согласно формуле M-шага, вероятность слова в теме увеличивается, если оно имеет переводы, имеющие высокую вероятность в данной теме:

$$\varphi_{wt} = \underset{w\in W^\ell}{\operatorname{norm}}\Big(n_{wt} + \tau\sum_{u} n_{ut}\Big).$$

Этот регуляризатор не учитывал, что перевод слова может зависеть от темы, и что среди переводов слова могут находиться переводы его омонимов. Поэтому в той же работе был предложен второй регуляризатор, который вводил в модель новые параметры $\pi_{uwt} = p(u\,|\,w,t)$ — вероятности того, что слово $u$ является переводом слова $w$ в теме $t$. Предполагается, что тема $t$, как распределение $\hat{p}(u\,|\,t) = \frac{n_{ut}}{n_t}$ над словами языка $k$, должна быть близка в смысле KL-дивергенции к вероятностной модели той же темы $p(u\,|\,t) = \sum_w \pi_{uwt}\varphi_{wt}$, построенной по переводам слов из языка $\ell$:

$$R(\Phi,\Pi) = \tau \sum_{u\in W^k}\sum_{t\in T} n_{ut} \ln \sum_{w\in W^\ell} \pi_{uwt}\varphi_{wt} \;\to\; \max_{\Phi,\Pi}.$$

Формула M-шага теперь учитывает вероятности переводов $\pi_{uwt}$. Кроме того, добавляется рекуррентная формула для оценивания этих вероятностей:

$$\varphi_{wt} = \underset{w\in W^\ell}{\operatorname{norm}}\Big(n_{wt} + \tau \sum_u \pi_{wut} n_{ut}\Big);$$
$$\pi_{uwt} = \underset{u\in W^k}{\operatorname{norm}}\big(\pi_{wut} n_{ut}\big).$$

Эксперименты показали, что связывание параллельных текстов сильнее улучшает качество поиска, чем оба способа учёта словарей. Второй способ немного лучше первого. Кроме того, он позволяет выбирать варианты перевода в зависимости от контекста, что может быть полезно для статистического машинного перевода.

**Модальности категорий и авторов.** Допустим, что распределения тем в документах $p(t\,|\,d)$ порождаются одной из модальностей, например, авторами, рубриками или категориями. Будем считать, что с каждым термом $w$ в каждом документе $d$ связана не только тема $t \in T$, но и категория $c$ из заданного множества категорий $C$. Расширим вероятностное пространство до множества $D \times W \times T \times C$. Пусть известно подмножество категорий $C_d \subseteq C$, к которым может относиться документ $d$.

Рассмотрим мультимодальную тематическую модель (22), в которой распределение вероятности тем документов $\theta_{td} = p(t\,|\,d)$ описывается смесью распределений тем категорий $\psi_{tc} = p(t\,|\,c)$ и категорий документов $\pi_{cd} = p(c\,|\,d)$:

$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t) \sum_{c\in C_d} p(t\,|\,c)p(c\,|\,d) = \sum_{t\in T}\sum_{c\in C_d} \varphi_{wt}\psi_{tc}\pi_{cd}. \tag{28}$$

Это также задача стохастического матричного разложения, только теперь требуется найти три матрицы: $\Phi$ — матрица термов тем, $\Psi = (\psi_{tc})_{T\times C}$ — *матрица тем категорий*, $\Pi = (\pi_{cd})_{C\times D}$ — *матрица категорий документов*.

Модель основана на двух гипотезах условной независимости:

$p(t\,|\,c,d) = p(t\,|\,c)$ — тематика документа $d$ зависит не от самого документа, а только от того, каким категориям он принадлежит;

$p(w\,|\,t,c,d) = p(w\,|\,t)$ — распределение термов полностью определяется тематикой документа и не зависит от самого документа и его категорий.

Кроме того, предполагается, что $\pi_{cd} = p(c\,|\,d) = 0$ для всех $c \notin C_d$.

Задача максимизации регуляризованного правдоподобия:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T}\sum_{c\in C_d} \varphi_{wt}\psi_{tc}\pi_{cd} + R(\Phi,\Psi,\Pi) \;\to\; \max_{\Phi,\Psi,\Pi}; \tag{29}$$

$$\sum_{w\in W} \varphi_{wt} = 1,\ \varphi_{wt} \geqslant 0; \qquad \sum_{t\in T} \psi_{tc} = 1,\ \psi_{tc} \geqslant 0; \qquad \sum_{c\in C_d} \pi_{cd} = 1,\ \pi_{cd} \geqslant 0. \tag{30}$$

**Теорема 3.** *Пусть функция $R(\Phi,\Psi,\Pi)$ непрерывно дифференцируема. Точка $(\Phi,\Psi,\Pi)$ локального экстремума задачи (29), (30) удовлетворяет системе уравнений со вспомогательными переменными $p_{tcdw} = p(t,c\,|\,d,w)$:*

$$p_{tcdw} = \underset{(t,c)\in T\times C_d}{\operatorname{norm}} \varphi_{wt}\psi_{tc}\pi_{cd};$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D}\sum_{c\in C_d} n_{dw}p_{tcdw};$$

$$\psi_{tc} = \underset{t\in T}{\operatorname{norm}}\left(n_{tc} + \psi_{tc}\frac{\partial R}{\partial \psi_{tc}}\right); \qquad n_{tc} = \sum_{d\in D}\sum_{w\in d} n_{dw}p_{tcdw};$$

$$\pi_{cd} = \underset{c\in C_d}{\operatorname{norm}}\left(n_{cd} + \pi_{cd}\frac{\partial R}{\partial \pi_{cd}}\right); \qquad n_{cd} = \sum_{w\in d}\sum_{t\in T} n_{dw}p_{tcdw}.$$

Данная модель, основанная на трёхматричном разложении, наиболее известна как *автор-тематическая модель* ATM (author-topic model), в которой порождающей модальностью являются авторы документов [105]. В *тематической модели тегирования документов* TWTM (tag weighted topic model) порождающей модальностью являются теги документа [64]. Аналогичная модель использовалась для обработки видеопотоков в [49]. Документы $d$ соответствовали последовательным 1-секундным видеоклипам, термы $w$ — элементарным визуальным событиям, темы $t$ — простым действиям, состоящим из сочетания событий, категории $c$ — более сложным поведениям, состоящим из сочетания действий, причём ставилась задача выделить в каждом клипе одно основное поведение.

Модель (28) можно упростить и свести снова к двухматричному разложению, если отождествить темы с категориями, $C \equiv T$, и взять единичную матрицу $\Psi$. Данная модель известна в литературе как Flat-LDA [106] и Labeled-LDA [102]. Её выразительные возможности беднее, чем у PLSA и LDA, так как значительная доля элементов матрицы $\Pi \equiv \Theta$ фиксированы и равны нулю.

Трёхматричные разложения пока не реализованы в библиотеке `BigARTM`.

**Темпоральные модели.** Время создания документов важно при анализе новостных потоков, научных публикаций, патентных баз, данных социальных сетей. Тематические модели, учитывающие время, называются *темпоральными*. Они позволяют выделять событийные и перманентные темы, детектировать новые темы, прослеживать развитие тем во времени, выделять тренды.

Пусть $I$ — конечное множество интервалов времени, и каждый документ относится к одному или нескольким интервалам, $D_i$ — подмножество документов, относящихся к интервалу $i$. Будем полагать, что темы как распределения $p(w\,|\,t)$ не меняются во времени. Требуется найти распределение каждой темы во времени $p(i\,|\,t)$.

Тривиальный подход заключается в том, чтобы построить тематическую модель без учёта времени, затем найти распределение тем в каждом интервале $p(t\,|\,i)$ как среднее $\theta_{td}$ по всем документам $d \in D_i$ и перенормировать условные вероятности: $p(i\,|\,t) = p(t\,|\,i)\frac{p(i)}{p(t)}$. Недостаток данного подхода в том, что информация о времени никак не используется при обучении модели и не влияет на формирование тем.

В ARTM эта проблема решается введением модальности времени $I$. Искомое распределение $p(i\,|\,t) = \varphi_{it}$ получается в столбце матрицы Φ. Дополнительные ограничения на поведение тем во времени можно вводить с помощью регуляризации.

В одной из первых темпоральных тематических моделей TOT (topics over time) [140] каждая тема моделировалась параметрическим $\beta$-распределением во времени. Это семейство монотонных и унимодальных непрерывных функций, с помощью которого можно описывать узкие пики событийных тем и ограниченный набор трендов. Темы, имеющие спорадические всплески, данная модель описывает плохо.

Непараметрические темпоральные модели способны описывать произвольные изменения тем во времени. Рассмотрим два естественных предположения и формализуем их с помощью регуляризации.

Во-первых, предположим, что многие темы являются событийными и имеют относительно небольшое «время жизни», поэтому в каждом интервале времени $i$ присутствуют не все темы. Потребуем разреженности распределений $p(t\,|\,i)$ с помощью кросс-энтропийного регуляризатора:

$$R_1(\Phi \text{ или } \Theta) = -\tau_1 \sum_{i\in I} \sum_{t\in T} \ln p(t\,|\,i).$$

Во-вторых, предположим, что распределения $p(i\,|\,t)$ как функции времени меняются не слишком быстро и введём регуляризатор сглаживания:

$$R_2(\Phi \text{ или } \Theta) = -\tau_2 \sum_{i\in I} \sum_{t\in T} \big|p(i\,|\,t) - p(i{-}1\,|\,t)\big|.$$

Оба регуляризатора можно записать и как функцию от Φ, и как функцию от Θ. В случае регуляризатора $R_2(\Phi)$ формула М-шага имеет вид[2]

$$\varphi_{it} = \underset{i\in I}{\text{norm}}\big(n_{it} + \tau_2\varphi_{it}\,\text{sign}(\varphi_{i-1,t} - \varphi_{it}) + \tau_2\varphi_{it}\,\text{sign}(\varphi_{i+1,t} - \varphi_{it})\big),$$

где функция sign возвращает $+1$ для положительного аргумента и $-1$ для отрицательного. Регуляризатор сглаживает значения в каждой точке временного ряда $p(i\,|\,t)$ по отношению к соседним точкам слева и справа.

# 8 Зависимости

**Классификация.** *Тематическая модель классификации* Dependency LDA [106] является байесовским аналогом модели (22) с модальностями термов $W$ и классов $C$.

[2]Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей. Бакалаврская диссертация, ВМК МГУ, 2015.
http://www.MachineLearning.ru/wiki/images/9/9f/2015_417_DoykovNV.pdf

Имеется обучающая выборка документов $d$, для каждого из которых известно подмножество классов $C_d \subset C$. Требуется классифицировать новые документы с неизвестным $C_d$. Для этого будем использовать *линейную вероятностную модель классификации*, в которой объектами являются документы $d$, признаки соответствуют темам $t$ и принимают значения $\theta_{td} = p(t\,|\,d)$:

$$\hat{C}_d = \Big\{ c \in C \;\Big|\; p(c\,|\,d) = \sum_{t \in T} \varphi_{ct}\theta_{td} \geqslant \gamma_c \Big\}.$$

Коэффициенты линейной модели $\varphi_{ct} = p(c\,|\,t)$ и пороги $\gamma_c$ обучаются по выборке документов с известными $C_d$. Признаковое описание нового документа $\theta_d$ вычисляется тематической моделью только по его термам.

Эксперименты в [106] показали, что тематические модели превосходят обычные методы многоклассовой классификации на больших текстовых коллекциях с большим числом несбалансированных, пересекающихся, взаимозависимых классов. В [125] те же выводы на тех же коллекциях были воспроизведены для мультимодальной ARTM. *Несбалансированность* означает, что классы могут содержать как малое, так и очень большое число документов. В случае *пересекающихся* классов документ может относиться как к одному классу, так и к большому числу классов. *Взаимозависимые* классы имеют общие термы и темы, поэтому при классификации документа могут вступать в конкуренцию.

В некоторых задачах классификации имеется информация о том, что документ $d$ из обучающей выборки не принадлежит подмножеству классов $C'_d \subset C$. Для этого случая запишем правдоподобие вероятностной модели бинарных данных:

$$L(\Phi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C_d} \ln \sum_{t \in T} \varphi_{ct}\theta_{td} + \tau \sum_{d \in D} \sum_{c \in C'_d} \ln\Big(1 - \sum_{t \in T} \varphi_{ct}\theta_{td}\Big) \;\to\; \max.$$

Первое слагаемое равняется log-правдоподобию модальности классов (22), если положить $n_{dc} = [c \in C_d]$. Второе слагаемое можно рассматривать как регуляризатор не-принадлежности документов классам.

**Регрессия.** Задачи предсказания числовой величины как функции от текста возникают во многих приложениях электронной коммерции: предсказание рейтинга товара, фильма или книги по тексту отзыва; предсказание числа кликов по тексту рекламного объявления; предсказание зарплаты по описанию вакансии; предсказание полезности (числа лайков) отзыва на отель, ресторан, сервис. Для восстановления числовых функций по конечной обучающей выборке пар «объект–ответ» используются регрессионные модели, однако все они принимают на входе векторные описания объектов. Тематическая модель позволяет заменить текст документа $d$ его векторным представлением $\theta_d$. С другой стороны, критерий оптимизации регрессионной модели можно использовать в качестве регуляризатора, чтобы найти темы, наиболее информативные с точки зрения точности предсказаний [74, 115].

Пусть для каждого документа $d$ обучающей выборки $D$ задано целевое значение $y_d \in \mathbb{R}$. Рассмотрим *линейную модель регрессии*, которая предсказывает математическое ожидание целевой величины:

$$\mathsf{E}(y\,|\,d) = \sum_{t \in T} v_t\theta_{td},$$

где $v \in \mathbb{R}^T$ — вектор коэффициентов. Применим метод наименьших квадратов для обучения вектора $v$ по выборке документов:

$$R(\Theta, v) = -\tau \sum_{d \in D} \Big( y_d - \sum_{t \in T} v_t \theta_{td} \Big)^2.$$

Подставляя этот регуляризатор в (13) и приравнивая нулю его производную по $v$, получим формулы M-шага:

$$\theta_{td} = \operatorname*{norm}_t \Big( n_{td} + \tau v_t \theta_{td} \Big( y_d - \sum_{s \in T} v_s \theta_{sd} \Big) \Big);$$
$$v = (\Theta \Theta^T)^{-1} \Theta y.$$

Заметим, что формула для вектора $v$ является стандартным решением задачи наименьших квадратов при фиксированной матрице $\Theta$. Вектор $v$ можно обновлять по окончании каждого прохода коллекции, либо после обработки каждого пакета документов в онлайновом EM-алгоритме.

В [115] показано, что качество восстановления регрессии на текстах может существенно зависеть от инициализации тематической модели, там же предложено несколько стратегий инициализации.

**Корреляции тем.** *Модель коррелированных тем* CTM (correlated topic model) предназначена для выявления связей между темами [21]. Например, статья по геологии более вероятно связана с археологией, чем с генетикой. Знание о том, какие темы чаще совместно встречаются в документах коллекции, позволяет точнее моделировать тематику отдельных документов в мультидисциплинарных коллекциях.

Для описания корреляций удобно использовать многомерное нормальное распределение. Оно не подходит для описания неотрицательных нормированных вектор-столбцов $\theta_d$, но неплохо описывает векторы их логарифмов $\eta_{td} = \ln \theta_{td}$. Поэтому в модель вводится многомерное лог-нормальное распределение (logistic normal) с двумя параметрами: вектором математического ожидания $\mu$ и ковариационной матрицей $\Sigma$:

$$p(\eta_d \,|\, \mu, \Sigma) = \frac{\exp\big(-\frac{1}{2}(\eta_d - \mu)^\intercal \Sigma^{-1} (\eta_d - \mu)\big)}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}}.$$

Изначально модель CTM была разработана в рамках байесовского подхода, где возникали дополнительные технические трудности из-за того, что лог-нормальное распределение не является сопряжённым к мультиномиальному. В рамках ARTM идея CTM формализуется и реализуется намного проще.

Определим регуляризатор как логарифм правдоподобия лог-нормальной модели для выборки векторов документов $\eta_d$:

$$R(\Theta, \mu, \Sigma) = \tau \sum_{d \in D} \ln p(\eta_d \,|\, \mu, \Sigma) = -\frac{\tau}{2} \sum_{d \in D} (\ln \theta_d - \mu)^\intercal \Sigma^{-1} (\ln \theta_d - \mu) + \text{const} \to \max_{\Theta, \mu, \Sigma}.$$

Согласно (13), формула M-шага для $\theta_{td}$ принимает вид

$$\theta_{td} = \operatorname*{norm}_{t \in T} \Big( n_{td} - \tau \sum_{s \in T} \Sigma^{-1}_{ts} \big( \ln \theta_{sd} - \mu_s \big) \Big), \tag{31}$$

где $\Sigma_{ts}^{-1}$ — элементы обратной ковариационной матрицы. Параметры $\Sigma, \mu$ нормального распределения обновляются после каждого прохода коллекции, либо после каждого пакета документов в онлайновом EM-алгоритме:

$$\mu = \frac{1}{|D|}\sum_{d\in D} \ln \theta_d;$$

$$\Sigma = \frac{1}{|D|}\sum_{d\in D} \big(\ln \theta_d - \mu\big)\big(\ln \theta_d - \mu\big)^{\intercal}.$$

Таким образом, трудоёмкая операция обращения ковариационной матрицы выполняется относительно редко. В [21] использовалась LASSO-регрессия, чтобы получать разреженную ковариационную матрицу.

# 9 Связи между документами

**Ссылки и цитирование.** Иногда имеется дополнительная информация о связях между документами и предполагается, что связанные документы имеют схожую тематику. Связь может означать, что два документа относятся к одной рубрике, совместно упоминаются или ссылаются друг на друга. Формализуем это предположение с помощью регуляризатора:

$$R(\Theta) = \tau \sum_{d,c} n_{dc} \sum_{t\in T} \theta_{td}\theta_{tc},$$

где $n_{dc}$ — вес связи между документами, например, число ссылок из $d$ на $c$. В [38] предложена похожая модель LDA-JS, в которой вместо максимизации ковариации минимизируется дивергенция Йенсена-Шеннона между распределениями $\theta_d$ и $\theta_c$. Формула M-шага для $\theta_{td}$, согласно (13), имеет вид

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(n_{td} + \tau\theta_{td}\sum_{c\in D} n_{dc}\theta_{tc}\Big).$$

Это ещё одна разновидность сглаживания. Вероятности $\theta_{td}$ в ходе итераций приближаются к вероятностям $\theta_{tc}$ документов, связанных с $d$.

Регуляризатор матрицы $\Theta$ становится неэффективным при пакетной обработке больших коллекций, когда документы $c$, на которые ссылается данный документ $d$, находятся в других пакетах. Проблема решается введением модальности документов, на которые есть ссылки из других документов. Этот способ порождает новую проблему: если мощность этой модальности окажется равной числу документов, то матрица $\Phi$ может не поместиться в оперативную память. Можно сократить эту модальность, оставив только наиболее влиятельные документы $c$, число ссылок на которые $n_c = \sum_d n_{dc}$ превышает выбранный порог.

Данная идея пришла из модели влияния научных публикаций LDA-post [38]. В ней используются две модальности: слова $W_1$ и цитируемые документы $W_2 \subseteq D$. Модель выявляет наиболее влиятельные документы внутри каждой темы. Ненулевые элементы в строке $c$ матрицы $\Phi_2$ показывают, на какие темы повлиял документ $c \in W_2$. Также модель позволяет различать, какие из ссылок существенно повлияли на научную статью, а какие являются второстепенными, чисто формальными или «данью вежливости». Считается, что документ $c$ повлиял на документ $d$, если $d$ ссылается на $c$ и они имеют значительную долю общей тематики.

**Геолокации.** Информация о географическом положении используется при анализе данных социальных сетей. Географическая привязка документа $d$ или его автора задаётся либо модальностями *геотегов* (названиями страны, региона, населённого пункта), либо *геолокацией* — парой географических координат $\ell_d = (x_d, y_d)$. В первом случае можно использовать обычную мультимодальную модель, во втором случае нужен дополнительный регуляризатор. ARTM позволяет совмещать в модели оба типа географических данных.

Целью моделирования может быть выделение региональных тем, определение «ареала обитания» каждой темы, поиск похожих тем в других регионах. Например, в качестве одной из иллюстраций в [150] определяются регионы популярности национальной кухни по постам пользователей Flickr. Другая иллюстрация из [75] показывает, что тематическая модель, учитывающая, из какого штата США пришло сообщение, точнее прослеживает путь урагана «Катрина».

Квадратичный регуляризатор матрицы $\Theta$, предложенный в [150], формализует предположение, что документы со схожими геолокациями имеют схожую тематику:

$$R(\Theta) = -\frac{\tau}{2} \sum_{(c,d)} w_{cd} \sum_{t \in T} \big(\theta_{td} - \theta_{tc}\big)^2,$$

где $w_{cd}$ — вес пары документов $(c, d)$, выражающий близость геолокаций. Например, $w_{cd} = \exp(-\gamma r_{cd}^2)$, где $r_{cd}^2 = (x_c - x_d)^2 + (y_c - y_d)^2$ — квадрат евклидова расстояния.

Этот регуляризатор требует при обработке каждого документа $d$ доступа к векторам $\theta_c$ других документов, что затрудняет пакетную обработку больших коллекций. Альтернативный способ сглаживания тематики географически близких сообщений основан на регуляризации матрицы $\Phi$.

Пусть $G$ — модальность геотегов, $\varphi_{gt} = p(g\,|\,t)$. Тематика геотега $g$ выражается по формуле Байеса: $p(t\,|\,g) = \varphi_{gt} \frac{n_t}{n_g}$, где $n_g$ — частота геотега $g$ в исходных данных, $n_t = \sum_g n_{gt}$ — частота темы $t$ в модальности геотегов, вычисляемая EM-алгоритмом.

Квадратичный регуляризатор матрицы $\Phi$ по модальности геотегов формализует предположение, что географически близкие геотеги имеют схожую тематику:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g' \in G} w_{gg'} \sum_{t \in T} n_t^2 \Big(\frac{\varphi_{gt}}{n_g} - \frac{\varphi_{g't}}{n_{g'}}\Big)^2,$$

где $w_{gg'}$ — вес пары геотегов $(g, g')$, выражающий их географическую близость. Ниже мы рассмотрим обобщение этого регуляризатора на более широкий класс задач.

**Графы и социальные сети.** В [75] предложена более общая тематическая модель NetPLSA, учитывающая произвольные графовые (сетевые) структуры на множестве документов. Пусть задан граф $\langle V, E\rangle$ с множеством вершин $V$ и множеством рёбер $E$. Каждой его вершине $v \in V$ соответствует подмножество документов $D_v \subset D$. Например, в роли $D_v$ может выступать отдельный документ, все статьи одного автора $v$, все посты из одного географического региона $v$, и т. д.

Тематика каждой вершины $v \in V$ выражается через параметры модели $\Theta$:

$$p(t\,|\,v) = \sum_{d \in D_v} p(t\,|\,d)\, p(d\,|\,v) = \frac{1}{|D_v|} \sum_{d \in D_v} \theta_{td}.$$

В модели NetPLSA используется квадратичный регуляризатор:

$$R(\Theta) = -\frac{\tau}{2} \sum_{(u,v)\in E} w_{uv} \sum_{t\in T} \big(p(t\,|\,v) - p(t\,|\,u)\big)^2,$$

где веса $w_{uv}$ рёбер графа $(u, v)$ задаются естественным образом, когда в задаче есть соответствующая дополнительная информация. Например, если $D_v$ — все статьи автора $v$, то в качестве веса ребра $w_{uv}$ естественно взять число статей, написанных авторами $u$ и $v$ в соавторстве. Если подобной информации нет, то вес полагается равным единице.

Этот регуляризатор требует при обработке каждого документа $d$ доступа к векторам $\theta_c$ других документов, что затрудняет эффективную пакетную обработку больших коллекций. Альтернативный путь состоит в том, чтобы множество вершин графа $V$ объявить модальностью и перейти к регуляризации матрицы Φ.

В каждый документ $d \in D_v$ добавим токен $v$ модальности $V$. Выразим тематику вершины $v$ через параметры Φ по формуле Байеса: $p(t\,|\,v) = p(v\,|\,t)\frac{p(t)}{p(v)} = \varphi_{vt}\frac{n_t}{|D_v|}$, где $n_t = \sum_v n_{vt}$ — частота темы $t$ в модальности $V$, вычисляемая EM-алгоритмом.

Регуляризатор NetPLSA сохраняет прежний вид, но становится функцией от Φ:

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v)\in E} w_{uv} \sum_{t\in T} n_t^2 \Big(\frac{\varphi_{vt}}{|D_v|} - \frac{\varphi_{ut}}{|D_u|}\Big)^2.$$

Во многих приложениях важны направленности связей, которые квадратичный регуляризатор не учитывает. Например, связь $(u, v)$ может означать ссылку из документа $u$ на документ $v$. В модели iTopicModel [117] предполагается, что если $(u, v) \in E$, то тематика $p(t\,|\,u)$ шире тематики $p(t\,|\,v)$. Поэтому минимизируется сумма дивергенций $\mathrm{KL}\big(p(t\,|\,v)\,\|\,p(t\,|\,u)\big)$, причём $p(t\,|\,v)$ можно выразить как через Θ, так и через Φ:

$$R(\Theta \text{ или } \Phi) = \frac{\tau}{2} \sum_{(u,v)\in E} w_{uv} \sum_{t\in T} p(t\,|\,v) \ln p(t\,|\,u).$$

Как показали эксперименты[3], регуляризация матрицы Φ приводит практически к тем же результатам, что и регуляризация Θ для моделей NetPLSA и iTopicModels.

# 10 Иерархии тем

Иерархические тематические модели рекурсивно делят темы на подтемы. Тематические иерархии служат для построения рубрикаторов, систематизации больших объёмов текстовой информации, информационного поиска и навигации по большим мультидисциплинарным коллекциям. Задача автоматической рубрикации текстов сложна своей неоднозначностью и субъективностью. Различия во мнениях экспертов относительно рубрикации документов могут достигать 40% [1]. Несмотря на обилие работ

[3]Виктор Булатов. Использование графовой структуры в тематическом моделировании. Магистерская диссертация, ФИВТ МФТИ, 2016.
http://www.MachineLearning.ru/wiki/images/4/4d/Bulatov-2016-ms.pdf

по иерархическим тематическим моделям [23, 65, 79, 151, 100, 135, 136, 137, 138], оптимизация размера и структуры иерархии остаётся открытой проблемой; более того, оценивание качества иерархий — также открытая проблема [151].

Стратегии построения тематических иерархий весьма разнообразны: нисходящие (дивизимные) и восходящие (агломеративные), представляющие иерархию деревом или многодольным графом, наращивающие граф по уровням или по вершинам, основанные на кластеризации документов или термов. Нельзя назвать какую-то из стратегий предпочтительной; у каждой есть свои достоинства и недостатки.

В [33] предложена нисходящая стратегия на основе ARTM. Иерархия представляется многодольным графом с увеличивающимся числом тем на каждом уровне. Модель строится по уровням сверху вниз. Число уровней и число тем каждого уровня задаётся вручную. Каждый уровень представляет собой обычную «плоскую» тематическую модель, поэтому время построения модели остаётся линейным по объёму коллекции.

Для моделирования связей между уровнями в модель вводятся параметры $\psi_{st} = p(s\,|\,t)$ — условные вероятности подтем в темах. В случае мультидисциплинарных коллекций подтемам разрешается иметь по несколько родительских тем. ARTM позволяет управлять разреженностью этого распределения с помощью дополнительного кросс-энтропийного регуляризатора. Можно усиливать разреженность распределений $p(t\,|\,s) = \psi_{st}\frac{n_t}{n_s}$ вплоть до вырожденности, тогда каждая подтема будет иметь ровно одну родительскую тему, а вся иерархия будет иметь вид дерева.

**Регуляризатор подтем.** На верхнем уровне иерархии строится обычная плоская тематическая модель. Пусть модель $\ell$-го уровня с множеством тем $T$ уже построена, и требуется построить модель уровня $\ell+1$ с множеством дочерних тем $S$ (subtopics) и бóльшим числом тем, $|S| > |T|$. Потребуем, чтобы родительские темы $t$ хорошо приближались вероятностными смесями дочерних тем $s$:

$$\sum_{t\in T} n_t\ \mathrm{KL}_w\Big(p(w\,|\,t)\ \Big\|\ \sum_{s\in S} p(w\,|\,s)\,p(s\,|\,t)\Big) = \sum_{t\in T} n_t\ \mathrm{KL}_w\Big(\tfrac{n_{wt}}{n_t}\ \Big\|\ \sum_{s\in S}\varphi_{ws}\psi_{st}\Big)\ \to\ \min_{\Phi,\Psi},$$

где $\Psi = (\psi_{st})_{S\times T}$ — *матрица связей*, которая становится дополнительной матрицей параметров для тематической модели дочернего уровня.

Это задача матричного разложения $\Phi^\ell = \Phi\Psi$ для матрицы $\Phi^\ell$ родительского уровня. Обычно мы используем низкоранговые разложения, приближая матрицу высокого ранга произведением матриц более низкого ранга. Однако в данном случае всё наоборот: предполагается, что матрицы $\Phi$ и $\Psi$ имеют полный ранг $|S|$, заведомо превышающий $\operatorname{rank}\Phi^\ell = |T|$. Среди матричных разложений обязательно имеются точные решения, но они нам не подходят. Матрице $\Phi$ выгодно иметь полный ранг, чтобы описывать коллекцию точнее, чем это делает матрица $\Phi^\ell$. Требование, чтобы она заодно приближала матрицу $\Phi^\ell$, вводится через регуляризатор:

$$R(\Phi,\Psi) = \tau\sum_{t\in T}\sum_{w\in W} n_{wt}\ln\sum_{s\in S}\varphi_{ws}\psi_{st}.$$

Задача максимизации $R(\Phi,\Psi)$ с точностью до обозначений совпадает с основной задачей тематического моделирования (8), если считать родительские темы $t$ *псевдодокументами* с частотами слов $\tau n_{wt} = \tau n_t\varphi_{wt}$. Это означает, что вместо добавления

слагаемого в формулы М-шага данный регуляризатор можно реализовать ещё проще. Построив родительский уровень, надо добавить в коллекцию ровно $|T|$ псевдодокументов, задав им в качестве частот термов значения $\tau n_{wt}$. Матрица $\Psi$ получится в столбцах матрицы $\Theta$, соответствующих псевдодокументам.

В библиотеке `BigARTM` этот подход реализован в виде отдельного класса `hARTM`.

# 11 Совстречаемость слов

Гипотеза «мешка слов» является одним из самых критикуемых постулатов тематического моделирования. Поэтому многие исследования направлены на создание более адекватных моделей, учитывающих порядок слов. Из них наиболее важными представляются три направления.

Первое направление связано с выделением *коллокаций* — статистически устойчивых *n-грамм* (последовательностей подряд идущих $n$ слов). Темы, построенные на $n$-граммах, намного лучше интерпретируются, чем построенные на униграммах (отдельных словах). Проблема в том, что число *n-грамм* катастрофически быстро растёт с ростом объёма коллекции.

Второе направление связано с анализом совместной встречаемости слов. Появление программы `word2vec` [76] стимулировало развитие *векторных представлений слов* (word embedding). Они находят массу применений благодаря свойству *дистрибутивности* — семантически близким словам соответствуют близкие векторы. Тематические модели способны строить векторные представления слов, обладающие свойствами интерпретируемости, разреженности и дистрибутивности.

Третье направление связано с *тематической сегментацией* и гипотезой, что текст на естественном языке состоит из последовательности монотематичных сообщений. В частности, каждое предложение чаще всего относится только к одной теме. Задачи сегментации рассматриваются в разделе 12.

**Коллокации.** Использование словосочетаний заметно улучшает интерпретируемость тем, что демонстрируется практически в каждой публикации по $n$-граммным тематическим моделям, см. например [53]. Первая биграммная тематическая модель BTM (bigram topic model) [132] представляла собой по сути мультимодальную модель, в которой каждому слову $v$ соответствовала отдельная модальность со словарём $W_v \subseteq W$, составленным из всех слов, встречающихся непосредственно после слова $v$. Запишем log-правдоподобие этой модели в виде регуляризатора:

$$R(\Phi, \Theta) = \sum_{d \in D} \sum_{v \in d} \sum_{w \in W_v} n_{dvw} \ln \sum_{t \in T} \varphi^v_{wt} \theta_{td},$$

где $\varphi^v_{wt} = p(w|v,t)$ — условная вероятность слов $w$ после слова $v$ в теме $t$; $n_{dvw}$ — частота биграммы «$vw$» в документе $d$. Главный недостаток модели BTM в том, что она учитывает только биграммы. Вторая проблема в том, что число всех биграмм быстро увеличивается с ростом коллекции, и использовать модель BTM на больших коллекциях затруднительно.

Модель TNG (topical $n$-grams) [141] устраняет эти недостатки. Условное распределение слов описывается вероятностной смесью $p(w|v,t) = \xi_{vwt}\varphi^v_{wt} + (1-\xi_{vwt})\varphi_{wt}$, где

$\xi_{vwt}$ — переменная, равная вероятности того, что пара слов «$vw$» является биграммой в теме $t$. При некоторых не особо жёстких предположениях log-правдоподобие этой модели оценивается снизу взвешенной суммой log-правдоподобий модальностей униграмм и биграмм в модели ARTM. Другими словами, мультимодальная ARTM может быть использована для поиска приближённого решения в модели TNG.

В ARTM $n$-граммная модель естественным образом определяется как мультимодальная, в которой для каждого $n$ выделяется отдельная модальность. Для предварительного сокращения словарей $n$-грамм подходит метод поиска коллокаций TopMine [40]. Он линейно масштабируется на большие коллекции и позволяет формировать словарь, в котором каждая $n$-грамма обладает тремя свойствами: (а) имеет высокую частоту в коллекции; (б) состоит из слов, неслучайно часто образующих $n$-грамму; (в) не содержится ни в какой $(n{+}1)$-грамме, обладающей свойствами (а) и (б). В последующих работах были предложены методы SegPhrase [67] и AutoPhrase [108], демонстрирующие ещё лучшие результаты.

**Битермы.** *Короткими текстами* (short text) называют документы, длина которых не достаточна для надёжного определения их тематики. Примерами коротких текстов являются сообщения Твиттера, заголовки новостных сообщений, рекламные объявления, реплики в записях диалогов контакт-центра, и т. д. Известны простые подходы к проблеме, но они не всегда применимы: объединять сообщения по какому-либо признаку (автору, времени, региону и т. д.); считать каждое сообщение отдельным документом, разреживая $p(t\,|\,d)$ вплоть до единственной темы; дополнять коллекцию длинными текстами (например, статьями Википедии). Одним из наиболее успешных и универсальных подходов к проблеме коротких текстов считается *тематическая модель битермов* (biterm topic model, BTM) [144].

*Битермом* называется пара слов, встречающихся рядом — в одном коротком сообщении или в одном предложении или в окне $\pm h$ слов. В отличие от биграммы, между двумя словами битерма могут находиться другие слова. Конкретизация понятия «рядом» зависит от постановки задачи и особенностей коллекции.

Модель BTM описывает вероятность совместного появления слов $(u, v)$. Исходными данными являются частоты $n_{uv}$ битермов $(u, v)$ в коллекции, или матрица вероятностей $P = (p_{uv})_{W\times W}$, где $p_{uv} = \underset{(u,v)\in W^2}{\operatorname{norm}}(n_{uv})$.

Примем гипотезу условной независимости $p(u, v\,|\,t) = p(u\,|\,t)\,p(v\,|\,t)$, то есть допустим, что слова битермов порождаются независимо друг от друга из одной и той же темы. Тогда, по формуле полной вероятности,

$$p(u,v) = \sum_{t\in T} p(u\,|\,t)\,p(v\,|\,t)\,p(t) = \sum_{t\in T} \varphi_{ut}\varphi_{vt}\pi_t,$$

где $\varphi_{wt} = p(w\,|\,t)$ и $\pi_t = p(t)$ — параметры тематической модели. Это трёхматричное разложение $P = \Phi\Pi\Phi^{\mathrm{T}}$, где $\Pi = \operatorname{diag}(\pi_1, \ldots, \pi_T)$ — диагональная матрица. Модель битермов не определяет тематику документов $\Theta$ и поэтому не подвержена влиянию эффектов, вызванных короткими текстами.

ARTM позволяет объединить модель битермов с обычной тематической моделью, чтобы всё-таки получить матрицу $\Theta$. Для этого возьмём log-правдоподобие модели

битермов в качестве регуляризатора с коэффициентом $\tau$:

$$R(\Phi, \Pi) = \tau \sum_{u,v} n_{uv} \ln \sum_t \varphi_{ut}\varphi_{vt}\pi_t.$$

Применение уравнений (12)–(13) к этому регуляризатору даёт формулы М-шага:

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\Big(n_{wt} + \tau \sum_{u \in W} n_{uw} p_{tuw}\Big);$$
$$p_{tuw} = \underset{t \in T}{\operatorname{norm}}\big(n_t \varphi_{wt} \varphi_{ut}\big).$$

Эти формулы интерпретируются как добавление *псевдо-документов*. Каждому слову $u \in W$ ставится в соответствие псевдо-документ $d_u$, объединяющий все контексты слова $u$, то есть это мешок слов, встретившихся рядом со словом $u$ по всей коллекции. Число вхождений слова $w$ в псевдо-документ $d_u$ равно $\tau n_{uw}$. Вспомогательные переменные $p_{tuw} = p(t \,|\, u, w)$ соответствуют формуле Е-шага для псевдо-документа $d_u$, если доопределить его тематику как $\theta_{tu} = \underset{t}{\operatorname{norm}}(n_t \varphi_{ut})$. Другими словами, в модели битермов столбцы матрицы $\Theta$, соответствующие псевдо-документам, образуются путём перенормировки строк матрицы $\Phi$ по формуле Байеса.

Увеличивая коэффициент $\tau$, можно добиться того, чтобы матрица $\Phi$ формировалась практически только по битермам. В таком случае модель ARTM переходит в модель битермов, которая строится по коллекции псевдо-документов, без использования исходных документов.

**Сеть слов.** Идея моделировать не документы, а связи между словами, была положена в основу тематических моделей совстречаемости слов WTM (word topic model) [31] и WNTM (word network topic model) [156]. Любопытно, что более ранняя публикация модели WTM осталась незамеченной (видимо, как не-байесовская), и во второй статье даже нет ссылки на неё. Модели WTM и WNTM сводятся к применению PLSA и LDA соответственно к коллекции псевдо-документов $d_u$:

$$p(w \,|\, d_u) = \sum_{t \in T} p(w \,|\, t)\, p(t \,|\, d_u) = \sum_{t \in T} \varphi_{wt}\theta_{tu}.$$

Запишем log-правдоподобие модели $p(w \,|\, d_u)$ в виде регуляризатора:

$$R(\Phi, \Theta) = \tau \sum_{u,w \in W} n_{uw} \ln \sum_{t \in T} \varphi_{wt}\theta_{tu},$$

где $n_{uw}$ — совстречаемость слов $u, w$ (кстати, $n_{uw} = n_{wu}$).

Основное отличие этих моделей от модели битермов в том, что здесь в явном виде строится матрица $\Theta$ для псевдо-коллекции, тогда как в модели битермов $\Theta = \operatorname{diag}(\pi_1, \ldots, \pi_t)\Phi^{\mathsf{T}}$ и количество параметров вдвое меньше. Как показали эксперименты на коллекциях коротких текстов, модель WNTM немного превосходит модель битермов и существенно превосходит обычные тематические модели [156]. На коллекциях длинных документов тематические модели совстречаемости слов не дают значимых преимуществ перед обычными тематическими моделями.

**Когерентность.** Тема называется *когерентной* (согласованной), если наиболее частые термы данной темы часто встречаются рядом в документах коллекции [87]. Совстречаемость термов может оцениваться по самой коллекции $D$ [81], или по сторонней коллекции, например, по Википедии [84]. Средняя когерентность тем считается хорошей мерой интерпретируемости тематической модели [88].

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

Запишем формулу полной вероятности $p(w\,|\,t) = \sum_v C_{wv}\varphi_{vt}$ и заменим в ней условную вероятность $\varphi_{vt}$ частотной оценкой: $\hat{p}(w\,|\,t) = \sum_v C_{wv}\frac{n_{vt}}{n_t}$. Введём регуляризатор, требующий, чтобы параметры $\varphi_{wt}$ тематической модели были согласованы с оценками $\hat{p}(w\,|\,t)$ в смысле кросс-энтропии:

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

Формула М-шага, согласно (12), принимает вид

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

Это сглаживающий регуляризатор. Он увеличивает вероятность терма в теме, если термы, с которыми он часто совместно встречается, относятся к данной теме. Точно такая же формула получилась в [81] для модели LDA и алгоритма сэмплирования Гиббса, но с более сложным обоснованием через обобщённую урновую схему Пойя, и с более сложной эвристической оценкой $C_{wv}$.

В работе [84] предложен другой регуляризатор когерентности:

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

в котором оценка совместной встречаемости $C_{uv} = N_{uv}\big[\mathrm{PMI}(u, v) > 0\big]$ определяется через *поточечную взаимную информацию* (pointwise mutual information)

$$\mathrm{PMI}(u, v) = \ln \frac{|D|N_{uv}}{N_u N_v}, \tag{32}$$

где $N_{uv}$ — число документов, в которых термы $u, v$ хотя бы один раз встречаются рядом (не далее, чем через 10 слов), $N_u$ — число документов, в которых терм $u$ встречается хотя бы один раз.

Таким образом, в литературе пока отсутствует единый подход к оптимизации когерентности. Предлагаемые критерии похожи на модели битермов и сети слов. Все они формализуют общую идею, что если слова часто совместно встречаются, то они имеют схожую тематику.

**Модели векторных представлений слов** ставят в соответствие каждому слову $w$ вектор $\nu_w$ фиксированной размерности. Основное требование к этому отображению — чтобы близким по смыслу словам соответствовали близкие векторы. Согласно *дистрибутивной гипотезе* (distributional hypothesis) смысл слова определяется

распределением слов, в окружении которых оно встречается [46]. Слова, встречающиеся в схожих контекстах, имеют схожую семантику и, соответственно, должны иметь близкие векторы. Для формализации этого принципа в [76, 77] предлагается несколько вероятностных моделей, и все они реализованы в программе word2vec. В частности, модель skip-gram предсказывает появление слова $w$ в контексте слова $u$, то есть при условии, что слово $u$ находится рядом:

$$p(w\,|\,u) = \operatorname*{SoftMax}_{w\in W} \langle \nu_w, \nu_u \rangle = \operatorname*{norm}_{w\in W}\big(\exp \langle \nu_w, \nu_u \rangle\big) = \frac{\exp \langle \nu_w, \nu_u \rangle}{\sum_v \exp \langle \nu_v, \nu_u \rangle},$$

где $\langle \nu_w, \nu_u \rangle = \sum_t \nu_{wt}\nu_{ut}$ — скалярное произведение векторов. В отличие от тематических моделей, нормировка вероятностей производится нелинейным преобразованием SoftMax, а сами векторные представления слов не нормируются.

Для обучения модели решается задача максимизации log-правдоподобия, как правило, градиентными методами:

$$\sum_{u,w\in W} n_{uw} \ln p(w\,|\,u) \;\to\; \max_{\{\nu_w\}}.$$

Постановка задачи очень похожа на тематические модели BTM и WNTM. Модели семейства word2vec и другие модели векторных представлений слов также являются матричными разложениями [63, 95, 68]. Главное отличие заключается в том, что в этих векторных представлениях координаты не интерпретируемы, не нормированы и не разрежены, тогда как в тематических моделях словам соответствуют разреженные дискретные распределения тем $p(t\,|\,w)$. С другой стороны, тематические модели изначально не предназначались для определения семантической близости слов, поэтому делают они это плохо.

В работе А. С. Попова[4] предложен способ построения *тематических векторных представлений слов* по псевдо-коллекции документов, аналогичный моделям BTM и WNTM. В задачах семантической близости слов они конкурируют с моделями word2vec и существенно превосходят обычные тематические модели. При этом тематические векторные представления являются интерпретируемыми и разреженными. Используя кросс-энтропийные регуляризаторы, разреженность векторов удаётся доводить до 93% без потери качества.

Кроме того, APTM позволяет обобщить тематические модели дистрибутивной семантики для мультимодальных коллекций. Используя данные о совстречаемости токенов различных модальностей, возможно строить интерпретируемые тематические векторные представления для всех модальностей. В то же время привлечение дополнительной информации о других модальностях повышает качество решения задачи близости слов.

# 12 Тематическая сегментация

Гипотеза «мешка слов» и предположение о статистической независимости соседних слов приводят к слишком частой хаотичной смене тематики между соседними словами. Если проследить, к каким темам относятся последовательные слова в тексте,

[4] Артём Попов. Регуляризация тематических моделей для векторных представлений слов. Бакалаврская диссертация, ВМК МГУ, 2017.
`http://www.MachineLearning.ru/wiki/images/4/45/2017PopovBsc.pdf`

то тематическая модель в целом покажется не настолько хорошо интерпретируемой, как ранжированные списки наиболее частотных слов в темах.

Тематические модели сегментации основаны на более реалистичных гипотезах о связном тексте. Каждое предложение относится к одной теме, иногда к небольшому числу тем. Следующее предложение часто продолжает тематику предыдущего. Смена темы чаще происходит между абзацами, ещё чаще между секциями документа. Каждое предложение можно считать «мешком термов».

**Тематическая модель предложений.** Допустим, что каждый документ $d$ разбит на множество сегментов $S_d$. Это могут быть предложения, абзацы или *фразы* — синтаксически корректные части предложений. Обозначим через $n_s$ длину сегмента $s$, через $n_{sw}$ — число вхождений терма $w$ в сегмент $s$.

Предположим, что все слова сегмента относятся к одной теме и запишем функцию вероятности сегмента $s \in S_d$ через параметры тематической модели $\varphi_{wt}$, $\theta_{td}$:

$$p(s\,|\,d) = \sum_{t\in T} p(t\,|\,d) \prod_{w\in s} p(w\,|\,t)^{n_{sw}} = \sum_{t\in T} \theta_{td} \prod_{w\in s} \varphi_{wt}^{n_{sw}}.$$

Будем считать каждый документ «мешком сегментов». Тогда функция вероятности выборки будет равна произведению функций вероятности сегментов. Поставим задачу максимизации суммы log-правдоподобия и регуляризатора $R$:

$$\sum_{d\in D} \sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \varphi_{wt}^{n_{sw}} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}, \tag{33}$$

при обычных ограничениях (9). В частном случае, когда каждый сегмент состоит только из одного слова, данная задача переходит в (10).

**Теорема 4.** *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка $(\Phi,\Theta)$ локального экстремума задачи* (33), (9) *удовлетворяет системе уравнений со вспомогательными переменными $p_{tds} \equiv p(t\,|\,d,s)$:*

$$p_{tds} = \underset{t\in T}{\operatorname{norm}}\big(\theta_{td} \prod_{w\in s} \varphi_{wt}^{n_{sw}}\big);$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D}\sum_{s\in S_d} [w\in s]\, p_{tds}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{s\in S_d} p_{tds}$$

Аналогичная задача ставилась для модели коротких сообщений Twitter-LDA [153], только в роли документов выступали авторы, в роли сегментов — все сообщения данного автора.

Тематическая модель предложений senLDA [19] имеет более важное структурное отличие: вместо матрицы параметров $\theta_{td} = p(t\,|\,d)$ в senLDA используется вектор параметров $\pi_t = p(t)$. Тем самым игнорируется разделение множества всех предложений коллекции по документам, что позволяет уменьшить число параметров модели. Если в senLDA нужно узнать тематику документа, то её нетрудно вычислить, усреднив тематику всех его предложений.

**Тематическая модель сегментации.** Теперь рассмотрим более сложный случай, когда текст состоит из предложений, и требуется объединить их в более крупные тематические сегменты, границы которых заранее не определены.

*Метод TopicTiling* [103] основан на пост-обработке распределений $p(t\,|\,d, w_i)$, $i = 1, \ldots, n$, получаемых какой-либо тематической моделью, например, LDA. Определим тематику предложения $s$ как среднюю тематику $p(t\,|\,d, w)$ всех его слов $w$. Посчитаем косинусную близость тематики для всех пар соседних предложений. Чем глубже локальный минимум близости, тем выше уверенность, что между данной парой предложений проходит граница сегментов. Метод TopicTiling использует набор эвристик для подбора числа предложений слева и справа от локального минимума близости, определения числа сегментов, подбора числа тем и числа итераций, игнорирования стоп-слов, фоновых тем и коротких предложений. Аккуратная настройка параметров этих эвристик позволяет достичь высокого качества сегментации [103].

TopicTiling не является полноценной тематической моделью сегментации текста, поскольку пост-обработка никак не влияет на сами темы. Чтобы найти темы, наиболее выгодные для сегментации, требуется специальный регуляризатор.

**Регуляризатор E-шага.** Некоторые требования к тематической модели удобнее выражать через распределения $p_{tdw} = p(t\,|\,d, w)$, а не через $\varphi_{wt}$ и $\theta_{td}$. Например, требования сходства тематики термов внутри предложений или соседних предложений внутри документа. Таким способом можно учитывать порядок слов внутри документов в обход гипотезы «мешка слов».

Рассмотрим регуляризатор $R(\Pi)$ как функцию от трёхмерной матрицы вспомогательных переменных $\Pi = (p_{tdw})_{T\times D\times W}$. Согласно уравнению (11), матрица $\Pi$ является функцией от $\Phi$ и $\Theta$. Поэтому к регуляризатору $R(\Pi(\Phi, \Theta))$ применима теорема 1.

Рассмотрим задачу максимизации регуляризованного log-правдоподобия с двумя регуляризаторами, один из которых зависит от $\Pi$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Pi(\Phi, \Theta)) + R'(\Phi, \Theta) \;\to\; \max_{\Phi,\Theta}, \tag{34}$$

при ограничениях неотрицательности и нормировки (9).

**Теорема 5.** *Пусть функции $R(\Pi(\Phi, \Theta))$ и $R'(\Phi, \Theta)$ непрерывно дифференцируемы и функция $R(\Pi)$ не зависит от переменных $p_{tdw}$ в случае $n_{dw} = 0$. Тогда точка $(\Phi, \Theta)$ локального экстремума задачи (34), (9) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} \equiv p(t\,|\,d, w)$ и $\tilde{p}_{tdw}$:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{35}$$

$$\tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right); \tag{36}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \varphi_{wt}\frac{\partial R'}{\partial \varphi_{wt}}\right); \tag{37}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R'}{\partial \theta_{td}}\right). \tag{38}$$

Таким образом, в ЕМ-алгоритме для каждого документа $d$ сначала вычисляются вспомогательные переменные $p_{tdw}$, затем они преобразуются в новые переменные $\tilde{p}_{tdw}$, которые подставляются в обычные формулы М-шага (12)–(13) вместо $p_{tdw}$. Такой способ вычислений будем называть *регуляризацией Е-шага*.

Переменные $\tilde{p}_{tdw}$ могут принимать отрицательные значения, поэтому в общем случае они не образуют вероятностных распределений. Тем не менее, условие нормировки для них выполнено всегда.

**Разреживающий регуляризатор Е-шага для сегментации.** Применим регуляризацию Е-шага для построения тематической модели сегментированного текста. Определим тематику сегмента $s \in S_d$ как среднюю тематику всех его термов:

$$p_{tds} \equiv p(t\,|\,d,s) = \sum_{w \in s} p(t\,|\,d,w)\,p(w\,|\,s) = \frac{1}{n_s} \sum_{w \in s} n_{sw} p_{tdw}.$$

Чтобы каждый сегмент относился к небольшому числу тем, будем минимизировать кросс-энтропию между распределениями $p(t\,|\,d,s)$ и равномерным распределением, что приведёт нас к разреживающему регуляризатору Е-шага:

$$R(\Pi) = -\tau \sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T} \ln \sum_{w \in s} n_{sw} p_{tdw}. \tag{39}$$

Опуская рутинные выкладки, приведём результат подстановки (39) в (36):

$$\tilde{p}_{tdw} = p_{tdw} \left( 1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s} \left( \frac{1}{p_{tds}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zds}} \right) \right).$$

Хотя формула выглядит громоздкой, эффект применения регуляризатора понять не трудно. Если вероятность $p_{tds}$ темы в сегменте окажется меньше некоторого порога, то вероятности $p_{tdw}$ будут уменьшаться для всех термов $w$ данного сегмента. В итоге тематика каждого сегмента сконцентрируется в небольшом числе тем.

В результате разреживания тематика соседних сегментов может оказаться близкой, и их можно будет объединить в один тематический сегмент. Назовём тему $t$ с максимальным значением $p(t\,|\,d,s)$ *доминирующей темой* сегмента $s$ документа $d$. Если тема доминирует в соседних сегментах, то она будет доминирующей и в их объединении. Если объединить последовательные сегменты с одинаковой доминирующей темой в один более крупный сегмент, то данная тема также останется в нём доминирующей. Это простая агломеративная стратегия тематической сегментации. В отличие от TopicTiling, у неё нет эвристических параметров, которые надо настраивать, и она почти не увеличивает время пост-обработки Е-шага.

# 13 Критерии качества

Количественное оценивание тематических моделей является нетривиальной проблемой. В отличие от задач классификации или регрессии здесь нет чёткого понятия «ошибки» или «потери». Критерии качества кластеризации типа средних внутрикластерных или межкластерных расстояний плохо подходят для оценивания «мягкой» совместной кластеризации документов и термов.

Критерии качества тематических моделей принято делить на внутренние (intrinsic) и внешние (extrinsic). *Внутренние критерии* характеризуют качество модели по исходной текстовой коллекции. *Внешние критерии* оценивают полезность модели с точки зрения приложения и конечных пользователей. Иногда для этого приходится собирать дополнительные данные, например, оценки асессоров.

**Внешние критерии** крайне разнообразны и зависят от решаемой прикладной задачи. Практически в каждой публикации по тематическому моделированию используется какой-либо внешний критерий: качество классификации документов [106], точность и полнота информационного поиска [147, 14, 7, 12], число найденных хорошо интерпретируемых тем [17], качество сегментации текстов [103]. В [34] предлагается методика диагностики моделей, основанная на сопоставлении найденных тем с заранее известными концептами.

**Перплексия.** Наиболее распространённым внутренним критерием является *перплексия* (perplexity), используемая для оценивания моделей языка в компьютерной лингвистике. Это мера несоответствия или «удивлённости» модели $p(w\,|\,d)$ токенам $w$, наблюдаемым в документах $d$ коллекции $D$. Она определяется через log-правдоподобие (8), а в случае мультимодальной модели — через log-правдоподобие (23) отдельно для каждой модальности:

$$\text{perplexity}_m(D;p) = \exp\biggl(-\frac{1}{n_m}\sum_{d\in D}\sum_{w\in W^m} n_{dw}\ln p(w\,|\,d)\biggr), \tag{40}$$

где $n_m = \sum\limits_{d\in D}\sum\limits_{w\in W^m} n_{dw}$ — длина коллекции по $m$-й модальности.

Чем меньше величина перплексии, тем лучше модель $p$ предсказывает появление токенов $w$ в документах $d$ коллекции $D$.

Перплексия имеет следующую интерпретацию. Если термы $w$ порождаются из равномерного распределения $p(w) = 1/V$ на словаре мощности $V$, то перплексия модели $p(w)$ на таком тексте сходится к $V$ с ростом его длины. Чем сильнее распределение $p(w)$ отличается от равномерного, тем меньше перплексия. В случае условных вероятностей $p(w\,|\,d)$ интерпретация немного другая: если каждый документ генерируется из $V$ равновероятных термов (возможно, различных в разных документах), то перплексия сходится к $V$.

Недостатком перплексии является неочевидность её численных значений, а также её зависимость не только от качества модели, но и от ряда посторонних факторов — длины документов, мощности и разреженности словаря. В частности, с помощью перплексии некорректно сравнивать тематические модели одной и той же коллекции, построенные на разных словарях.

Обозначим через $p_D(w\,|\,d)$ модель, построенную по обучающей коллекции документов $D$. Перплексия обучающей выборки $\mathscr{P}_m(D;p_D)$ является оптимистично смещённой (заниженной) характеристикой качества модели из-за эффекта переобучения. Обобщающую способность тематических моделей принято оценивать *перплексией контрольной выборки* (hold-out perplexity) $\mathscr{P}_m(D';p_D)$. Обычно коллекцию разделяют на обучающую и контрольную случайным образом в пропорции $9:1$ [26].

Недостатком контрольный перплексии является высокая чувствительность к редким и новым словам, которые практически бесполезны для тематических моделей.

В ранних экспериментах было показано, что LDA существенно превосходит PLSA по перплексии, откуда был сделан вывод, что LDA меньше переобучается [26]. В [4, 98, 5] были предложены *робастные тематические модели*, описывающие редкие слова специальным «фоновым» распределением. Перплексия робастных вариантов PLSA и LDA оказалась существенно меньшей и практически одинаковой.

**Когерентность.** 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 [29]. Свойство интерпретируемости важно в информационно-поисковых системах для систематизации и визуализации результатов тематического поиска или категоризации документов.

Большинство существующих методов оценивания интерпретируемости основано на привлечении экспертов-асессоров. В [86] экспертам предлагалось непосредственно оценивать полезность тем по трёхбалльной шкале. В методе интрузий [29] для каждой найденной темы составляется список из 10 наиболее частотных слов, в который внедряется одно случайное слово. Тема считается интерпретируемой, если подавляющее большинство экспертов правильно указывают лишнее слово. Экспертные подходы необходимы на стадии исследований, но они затрудняют автоматическое построение тематических моделей. В серии работ [86, 87, 87, 81] показано, что среди величин, вычисляемых по коллекции автоматически, лучше всего коррелирует с экспертными оценками интерпретируемости *когерентность* (coherence).

Тема называется *когерентной* (согласованной), если термы, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции [87, 88]. Численной мерой когерентности темы $t$ является поточечная взаимная информация (32), вычисляемая по $k$ наиболее вероятным словам темы:

$$\mathrm{PMI}(t) = \sum_{i=1}^{k-1} \sum_{j=i}^{k} \mathrm{PMI}(w_i, w_j),$$

где $w_i$ — $i$-й терм в порядке убывания $\varphi_{wt}$, число $k$ обычно полагается равным 10.

Когерентность модели определяется как средняя когерентность тем. Когерентность может оцениваться по сторонней коллекции (например, по Википедии) [84], либо по той же коллекции, по которой строится модель [81].

**Разреженность и различность тем.** Разреженность модели измеряется долей нулевых элементов в матрицах $\Phi$ и $\Theta$. В моделях, разделяющих множество тем $T$ на предметные $S$ и фоновые $B$, разреженность оценивается только по частям матриц $\Phi, \Theta$, соответствующим предметным темам.

В [127] вводятся косвенные меры интерпретируемости тем, не требующие привлечения асессоров. Предполагается, что интерпретируемая тема должна содержать *лексическое ядро* — множество слов, которые с большой вероятностью употребляются в данной теме и редко употребляются в других темах. В таком случае матрицы $\Phi$ и $\Theta$ должны обладать структурой разреженности, аналогичной рис. 4.

Ядро $W_t = \{w \in W \mid p(t|w) > 0.25\}$ темы $t$ определяется как множество термов, которые имеют высокую условную вероятность $p(t|w) = \varphi_{wt}\frac{n_t}{n_w}$ для данной темы. Затем по ядру определяется три показателя интерпретируемости темы $t$:

$\text{pur}_t = \sum\limits_{w \in W_t} p(w|t)$ — *чистота темы* (чем выше, тем лучше);
$\text{con}_t = \frac{1}{|W_t|} \sum\limits_{w \in W_t} p(t|w)$ — *контрастность темы* (чем выше, тем лучше);
$\text{ker}_t = |W_t|$ — *размер ядра* (ориентировочный оптимум $\frac{|W|}{|T|}$).

Показатели размера ядра, чистоты и контрастности для модели в целом определяются как средние по всем предметным темам $t \in S$.

*Доля фоновых слов* во всей коллекции

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

принимает значения от 0 до 1. Значения, близкие к 0, говорят о том, что модель не способна отделять слова общей лексики от специальной терминологии. Значения, близкие к 1, свидетельствуют о вырождении тематической модели.

Такие критерии, как размер ядра или доля фоновых слов, могут использоваться для контроля адекватности модели. Чрезмерная регуляризация может приводить к деградации тем или вырождению модели для слишком большой доли документов. Образно говоря, регуляризаторы в малых дозах являются лекарствами, но в случае передозировки могут превращаться в яд. Многие критерии, включая перплексию, слабо чувствительны к некоторым типам вырождения, например, когда в предметных темах остаётся слишком мало слов.

На практике к тематическим моделям предъявляются сочетания разнообразных требований. Задачи тематического моделирования по сути являются многокритериальными, поэтому и качество модели должно оцениваться по многим критериям.

В проекте `BigARTM` поддерживается библиотека стандартных метрик качества и механизмы добавления новых пользовательских метрик.

# 14 Разведочный информационный поиск

Важным приложением тематического моделирования является *информационный поиск* (information retrieval) [147, 14]. Современные поисковые системы предназначены, главным образом, для поиска конкретных ответов на короткие текстовые запросы. Другие поисковые потребности возникают у пользователей, которым необходимо разобраться в новой предметной области или пополнить свой багаж знаний. Пользователь может не владеть терминологией, слабо понимать структуру предметной области, не иметь точных формулировок запроса и не подразумевать единственный правильный ответ. В таких случаях нужен поиск не по ключевым словам, а по смыслу. Запросом может быть длинный фрагмент текста, документ или подборка документов. Результатом поиска должна быть удобно систематизированная информация, «дорожная карта» предметной области.

Для этих случаев подходит парадигма *разведочного информационного поиска* (exploratory search) [71, 142]. Его целью является получение ответов на сложные вопросы: «какие темы представлены в тексте запроса», «что читать в первую очередь по этим темам», «что находится на стыке этих тем со смежными областями», «какова тематическая структура данной предметной области», «как она развивалась во времени», «каковы последние достижения», «где находятся основные центры компетентности», «кто является экспертом по данной теме» и т. д. Пользователь обычной

поисковой системы вынужден итеративно переформулировать свои короткие запросы, расширяя зону поиска по мере усвоения терминологии предметной области, периодически пересматривая и систематизируя результаты поиска. Это требует затрат времени и высокой квалификации. При отсутствии инструмента для получения «общей картины» остаётся сомнение, что какие-то важные аспекты изучаемой проблемы так и не были найдены. Если образно представить итеративный поиск как блуждание по лабиринту знаний, то разведочный поиск — это средство автоматического построения карты для любой части этого лабиринта.

**Тематический разведочный поиск.** Обычные (полнотекстовые) поисковые системы основаны на инвертированных индексах, в которых для каждого слова хранится список содержащих его документов [9]. Поисковая система ищет документы, содержащие все слова запроса, поэтому по длинному запросу, скорее всего, ничего не будет найдено.

Система тематического разведочного поиска сначала строит тематическую модель запроса и определяет короткий список тем запроса. Затем для поиска документов схожей тематики применяются те же механизмы индексирования и поиска, только в роли слов выступают темы. Поскольку число тем на несколько порядков меньше объёма словаря, тематический поиск требует намного меньше памяти по сравнению с полнотекстовым поиском и может быть реализован на весьма скромной технике. Технологии информационного поиска на основе тематического моделирования в настоящее время находятся в стадии исследований и разработок [116, 21, 94, 28, 13, 134].

В литературе по разведочному поиску тематическое моделирование стали использовать относительно недавно [107, 45, 104, 124], а многие обзоры о нём вообще не упоминают [41, 101, 114, 54, 72, 51]. В недавней статье [124] важными преимуществами тематических моделей называются гибкость, возможности визуализации и навигации. В то же время, в качестве недостатков отмечаются проблемы с интерпретируемостью тем, трудности с модификацией тематической модели при поступлении новых документов и высокая вычислительная сложность. Эти проблемы относятся к устаревшим методам и успешно решены в последние годы: десятки новых моделей разработаны для улучшения интерпретируемости; онлайновые алгоритмы способны обрабатывать большие коллекции и потоки документов за линейное время [78, 20, 125]. С другой стороны, в работах по тематическому моделированию разведочный поиск часто называют одним из важнейших приложений, а оценки качества поиска используют для валидации моделей [147, 14]. Однако эти исследования пока не привели к созданию общедоступных систем разведочного поиска. Всё это говорит о разобщённости научных сообществ, разрабатывающих эти два направления. Тенденция к их сближению наметилась лишь в последние годы.

Такие приложения, как разведочный поиск, стимулируют развитие многокритериального тематического моделирования. Тематическая модель для разведочного поиска в идеале должна быть интерпретируемой, разреженной, мультиграммной, мультимодальной, мультиязычной, иерархической, динамической, сегментирующей, обучаемой по оценкам асессоров или логам пользователей. Также она должна автоматически определять число тем на каждом уровне иерархии и автоматически создавать и именовать новые темы. Наконец, она должна быть онлайновой, параллельной и распределённой, чтобы эффективно обрабатывать большие коллекции текстов.

Таким образом, многие из рассмотренных в данном обзоре моделей должны быть скомбинированы для создания полнофункционального разведочного поиска.

**Качество разведочного поиска.** Модель ARTM для разведочного поиска была предложена в [12] и улучшена в [145]. Для измерения качества разведочного тематического поиска использовались критерии точности и полноты на основе оценок асессоров. Для оценивания была составлена выборка запросов — заданий разведочного поиска. Каждый запрос представлял собой текст объёмом около одной страницы формата A4, описывающий тематику поиска. Каждое задание сначала выполнялось независимо несколькими асессорами, затем системой тематического поиска, затем релевантность найденных системой документов снова оценивалась асессорами. Данная методика позволяет, единожды сделав разметку результатов поиска, многократно оценивать качество различных тематических моделей и алгоритмов поиска. Эксперименты на коллекциях 175 тысяч статей русскоязычного коллективного блога `habrahabr.ru` и 760 тысяч статей англоязычного блога `techcrunch.com` показали, что тематический поиск находит больше релевантных документов, чем асессоры, сокращая среднее время поиска с получаса до секунды. Комбинирование регуляризаторов декоррелирования, разреживания и сглаживания вместе с модальностями $n$-грамм, авторов и категорий значимо улучшает качество поиска и позволяет достичь точности выше 80% и полноты выше 90%.

**Визуализация.** Систематизация результатов тематического поиска невозможна без интерактивного графического представления. В обзоре [2] описываются и сравниваются 16 средств визуализации тематических моделей на основе веб-интерфейсов. Ещё больше идей можно почерпнуть из интерактивного обзора[5], который на момент написания данной статьи насчитывал 380 средств визуализации текстов. Несмотря на такое богатство технических решений, основных идей визуализации тематических моделей не так много: это либо двумерное отображение семантической близости тем в виде графа или «дорожной карты», либо тематическая иерархия, либо динамика развития тем во времени, либо графовая структура взаимосвязей между темами, документами, авторами или иными модальностями, либо сегментная структура отдельных документов.

Статичные визуализации практически бесполезны при графической визуализации больших данных. Это было понято более 20 лет назад и сформулировано Беном Шнейдерманом в виде *мантры визуального поиска информации*: «сначала крупный план, затем масшабирование и фильтрация, детали по требованию»[6] [112].

Отображение результатов тематического моделирования и разведочного поиска соответствует концепции дальнего чтения (distant reading) социолога литературы Франко Моретти [82]. Он противопоставляет этот способ изучения текстов нашему обычному чтению (close reading). Невозможно прочитать сотни миллионов книг или статей, но вполне возможно применить статистические методы и графическую визуализацию, чтобы понять в общих чертах, о чём вся эта литература, и научиться быстрее отыскивать нужное. «*Дальнее чтение* — это специальная форма представ-

[5] `http://textvis.lnu.se` — интерактивный обзор средств визуализации текстов.

[6] Visual Information Seeking Mantra: «Overview first, zoom and filter, details on demand» [112].

ления знаний, в которой меньше элементов, грубее их взаимосвязи, остаются лишь формы, очертания, структуры, модели»[7].

Для библиотеки `BigARTM` в настоящее время развивается собственный инструмент визуализации на основе веб-интерфейса VisARTM[8], поддерживающий важнейшие формы представления тематических моделей. Интересной возможностью VisARTM является построение *спектра тем* — оптимальное ранжирование списка тем, при котором семантически близкие темы оказываются в списке рядом. Это помогает пользователям быстрее находить темы и группировать их по смыслу.

# 15 Заключение

Данный обзор написан по материалам спецкурса «Вероятностное тематическое моделирование»[9], который автор читает на факультете ВМК Московского Государственного Университета им. М. В. Ломоносова. Обновляемая электронная версия доступна на сайте MachineLearning.ru[10],

Что не вошло в этот обзор, но может оказаться в ближайших обновлениях: доказательства пяти теорем; стратегии подбора коэффициентов регуляризации; методы суммаризации и автоматического именования тем; примеры применения тематических моделей для автоматического выделения терминов, обнаружения новых тем и отслеживания сюжетов, анализа тональности и выявления мнений, анализа записей разговоров контакт-центра, анализа банковских транзакционных данных, агрегации и категоризации научного контента.

# Список литературы


[1] *Агеев М. С., Добров Б. В., Лукашевич Н. В.* Автоматическая рубрикация текстов: методы и проблемы // *Учёные записки Казанского государственного университета. Серия Физико-математические науки.* — 2008. — Т. 150, № 4. — С. 25–40.

[2] *Айсина Р. М.* Обзор средств визуализации тематических моделей коллекций текстовых документов // *Машинное обучение и анализ данных (http://jmlda.org).* — 2015. — Т. 1, № 11. — С. 1584–1618.

[3] *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // *Доклады РАН.* — 2014. — Т. 456, № 3. — С. 268–271.

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

[5] *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.


[7] «*Distant reading* is not an obstacle but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models» [82].

[8] Дмитрий Федоряка. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация, ФУПМ МФТИ, 2017.
`http://www.MachineLearning.ru/wiki/images/d/d8/Fedoriaka17bsc.pdf`

[9] `http://www.MachineLearning.ru/wiki?title=BTM`.

[10] `http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf`.

[6] *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.

[7] *Дударенко М. А.* Регуляризация многоязычных тематических моделей // *Вычислительные методы и программирование.* — 2015. — Т. 16. — С. 26–38.

[8] *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска. — Издательство МГУ имени М. В. Ломоносова, 2011.

[9] *Маннинг К. Д., Рагхаван П., Шютце Х.* Введение в информационный поиск. — Вильямс, 2011.

[10] *Павлов А. С., Добров Б. В.* Метод обнаружения массово порожденных неестественных текстов на основе анализа тематической структуры // *Вычислительные методы и программирование: новые вычислительные технологии.* — 2011. — Т. 12. — С. 58–72.

[11] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1986.

[12] *Янина А. О., Воронцов К. В.* Мультимодальные тематические модели для разведочного поиска в коллективном блоге // *Машинное обучение и анализ данных.* — 2016. — Т. 2, № 2. — С. 173–186.

[13] *Airoldi E. M., Erosheva E. A., Fienberg S. E., Joutard C., Love T., Shringarpure S.* Reconceptualizing the classification of PNAS articles // *Proceedings of The National Academy of Sciences.* — 2010. — Vol. 107. — Pp. 20899–20904.

[14] *Andrzejewski D., Buttler D.* Latent topic feedback for information retrieval // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '11. — 2011. — Pp. 600–608.

[15] *Andrzejewski D., Zhu X.* Latent Dirichlet allocation with topic-in-set knowledge // Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. — SemiSupLearn '09. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. — Pp. 43–48.

[16] *Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K.* Additive regularization for topic modeling in sociological studies of user-generated text content // MICAI 2016, 15th Mexican International Conference on Artificial Intelligence. — Vol. 10061. — Springer, Lecture Notes in Artificial Intelligence, 2016. — P. 166–181.

[17] *Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K.* Mining ethnic content online with additively regularized topic models // *Computacion y Sistemas.* — 2016. — Vol. 20, no. 3. — P. 387–403.

[18] *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. — 2009. — Pp. 27–34.

[19] *Balikas G., Amini M., Clausel M.* On a topic model for sentences // Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. — SIGIR '16. — New York, NY, USA: ACM, 2016. — Pp. 921–924.

[20] *Bassiou N., Kotropoulos C.* Online PLSA: Batch updating techniques including out-of-vocabulary words // *Neural Networks and Learning Systems, IEEE Transactions on.* — Nov 2014. — Vol. 25, no. 11. — Pp. 1953–1966.

[21] *Blei D., Lafferty J.* A correlated topic model of Science // *Annals of Applied Statistics.* — 2007. — Vol. 1. — Pp. 17–35.

[22] *Blei D. M.* Probabilistic topic models // *Communications of the ACM.* — 2012. — Vol. 55, no. 4. — Pp. 77–84.

[23] *Blei D. M., Griffiths T., Jordan M., Tenenbaum J.* Hierarchical topic models and the nested chinese restaurant process // NIPS. — 2003.

[24] *Blei D. M., Griffiths T. L., Jordan M. I.* The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies // *J. ACM.* — 2010. — Vol. 57, no. 2. — Pp. 7:1–7:30.

[25] *Blei D. M., Jordan M. I.* Modeling annotated data // Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. — New York, NY, USA: ACM, 2003. — Pp. 127–134.

[26] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // *Journal of Machine Learning Research.* — 2003. — Vol. 3. — Pp. 993–1022.

[27] *Bodrunova S., Koltsov S., Koltsova O., Nikolenko S. I., Shimorina A.* Interval semi-supervised LDA: Classifying needles in a haystack // MICAI (1) / Ed. by F. C. Espinoza, A. F. Gelbukh, M. Gonzalez-Mendoza. — Vol. 8265 of *Lecture Notes in Computer Science.* — Springer, 2013. — Pp. 265–274.

[28] *Bolelli L., Ertekin S., Giles C. L.* Topic and trend detection in text collections using latent Dirichlet allocation // ECIR. — Vol. 5478 of *Lecture Notes in Computer Science.* — Springer, 2009. — Pp. 776–780.

[29] *Chang J., Gerrish S., Wang C., Boyd-Graber J. L., Blei D. M.* Reading tea leaves: How humans interpret topic models // Neural Information Processing Systems (NIPS). — 2009. — Pp. 288–296.

[30] *Chemudugunta C., Smyth P., Steyvers M.* Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems. — Vol. 19. — MIT Press, 2007. — Pp. 241–248.

[31] *Chen B.* Word topic models for spoken document retrieval and transcription. — 2009. — Vol. 8, no. 1. — Pp. 2:1–2:27.

[32] *Chien J.-T., Chang Y.-L.* Bayesian sparse topic model // *Journal of Signal Processessing Systems.* — 2013. — Vol. 74. — Pp. 375–389.

[33] *Chirkova N. A., Vorontsov K. V.* Additive regularization for hierarchical multimodal topic modeling // *Journal Machine Learning and Data Analysis.* — 2016. — Vol. 2, no. 2. — Pp. 187–200.

[34] *Chuang J., Gupta S., Manning C., Heer J.* Topic model diagnostics: Assessing domain relevance via topical alignment // Proceedings of the 30th International Conference on Machine Learning (ICML-13) / Ed. by S. Dasgupta, D. Mcallester. — Vol. 28. — JMLR Workshop and Conference Proceedings, 2013. — Pp. 612–620.

[35] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // *Frontiers of Computer Science in China.* — 2010. — Vol. 4, no. 2. — Pp. 280–301.

[36] *De Smet W., Moens M.-F.* Cross-language linking of news stories on the web using interlingual topic modelling // Proceedings of the 2Nd ACM Workshop on Social Web Search and Mining. — SWSM '09. — New York, NY, USA: ACM, 2009. — Pp. 57–64.

[37] *Dempster A. P., Laird N. M., Rubin D. B.* Maximum likelihood from incomplete data via the EM algorithm // *J. of the Royal Statistical Society, Series B.* — 1977. — no. 34. — Pp. 1–38.

[38] *Dietz L., Bickel S., Scheffer T.* Unsupervised prediction of citation influences // Proceedings of the 24th international conference on Machine learning. — ICML '07. — New York, NY, USA: ACM, 2007. — Pp. 233–240.

[39] *Eisenstein J., Ahmed A., Xing E. P.* Sparse additive generative models of text // ICML'11. — 2011. — Pp. 1041–1048.

[40] *El-Kishky A., Song Y., Wang C., Voss C. R., Han J.* Scalable topical phrase mining from text corpora // *Proc. VLDB Endowment.* — 2014. — Vol. 8, no. 3. — Pp. 305–316.

[41] *Feldman S. E.* The answer machine // Synthesis Lectures on Information Concepts, Retrieval, and Services. — Morgan & Claypool Publishers, 2012. — Vol. 4. — Pp. 1–137.

[42] *Feng Y., Lapata M.* Topic models for image annotation and text illustration // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Association for Computational Linguistics, 2010. — Pp. 831–839.

[43] *Frei O., Apishev M.* Parallel non-blocking deterministic algorithm for online topic modeling // AIST'2016, Analysis of Images, Social networks and Texts. — Vol. 661. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2016. — P. 132–144.

[44] *Girolami M., Kabán A.* On an equivalence between PLSI and LDA // SIGIR'03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. — 2003. — Pp. 433–434.

[45] *Grant C. E., George C. P., Kanjilal V., Nirkhiwale S., Wilson J. N., Wang D. Z.* A topic-based search, visualization, and exploration system // FLAIRS Conference. — AAAI Press, 2015. — Pp. 43–48.

[46] *Harris Z.* Distributional structure // *Word.* — 1954. — Vol. 10, no. 23. — Pp. 146–162.

[47] *Hoffman M. D., Blei D. M., Bach F. R.* Online learning for latent Dirichlet allocation // NIPS. — Curran Associates, Inc., 2010. — Pp. 856–864.

[48] *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. — Pp. 50–57.

[49] *Hospedales T., Gong S., Xiang T.* Video behaviour mining using a dynamic topic model // *International Journal of Computer Vision.* — 2012. — Vol. 98, no. 3. — Pp. 303–323.

[50] *Huang P.-S., He X., Gao J., Deng L., Acero A., Heck L.* Learning deep structured semantic models for web search using clickthrough data // Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management. — CIKM '13. — New York, NY, USA: ACM, 2013. — Pp. 2333–2338.

[51] *Jacksi K., Dimililer N., Zeebaree S. R. M.* A survey of exploratory search systems based on LOD resources // Proceedings of the 5th International Conference on Computing and Informatics, ICOCI 2015. — School of Computing, Universiti Utara Malaysia, 2015. — Pp. 501–509.

[52] *Jagarlamudi J., Daumé III H., Udupa R.* Incorporating lexical priors into topic models // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. — EACL'12. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. — Pp. 204–213.

[53] *Jameel S., Lam W.* An N-gram topic model for time-stamped documents // 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. — Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 2013. — Pp. 292–304.

[54] *Jiang T.* Exploratory Search: A Critical Analysis of the Theoretical Foundations, System Features, and Research Trends // Library and Information Sciences: Trends and Research / Ed. by C. Chen, R. Larsen. — Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. — Pp. 79–103.

[55] *Kataria S., Mitra P., Caragea C., Giles C. L.* Context sensitive topic models for author influence in document networks // Proceedings of the Twenty-Second international joint conference on Artificial Intelligence — Volume 3. — IJCAI'11. — AAAI Press, 2011. — Pp. 2274–2280.

[56] *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. — WebSci'14. — New York, NY, USA: ACM, 2014. — Pp. 161–165.

[57] *Konietzny S., Dietz L., McHardy A.* Inferring functional modules of protein families with probabilistic topic models // *BMC Bioinformatics.* — 2011. — Vol. 12, no. 1. — P. 141.

[58] *Krestel R., Fankhauser P., Nejdl W.* Latent Dirichlet allocation for tag recommendation // Proceedings of the third ACM conference on Recommender systems. — ACM, 2009. — Pp. 61–68.

[59] *La Rosa M., Fiannaca A., Rizzo R., Urso A.* Probabilistic topic modeling for the analysis and classification of genomic sequences // *BMC Bioinformatics.* — 2015. — Vol. 16, no. Suppl 6. — P. S2.

[60] *Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C.* Neural architectures for named entity recognition // HLT-NAACL / Ed. by K. Knight, A. Nenkova, O. Rambow. — The Association for Computational Linguistics, 2016. — Pp. 260–270.

[61] *Larsson M. O., Ugander J.* A concave regularization technique for sparse mixture models // Advances in Neural Information Processing Systems 24 / Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K. Weinberger. — 2011. — Pp. 1890–1898.

[62] *Lee S. S., Chung T., McLeod D.* Dynamic item recommendation by topic modeling for social networks // Information Technology: New Generations (ITNG), 2011 Eighth International Conference on. — IEEE, 2011. — Pp. 884–889.

[63] *Levy O., Goldberg Y.* Neural Word Embedding as Implicit Matrix Factorization // Advances in Neural Information Processing Systems 27 / Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger. — Curran Associates, Inc., 2014. — Pp. 2177–2185.

[64] *Li S., Li J., Pan R.* Tag-weighted topic model for mining semi-structured documents // IJCAI'13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. — AAAI Press, 2013. — Pp. 2855–2861.

[65] *Li W., McCallum A.* Pachinko allocation: Dag-structured mixture models of topic correlations // ICML. — 2006.

[66] *Li X.-X., Sun C.-B., Lu P., Wang X.-J., Zhong Y.-X.* Simultaneous image classification and annotation based on probabilistic model // *The Journal of China Universities of Posts and Telecommunications.* — 2012. — Vol. 19, no. 2. — Pp. 107–115.

[67] *Liu J., Shang J., Wang C., Ren X., Han J.* Mining quality phrases from massive text corpora // Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. — SIGMOD '15. — New York, NY, USA: ACM, 2015. — Pp. 1729–1744.

[68] *Liu Y., Liu Z., Chua T.-S., Sun M.* Topical word embeddings // Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. — AAAI'15. — AAAI Press, 2015. — Pp. 2418–2424.

[69] *Lu Y., Mei Q., Zhai C.* Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA // *Information Retrieval.* — 2011. — Vol. 14, no. 2. — Pp. 178–203.

[70] *M. A. Basher A. R., Fung B. C. M.* Analyzing topics and authors in chat logs for crime investigation // *Knowledge and Information Systems.* — 2014. — Vol. 39, no. 2. — Pp. 351–381.

[71] *Marchionini G.* Exploratory search: From finding to understanding // *Commun. ACM.* — 2006. — Vol. 49, no. 4. — Pp. 41–46.

[72] *Marie N., Gandon F.* Survey of linked data based exploration systems // Proceedings of the 3rd International Workshop on Intelligent Exploration of Semantic Data (IESD 2014) co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 20, 2014. — 2014.

[73] *Masada T., Kiyasu S., Miyahara S.* Comparing LDA with pLSI as a dimensionality reduction method in document clustering // Proceedings of the 3rd International Conference on Large-scale knowledge resources: construction and application. — LKR'08. — Springer-Verlag, 2008. — Pp. 13–26.

[74] *McAuliffe J. D., Blei D. M.* Supervised topic models // Advances in Neural Information Processing Systems 20 / Ed. by J. C. Platt, D. Koller, Y. Singer, S. T. Roweis. — Curran Associates, Inc., 2008. — Pp. 121–128.

[75] *Mei Q., Cai D., Zhang D., Zhai C.* Topic modeling with network regularization // Proceedings of the 17th International Conference on World Wide Web. — WWW'08. — New York, NY, USA: ACM, 2008. — Pp. 101–110.

[76] *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space // *CoRR.* — 2013. — Vol. abs/1301.3781.

[77] *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* Distributed representations of words and phrases and their compositionality // *CoRR.* — 2013. — Vol. abs/1310.4546.

[78] *Mimno D., Hoffman M., Blei D.* Sparse stochastic inference for latent Dirichlet allocation // Proceedings of the 29th International Conference on Machine Learning (ICML-12) / Ed. by J. Langford, J. Pineau. — New York, NY, USA: Omnipress, July 2012. — Pp. 1599–1606.

[79] *Mimno D., Li W., McCallum A.* Mixtures of hierarchical topics with pachinko allocation // ICML. — 2007.

[80] *Mimno D., Wallach H. M., Naradowsky J., Smith D. A., McCallum A.* Polylingual topic models // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2. — EMNLP '09. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. — Pp. 880–889.

[81] *Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A.* Optimizing semantic coherence in topic models // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — EMNLP '11. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. — Pp. 262–272.

[82] *Moretti F.* Graphs, maps, trees : abstract models for literary history. — London; New York: Verso, 2007.

[83] *Nadeau D., Sekine S.* A survey of named entity recognition and classification // *Linguisticae Investigationes*. — 2007. — Vol. 30, no. 1. — Pp. 3–26.

[84] *Newman D., Bonilla E. V., Buntine W. L.* Improving topic coherence with regularized topic models // Advances in Neural Information Processing Systems 24 / Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K. Weinberger. — 2011. — Pp. 496–504.

[85] *Newman D., Chemudugunta C., Smyth P.* Statistical entity-topic models // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '06. — New York, NY, USA: ACM, 2006. — Pp. 680–686.

[86] *Newman D., Karimi S., Cavedon L.* External evaluation of topic models // Australasian Document Computing Symposium. — December 2009. — Pp. 11–18.

[87] *Newman D., Lau J. H., Grieser K., Baldwin T.* Automatic evaluation of topic coherence // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — HLT '10. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. — Pp. 100–108.

[88] *Newman D., Noh Y., Talley E., Karimi S., Baldwin T.* Evaluating topic models for digital libraries // Proceedings of the 10th annual Joint Conference on Digital libraries. — JCDL '10. — New York, NY, USA: ACM, 2010. — Pp. 215–224.

[89] *Ni J., Dinu G., Florian R.* Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection // The 55th Annual Meeting of the Association for Computational Linguistics (ACL). — 2017.

[90] *Ni X., Sun J.-T., Hu J., Chen Z.* Mining multilingual topics from wikipedia // Proceedings of the 18th International Conference on World Wide Web. — WWW '09. — New York, NY, USA: ACM, 2009. — Pp. 1155–1156.

[91] *Nikolenko S. I., Koltcov S., Koltsova O.* Topic modelling for qualitative studies // *Journal of Information Science*. — 2017. — Vol. 43, no. 1. — Pp. 88–102.

[92] *Paul M. J., Dredze M.* Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models // Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9–14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. — 2013. — Pp. 168–178.

[93] *Paul M. J., Dredze M.* Discovering health topics in social media using topic models // *PLoS ONE*. — 2014. — Vol. 9, no. 8.

[94] *Paul M. J., Girju R.* Topic modeling of research fields: An interdisciplinary perspective // RANLP. — RANLP 2009 Organising Committee / ACL, 2009. — Pp. 337–342.

[95] *Pennington J., Socher R., Manning C. D.* Glove: Global vectors for word representation // Empirical Methods in Natural Language Processing (EMNLP). — 2014. — Pp. 1532–1543.

[96] *Phuong D. V., Phuong T. M.* A keyword-topic model for contextual advertising // Proceedings of the Third Symposium on Information and Communication Technology. — SoICT '12. — New York, NY, USA: ACM, 2012. — Pp. 63–70.

[97] *Pinto J. C. L., Chahed T.* Modeling multi-topic information diffusion in social networks using latent Dirichlet allocation and Hawkes processes // Tenth International Conference on Signal-Image Technology & Internet-Based Systems. — 2014. — Pp. 339–346.

[98] *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.

[99] *Pritchard J. K., Stephens M., Donnelly P.* Inference of population structure using multilocus genotype data // *Genetics*. — 2000. — Vol. 155. — Pp. 945–959.

[100] *Pujara J., Skomoroch P.* Large-scale hierarchical topic models // NIPS Workshop on Big Learning. — 2012.

[101] *Rahman M.* Search engines going beyond keyword search: A survey // *International Journal of Computer Applications*. — August 2013. — Vol. 75, no. 17. — Pp. 1–8.

[102] *Ramage D., Hall D., Nallapati R., Manning C. D.* Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. — EMNLP '09. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. — Pp. 248–256.

[103] *Riedl M., Biemann C.* TopicTiling: A text segmentation algorithm based on LDA // Proceedings of ACL 2012 Student Research Workshop. — ACL '12. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. — Pp. 37–42.

[104] *Rönnqvist S.* Exploratory topic modeling with distributional semantics // Advances in Intelligent Data Analysis XIV: 14th International Symposium, IDA 2015, Saint Etienne. France, October 22 -24, 2015. Proceedings / Ed. by E. Fromont, T. De Bie, M. van Leeuwen. — Springer International Publishing, 2015. — Pp. 241–252.

[105] *Rosen-Zvi M., Griffiths T., Steyvers M., Smyth P.* The author-topic model for authors and documents // Proceedings of the 20th conference on Uncertainty in artificial intelligence. — UAI '04. — Arlington, Virginia, United States: AUAI Press, 2004. — Pp. 487–494.

[106] *Rubin T. N., Chambers A., Smyth P., Steyvers M.* Statistical topic models for multi-label document classification // *Machine Learning*. — 2012. — Vol. 88, no. 1-2. — Pp. 157–208.

[107] *Scherer M., von Landesberger T., Schreck T.* Topic modeling for search and exploration in multivariate research data repositories // Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings / Ed. by T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, C. J. Farrugia. — Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. — Pp. 370–373.

[108] *Shang J., Liu J., Jiang M., Ren X., Voss C. R., Han J.* Automated phrase mining from massive text corpora // *CoRR*. — 2017. — Vol. abs/1702.04457.

[109] *Sharma A., Pawar D. M.* Survey paper on topic modeling techniques to gain usefull forcasting information on violant extremist activities over cyber space // *International Journal of Advanced Research in Computer Science and Software Engineering*. — 2015. — Vol. 5, no. 12. — Pp. 429–436.

[110] *Shashanka M., Raj B., Smaragdis P.* Sparse overcomplete latent variable decomposition of counts data // Advances in Neural Information Processing Systems, NIPS-2007 / Ed. by J. C. Platt, D. Koller, Y. Singer, S. Roweis. — Cambridge, MA: MIT Press, 2008. — Pp. 1313–1320.

[111] *Shivashankar S., Srivathsan S., Ravindran B., Tendulkar A. V.* Multi-view methods for protein structure comparison using latent dirichlet allocation. // *Bioinformatics [ISMB/ECCB]*. — 2011. — Vol. 27, no. 13. — Pp. 61–68.

[112] *Shneiderman B.* The eyes have it: A task by data type taxonomy for information visualizations // Proceedings of the 1996 IEEE Symposium on Visual Languages. — VL'96. — Washington, DC, USA: IEEE Computer Society, 1996. — Pp. 336–343.

[113] *Si X., Sun M.* Tag-LDA for scalable real-time tag recommendation // *Journal of Information & Computational Science.* — 2009. — Vol. 6. — Pp. 23–31.

[114] *Singh R., Hsu Y.-W., Moon N.* Multiple perspective interactive search: a paradigm for exploratory search and information retrieval on the Web // *Multimedia Tools and Applications.* — 2013. — Vol. 62, no. 2. — Pp. 507–543.

[115] *Sokolov E., Bogolubsky L.* Topic models regularization and initialization for regression problems // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. — New York, NY, USA: ACM, 2015. — Pp. 21–27.

[116] *Steyvers M., Griffiths T.* Finding scientific topics // *Proceedings of the National Academy of Sciences.* — 2004. — Vol. 101, no. Suppl. 1. — Pp. 5228–5235.

[117] *Sun Y., Han J., Gao J., Yu Y.* iTopicModel: Information network-integrated topic modeling // 2009 Ninth IEEE International Conference on Data Mining. — 2009. — Pp. 493–502.

[118] *Tan Y., Ou Z.* Topic-weak-correlated latent Dirichlet allocation // 7th International Symposium Chinese Spoken Language Processing (ISCSLP). — 2010. — Pp. 224–228.

[119] *Teh Y. W., Jordan M. I., Beal M. J., Blei D. M.* Hierarchical Dirichlet processes // *Journal of the American Statistical Association.* — 2006. — Vol. 101, no. 476. — Pp. 1566–1581.

[120] *Teh Y. W., Newman D., Welling M.* A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation // NIPS. — 2006. — Pp. 1353–1360.

[121] TextFlow: Towards better understanding of evolving topics in text. / W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, X. Tong // *IEEE transactions on visualization and computer graphics.* — 2011. — Vol. 17, no. 12. — Pp. 2412–2421.

[122] *Varadarajan J., Emonet R., Odobez J.-M.* A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions. — 2010.

[123] *Varshney D., Kumar S., Gupta V.* Modeling information diffusion in social networks using latent topic information // Intelligent Computing Theory / Ed. by D.-S. Huang, V. Bevilacqua, P. Premaratne. — Springer International Publishing, 2014. — Vol. 8588 of *Lecture Notes in Computer Science.* — Pp. 137–148.

[124] *Veas E. E., di Sciascio C.* Interactive topic analysis with visual analytics and recommender systems // 2nd Workshop on Cognitive Computing and Applications for Augmented Human Intelligence, CCAAHI2015, International Joint Conference on Artificial Intelligence, IJCAI, Buenos Aires, Argentina, July 2015. — Aachen, Germany, Germany: CEUR-WS.org, 2015.

[125] *Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Yanina A.* Non-bayesian additive regularization for multimodal topic modeling of large collections // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. — New York, NY, USA: ACM, 2015. — Pp. 29–37.

[126] *Vorontsov K. V., Potapenko A. A.* Additive regularization of topic models // *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization.* — 2014.

[127] *Vorontsov K. V., Potapenko A. A.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // AIST'2014, Analysis of Images, Social networks and Texts. — Vol. 436. — Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2014. — Pp. 29–46.

[128] *Vorontsov K. V., Potapenko A. A.* Additive regularization of topic models // *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications.* — 2015. — Vol. 101, no. 1. — Pp. 303–323.

[129] *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). April 20-22, 2015. Royal Holloway, University of London, UK. / Ed. by A. G. et al. — Springer International Publishing Switzerland 2015, 2015. — Pp. 193–202.

[130] *Vulic I., De Smet W., Tang J., Moens M.-F.* Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications // *Information Processing & Management.* — 2015. — Vol. 51, no. 1. — Pp. 111–147.

[131] *Vulić I., Smet W., Moens M.-F.* Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora // *Information Retrieval.* — 2012. — Pp. 1–38.

[132] *Wallach H. M.* Topic modeling: Beyond bag-of-words // Proceedings of the 23rd International Conference on Machine Learning. — ICML '06. — New York, NY, USA: ACM, 2006. — Pp. 977–984.

[133] *Wang C., Blei D. M.* Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process // NIPS. — Curran Associates, Inc., 2009. — Pp. 1982–1989.

[134] *Wang C., Blei D. M.* Collaborative topic modeling for recommending scientific articles // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — New York, NY, USA: ACM, 2011. — Pp. 448–456.

[135] *Wang C., Danilevsky M., Desai N., Zhang Y., Nguyen P., Taula T., Han J.* A phrase mining framework for recursive construction of a topical hierarchy // Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '13. — New York, NY, USA: ACM, 2013. — Pp. 437–445.

[136] *Wang C., Liu J., Desai N., Danilevsky M., Han J.* Constructing topical hierarchies in heterogeneous information networks // *Knowledge and Information Systems.* — 2014. — Vol. 44, no. 3. — Pp. 529–558.

[137] *Wang C., Liu X., Song Y., Han J.* Scalable and robust construction of topical hierarchies // *CoRR.* — 2014. — Vol. abs/1403.3460.

[138] *Wang C., Liu X., Song Y., Han J.* Towards interactive construction of topical hierarchy: A recursive tensor decomposition approach // Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '15. — New York, NY, USA: ACM, 2015. — Pp. 1225–1234.

[139] *Wang H., Zhang D., Zhai C.* Structural topic model for latent topical structure analysis // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. — HLT '11. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. — Pp. 1526–1535.

[140] *Wang X., McCallum A.* Topics over time: A non-markov continuous-time model of topical trends // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '06. — New York, NY, USA: ACM, 2006. — Pp. 424–433.

[141] *Wang X., McCallum A., Wei X.* Topical n-grams: Phrase and topic discovery, with an application to information retrieval // Proceedings of the 2007 Seventh IEEE International Conference on Data Mining. — Washington, DC, USA: IEEE Computer Society, 2007. — Pp. 697–702.

[142] *White R. W., Roth R. A.* Exploratory Search: Beyond the Query-Response Paradigm. Synthesis Lectures on Information Concepts, Retrieval, and Services. — Morgan and Claypool Publishers, 2009.

[143] *Wu Y., Ding Y., Wang X., Xu J.* A comparative study of topic models for topic clustering of Chinese web news // Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on. — Vol. 5. — july 2010. — Pp. 236–240.

[144] *Yan X., Guo J., Lan Y., Cheng X.* A biterm topic model for short texts // Proceedings of the 22Nd International Conference on World Wide Web. — WWW '13. — Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. — Pp. 1445–1456.

[145] *Yanina A., Vorontsov K.* Multi-objective topic modeling for exploratory search in tech news // AINL. — 2016 (to appear).

[146] *Yeh J.-h., Wu M.-l.* Recommendation based on latent topics and social network analysis // Proceedings of the 2010 Second International Conference on Computer Engineering and Applications. — Vol. 1. — IEEE Computer Society, 2010. — Pp. 209–213.

[147] *Yi X., Allan J.* A comparative study of utilizing topic models for information retrieval // Advances in Information Retrieval. — Springer Berlin Heidelberg, 2009. — Vol. 5478 of *Lecture Notes in Computer Science*. — Pp. 29–41.

[148] *Yin H., Cui B., Chen L., Hu Z., Zhang C.* Modeling location-based user rating profiles for personalized recommendation // *ACM Transactions of Knowledge Discovery from Data*. — 2015.

[149] *Yin H., Cui B., Sun Y., Hu Z., Chen L.* LCARS: A spatial item recommender system // *ACM Transaction on Information Systems*. — 2014.

[150] *Yin Z., Cao L., Han J., Zhai C., Huang T.* Geographical topic discovery and comparison // Proceedings of the 20th international conference on World wide web / ACM. — 2011. — Pp. 247–256.

[151] *Zavitsanos E., Paliouras G., Vouros G. A.* Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes // *Journal of Machine Learning Research*. — 2011. — Vol. 12. — Pp. 2749–2775.

[152] *Zhang J., Song Y., Zhang C., Liu S.* Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2010. — Pp. 1079–1088.

[153] *Zhao W. X., Jiang J., Weng J., He J., Lim E.-P., Yan H., Li X.* Comparing Twitter and traditional media using topic models // Proceedings of the 33rd European Conference on Advances in Information Retrieval. — ECIR'11. — Berlin, Heidelberg: Springer-Verlag, 2011. — Pp. 338–349.

[154] *Zhao X. W., Wang J., He Y., Nie J.-Y., Li X.* Originator or propagator?: Incorporating social role theory into topic models for Twitter content analysis // Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management. — CIKM '13. — New York, NY, USA: ACM, 2013. — Pp. 1649–1654.

[155] *Zhou S., Li K., Liu Y.* Text categorization based on topic model // *International Journal of Computational Intelligence Systems*. — 2009. — Vol. 2, no. 4. — Pp. 398–409.

[156] *Zuo Y., Zhao J., Xu K.* Word network topic model: A simple but general solution for short and imbalanced texts // *Knowledge and Information Systems*. — 2016. — Vol. 48, no. 2. — Pp. 379–398.