Вероятностные тематические модели
Лекция 5. Оценивание качества тематических моделей

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

МФТИ – ФИЦ ИУ РАН • 5 октября 2023

Содержание

Задача тематического моделирования

Дано: коллекция текстовых документов, $p(w|d) = \frac{n_{dw}}{n_d}$

Найти: матрицы параметров $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ вероятностной тематической модели

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

Критерий: максимум регуляризованного правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}; \quad R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$$

Задача BTM по природе своей многокритериальная:

- критерии регуляризации гладкие для удобства оптимизации
- критерии для измерения различных аспектов качества модели более интерпретируемые, но не всегда гладкие

Aly Abdelrazeka et al. Topic modeling algorithms and applications: a survey, 2022.
Caitlin Doogan, Wray Buntine. Topic model or topic twaddle? Re-evaluating semantic interpretability measures. 2021.

Критерии (метрики, меры) качества тематических моделей

Внешние критерии используют внешние данные

- Полнота и точность тематического поиска
- Качество ранжирования при тематическом поиске
- Качество решения прикладной задачи: классификации, категоризации, суммаризации, сегментации и т.п.
- Экспертные оценки качества (интерпретируемости) тем

Внутренние критерии используют только матрицы Φ и Θ

- Правдоподобие и перплексия
- Различные косвенные меры интерпретируемости:
  — когерентность (согласованность) тем,
  — разреженность матриц Φ и Θ,
  — различность, чистота, контрастность тем,
  — объём семантических ядер тем, невырожденность тем
- Статистический тест условной независимости

Правдоподобие и перплексия (perplexity)

Правдоподобие языковой модели $p(w|d)$ (чем выше, тем лучше):

$$\mathscr{L}(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d), \qquad p(w|d) = \sum_t \phi_{wt}\theta_{td}$$

Перплексия языковой модели $p(w|d)$ (чем меньше, тем лучше):

$$\mathcal{P}(D) = \exp\Big(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d)\Big), \quad n = \sum_{d\in D}\sum_{w\in d} n_{dw}$$

Интерпретация перплексии:

- если распределение $p(w|d) = \frac{1}{|W|}$ равномерное, то $\mathcal{P} = |W|$
- мера «удивлённости» модели словам текста
- коэффициент ветвления (branching factor) текста
- известные оценки человеческой перплексии: 8–12

Перплексия тестовой (отложенной) коллекции

Проблема: перплексия может быть оптимистично занижена из-за *эффекта переобучения*.

Перплексия тестовой коллекции D' (hold-out perplexity):

$$\mathcal{P}(D') = \exp\Big(-\frac{1}{n''}\sum_{d\in D'}\sum_{w\in d''} n_{dw}\ln p(w|d)\Big), \quad n'' = \sum_{d\in D'}\sum_{w\in d''} n_{dw}$$

$d = d' \sqcup d''$ — случайное разбиение тестового документа на две половины равной длины;

параметры ϕ_{wt} оцениваются по обучающей коллекции D;
параметры θ_{td} оцениваются по первой половине d';
перплексия вычисляется по второй половине d''.

Проблема: как разбивать документ на две половины?

Измерение интерпретируемости тем

Тема интерпретируемая, если по топовым словам темы эксперт может определить, о чём эта тема, и дать ей название.

- *Экспертные оценки*:
  — интерпретируемость темы по балльной шкале;
  — каждую тему оценивают несколько экспертов.
- *Метод интрузий* (intrusion):
  — в список топовых слов внедряется лишнее слово;
  — измеряется доля ошибок экспертов при его определении

Задача: найти внутренний критерий интерпретируемости, наиболее коррелирующий с экспертными оценками

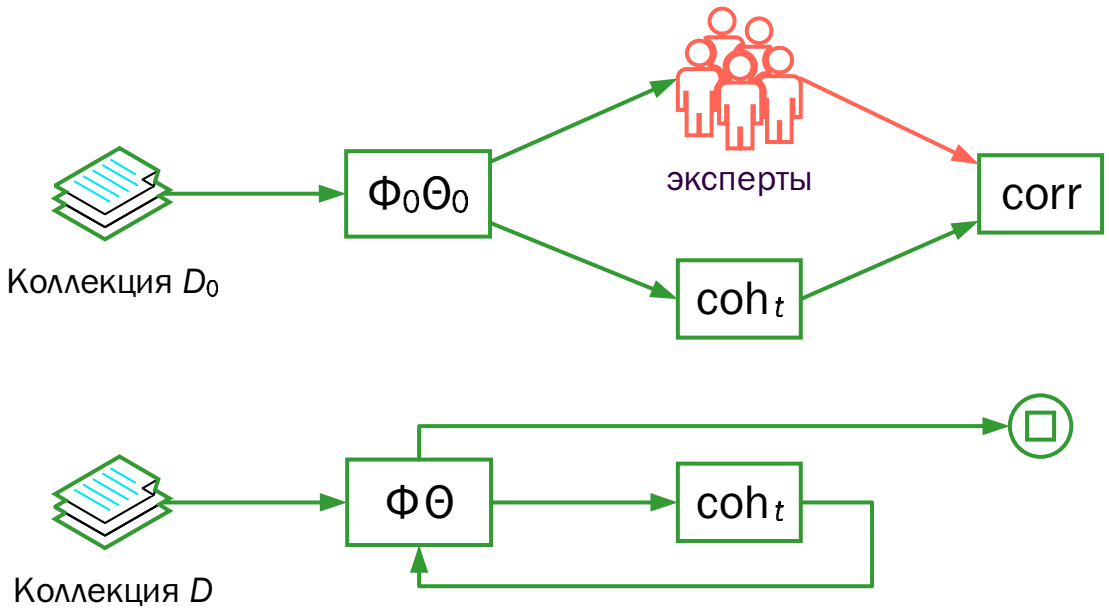
Решение: *когерентность* (согласованность) тем (topic coherence)

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Схема эксперимента Ньюмана

1. берём коллекцию D_0 для калибровки внутреннего критерия
2. строим тематическую модель $\Phi_0\Theta_0$
3. эксперты оценивают темы (рейтингами или интрузиями)
4. ищем критерий, коррелирующий с оценками экспертов

На новой коллекции D используем откалиброванный критерий (когерентность тем coh_t) для оценивания и выбора моделей $\Phi\Theta$



Эксперимент. Связь когерентности и интерпретируемости

Измерялась ранговая корреляция Спирмена каждой из 15 метрик и экспертными оценками интерпретируемости.

PMI — лучшая метрика.

Gold-standard — средняя корреляция Спирмена между оценками разных экспертов.

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	−0.20	0.19
	LCH	−0.31	−0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	−0.08	0.27
	WUP	0.41	0.26
Wikipedia	RACO	0.62	0.69
	MIW	0.68	0.70
	DOCSIM	0.59	0.60
	PMI	**0.74**	**0.77**
Google	TITLES	0.51	
	LOGHITS	−0.19	
Gold-standard	IAA	0.82	0.78

Вывод: когерентность близка к «золотому стандарту».

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Когерентность как внутренний критерий интерпретируемости

Когерентность (согласованность) темы t по k топовым словам:

$$\text{coh}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{PMI}(w_i, w_j)$$

где w_i — i-е слово в порядке убывания ϕ_{wt},

$\text{PMI}(u,v) = \ln \frac{P_{uv}}{P_u P_v}$ — *поточечная взаимная информация* (pointwise mutual information),

P_{uv} — доля документов, в которых слова u, v хотя бы один раз встречаются рядом (в одном предложении или в окне 10 слов),

P_u — доля документов, в которых u встретился хотя бы 1 раз,

P_{uv}, P_u можно вычислять по другой коллекции (Википедии).

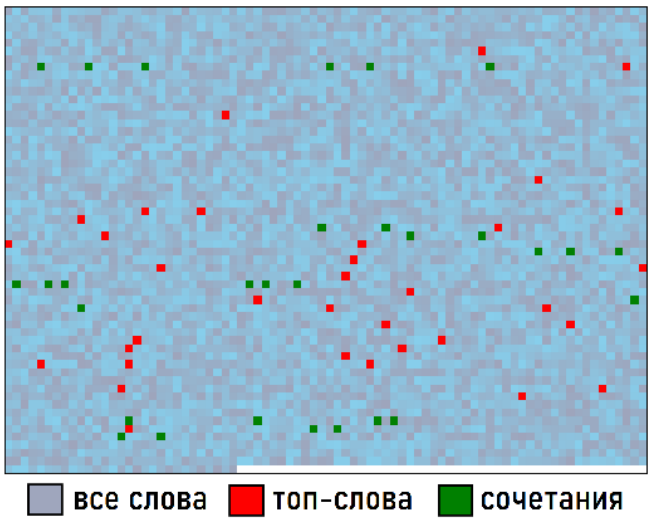
Когерентность модели = средняя когерентность всех тем.

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Недостаток когерентности

Обычно берут $k = 10..20$ топовых (наиболее частотных) слов, но они занимают лишь 1–2% текста совместно по всем темам, а пары с большим N_{uv} образуются из топовых слов ещё реже!

Более 99% текста игнорируется оценкой когерентности модели, и «золотой стандарт» Ньюмана страдает тем же недостатком!



Напротив, если предположить существование суперсимметрии, то введение новых **частиц** приводит как раз к такому объединению. Оказывается, что суперсимметрия не только обеспечивает объединение взаимодействий, но и стабилизирует объединённую теорию, в которой присутствуют два совершенно разных масштаба: масштаб масс обычных **частиц** (порядка 100 масс протона) и масштаб великого объединения (порядка 10^{16} масс протона). Последний масштаб уже близок к так называемому планковскому масштабу, равному обратной ньютоновской константе тяготения, что составляет порядка 10^{19} масс протона. На этом масштабе мы ожидаем проявление эффектов квантовой гравитации. В этом моменте нас ожидает приятный сюрприз. Дело в том, что гравитация всегда стояла несколько особняком по отношению к остальным взаимодействиям. Переносчик гравитации, гравитон, имеет спин 2, в то время как переносчики остальных взаимодействий имеют спин 1. Однако суперсимметрия перемешивает спины.

first top words of *topic 3: физика* with top 10 in bold: **частица**, **электрон**, **кварк**, **атом**, **энергия**, **вселенная**, **фотон**, **физика**, **физик**, **эксперимент**, масса, теория, свет, симметрия, протон, эйнштейн, нейтрино, вещество, квантовый, ускоритель, детектор, волна, эффект, свойство, спин, гравитация, материя, адрон, поль, частота

V.A.Alekseev, V.G.Bulatov, K.V.Vorontsov. Intra-text coherence as a measure of topic models interpretability // Dialogue, 2018.

Обобщение — семейство средневзвешенных когерентностей

Средневзвешенная когерентность темы:

$$\mathrm{coh}_t = \frac{\sum_{u,v} \mathrm{rel}_t(u,v)\,\mathrm{coh}(u,v)}{\sum_{u,v} \mathrm{rel}_t(u,v)},$$

$\mathrm{coh}(u,v)$ — *сочетаемость* пары слов u, v в текстах,
$\mathrm{rel}_t(u,v)$ — *релевантность* слов u и v теме t, в частности,
$\mathrm{rel}_t(u,v) = \big[\phi_{ut}, \phi_{vt} > \mathrm{top}_k \phi_{wt}\big]$ — *когерентность Ньюмана*

Возможные модификации:

- сделать rel ненулевым для большего числа пар u, v:
  $\mathrm{rel}_t(u,v) = \phi_{ut} + \phi_{vt}$ или $\sqrt{\phi_{ut}\phi_{vt}}$ или $[\phi_{ut}\phi_{vt} \geqslant \varepsilon]$
- можно поэкспериментировать также с выбором coh:
  $\mathrm{coh}(u,v) = (\mathrm{PMI} - \delta)_+$ или $\mu\big(\frac{P_{uv}}{P_u P_v}\big)$ или $\frac{P_{uv} - P_u P_v}{\sqrt{P_{uv}}}$

Проблема: большой объём вычислений по всем парам слов

Внутритекстовая когерентность

Средневзвешенная когерентность темы:

$$\mathrm{coh}_t = \frac{\sum_{u,v} \mathrm{rel}_t(u,v)\,\mathrm{coh}(u,v)}{\sum_{u,v} \mathrm{rel}_t(u,v)},$$

но теперь суммирование не по парам слов словаря $(u,v) \in W^2$, а по парам слов, находящихся в общих *контекстах*, например, в одном предложении или на расстоянии не более 10 слов.

Теперь все $\mathrm{rel}_t(u,v)$ можно брать ненулевыми.

Новая возможность: $\mathrm{rel}_t(u,v) = \sqrt{p(t|d,u)\,p(t|d,v)}$.

Вычисление: за один проход по коллекции для каждой темы t аккумулируются суммы в числителе и в знаменателе.

Василий Алексеев. Внутритекстовая когерентность как мера интерпретируемости тематических моделей текстовых коллекций. МФТИ, 2018.

Как проверить адекватность внутритекстовой когерентности

. . . если «золотой стандарт» Ньюмана столь же неадекватен?

Идея:

- эксперты размечают в текстах *тематические цепочки слов*
- тексты — научно-популярные, междисциплинарные

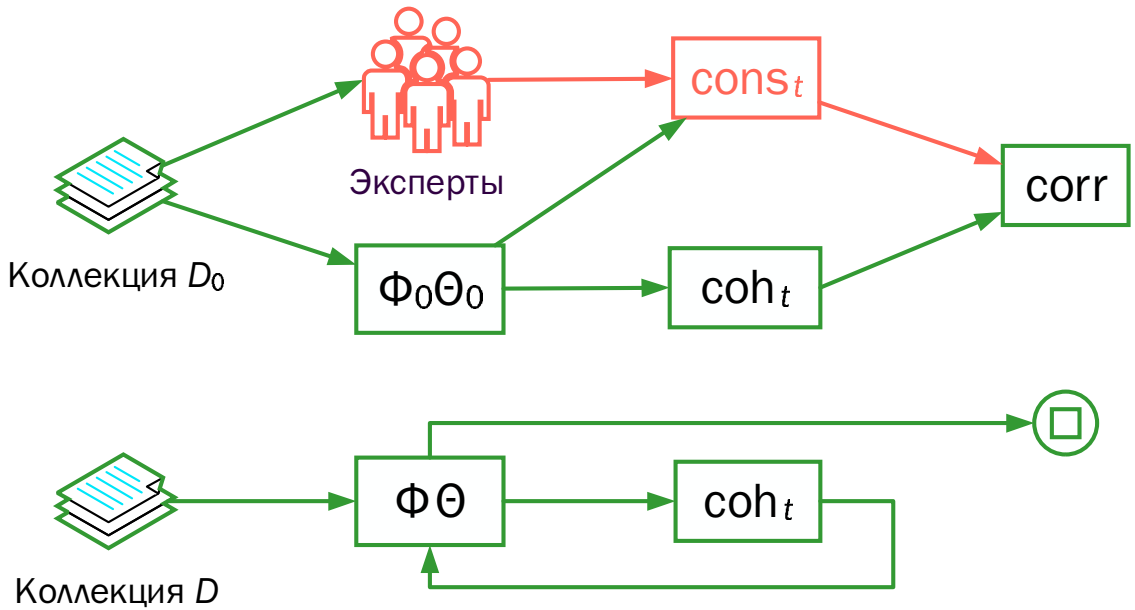
Пример разметки:

транспорт	психология	общенаучная лексика	общеупотребительная лексика

В исследованиях мы действительно можем находить корреляции между стилем вождения и особенностями личности. Например, склонные к экстраверсии водители могут больше отвлекаться на внешние факторы и стимулы внешней среды и в этом отношении представляют большую опасность. В свою очередь, люди, которым требуется большее количество психических ресурсов, для того чтобы справиться с тревогой, будут вести себя осторожнее в условиях трафика. Вместе с тем есть и обратная сторона: та же характеристика интроверсии за счет высокого уровня тревожности приводит к чрезмерной осторожности. Для таких водителей характерен крадущийся тип вождения, что будет влиять на общее тревожное поведение всех участников трафика.

Схема калибровки внутритекстовой когерентности

1. выбираем из коллекции D_0 фрагменты для разметки
2. эксперты размечают тематические цепочки во фрагментах
3. строим тематическую модель $\Phi_0\Theta_0$ (или несколько разных)
4. ищем критерий, коррелирующий с согласованностью cons_t между темами t и размеченными тематическими цепочками



Мера согласованности темы с размеченными цепочками

C_{di} — i-я цепочка в размеченном фрагменте d

Тематика цепочки C как подмножества слов:

$$p(t|C) = \sum_{w \in C} p(t|w)p(w|C) = \operatorname*{mean}_{w \in C} p(t|w),$$

где $p(t|w) = p(w|t)\frac{p(t)}{p(w)} = \phi_{wt}\frac{n_t}{n_w}$ (по формуле Байеса)

Множество цепочек, *согласованных* (consistent) с темой t:

$$C(t) = \left\{ C_{di} \colon t = \arg\max_t p(t|C_{di}) \right\}$$

Мера согласованности темы с размеченными цепочками:

$$\text{cons}_t = \operatorname*{mean}_{C_{di} \in C(t)} p(t|C_{di})$$

Различимость цепочек во фрагментах d, должна быть близка к 1:

$$\text{diff} = \frac{\sum_d \#\{t \colon C_{di} \in C(t)\}}{\sum_d \#\{C_{di}\}}$$

Критерии разреженности матриц Φ и Θ

Разреженность — доля нулевых элементов в Φ и Θ

Однако ϕ_{wt} и θ_{td} не всегда разреживаются до нуля

- *Доля существенных слов* в темах (Word Ratio):

$$\mathrm{WR}_t = \frac{1}{|W|} \sum_{w \in W} \left[\phi_{wt} > \frac{1}{|W|}\right] \qquad \mathrm{WR} = \frac{1}{|T|} \sum_{t \in T} \mathrm{WR}_t$$

- *Доля существенных тем* в документах (Document Ratio):

$$\mathrm{DR}_d = \frac{1}{|T|} \sum_{t \in T} \left[\theta_{td} > \frac{1}{|T|}\right] \qquad \mathrm{DR} = \frac{1}{|D|} \sum_{d \in D} \mathrm{DR}_d$$

Естественная разреженность матриц Φ и Θ в экспериментах:

- $\mathrm{WR} = 3.5\%$, $\mathrm{DR} = 11.5\%$
- Если оставить слова w: $\phi_{wt} > \frac{1}{|W|}$ хотя бы в одной теме, то *сокращение словаря* (vocabulary reduction): 154 K $\to$ 8 K

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Лексическое ядро, чистота и контрастность темы

Лексическое ядро W_t *темы* t, варианты определения:

- W_t — top-k термов с наибольшими значениями $p(w|t)$
- $W_t = \{w : p(w|t) > p(w)\}$
- $W_t = \{w : p(w|t) > \frac{1}{|W|}\}$ [Кольцов и др., 2014]
- $W_t = \{w : p(t|w) > 0.25\}$ [Воронцов, Потапенко, 2014]

Характеристики лексического ядра темы:

- $|W_t|$ — *размер ядра* темы, ориентировочно $|W_t| \sim \frac{|W|}{|T|}$
- $\sum_{w \in W_t} p(w|t)$ — *чистота* темы, из $[0,1]$, лучше больше
- $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$ — *контрастность* темы, $[0,1]$, лучше больше
- $\frac{1}{|W_t|} \sum_{w \in W_t} \log \frac{p(w|t)}{p(w)}$ — logLift, лучше больше [Taddy, 2012]

Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST, 2014.

Критерии различности тем

Среднее расстояние от темы t до ближайшей к ней темы

$$\text{minDist}_t = \min_{s \in T \setminus t} \rho(\phi_t, \phi_s) \qquad \text{minDist} = \frac{1}{|T|} \sum_{t \in T} \text{minDist}_t$$

Расстояния между вероятностными распределениями (от 0 до 1):

- $\rho(\phi_t, \phi_s) = 1 - \frac{\sum_w \phi_{ws}\phi_{wt}}{\left(\sum_w \phi_{ws}^2\right)^{1/2}\left(\sum_w \phi_{wt}^2\right)^{1/2}}$ — косинусное
- $\rho(\phi_t, \phi_s) = 1 - |W_t \cap W_s| : |W_t \cup W_s|$ — Жаккара
- $\rho^2(\phi_t, \phi_s) = \frac{1}{2}\sum_w \left(\sqrt{\phi_{ws}} - \sqrt{\phi_{wt}}\right)^2$ — Хеллингера

Дивергенции — несимметричные меры «вложенности» ϕ_t в ϕ_s:

- $\rho(\phi_t, \phi_s) = \sum_w \phi_{wt} \ln\left(\frac{\phi_{wt}}{\phi_{ws}}\right)$ — Кульбака–Лейблера
- $\rho(\phi_t, \phi_s) = \frac{1}{\lambda(\lambda+1)} \sum_w \phi_{wt}\left(\left(\frac{\phi_{wt}}{\phi_{ws}}\right)^\lambda - 1\right)$ — Кресси–Рида

Критерии вырожденности тематической модели

Тематичность терма (чем выше кросс-энтропия, тем тематичнее):

$$H(w) = -\sum_{t \in T} p(t) \ln p(t|w)$$

Доля нетематичных термов:

- $\frac{1}{|W|} \sum_w \left[H(w) < H_0 \right]$ — в словаре W
- $\frac{1}{n_d} \sum_w n_{dw} \left[H(w) < H_0 \right]$ — в документе d
- $\frac{1}{n} \sum_d \sum_w n_{dw} \left[H(w) < H_0 \right]$ — в коллекции D

Доля фоновых термов (при сглаживании фоновых тем $B \subset T$):

- $\frac{1}{|W|} \sum_w \sum_{t \in B} p(t|w)$ — в словаре W
- $\sum_{t \in B} p(t|d)$ — в документе d
- $\frac{1}{n} \sum_d n_d \sum_{t \in B} p(t|d)$ — в коллекции D

Напоминание. Регуляризаторы сглаживания и разреживания

Общий вид регуляризаторов сглаживания и разреживания:

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

где $\beta_0 > 0$, $\alpha_0 > 0$ — коэффициенты регуляризации,
β_{wt}, α_{td} — параметры, задаваемые пользователем:

- $\beta_{wt} > 0$, $\alpha_{td} > 0$ — сглаживание
- $\beta_{wt} < 0$, $\alpha_{td} < 0$ — разреживание
- $\beta_{wt} > -1$, $\alpha_{td} > -1$ — модель LDA

Возможные применения сглаживания и разреживания:

- задать фоновые темы с общей лексикой языка
- задать шумовую тему для нетематичных термов
- задать псевдо-документ с ключевыми термами темы
- скорректировать состав термов и документов темы

Напоминание. Регуляризатор декоррелирования тем

Цель: сделать темы как можно более различными, выделить для каждой темы *лексическое ядро* — набор термов, отличающий её от других тем.

Минимизируем ковариации между вектор-столбцами ϕ_t:

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

Подставляем, получаем ещё один вариант разреживания — постепенное контрастирование строк матрицы Φ (малые вероятности ϕ_{wt} в строке становятся ещё меньше):

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

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation. 2010.

Напоминание. Разреживающий регуляризатор для отбора тем

Цель: избавиться от незначимых тем (topic selection).

Разреживаем распределение $p(t) = \sum_d p(d)\theta_{td}$, максимизируя кросс-энтропию между $p(t)$ и равномерным распределением:

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

Подставляем, получаем:

$$\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} - \tau \frac{n_d}{n_t}\theta_{td}\right), \text{ вариант: } \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td}\left(1 - \frac{\tau}{n_t}\right)\right).$$

Эффект: обнуляются строки матрицы Θ с малыми n_t, заодно (неожиданно) удаляются зависимые и расщеплённые темы.

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization. SLDS 2015.

Разреживание, сглаживание, декоррелирование, отбор тем

M-шаг при комбинировании 6 регуляризаторов:

$$\phi_{wt} = \underset{w}{\text{norm}}\Big(n_{wt} + \tau_1 \underbrace{\beta_w[t \in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \tau_2 \underbrace{\beta_w[t \in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} - \tau_3 \underbrace{\phi_{wt} \sum_{s \in S \setminus t} \phi_{ws}}_{\text{декоррелирование}} \Big)$$

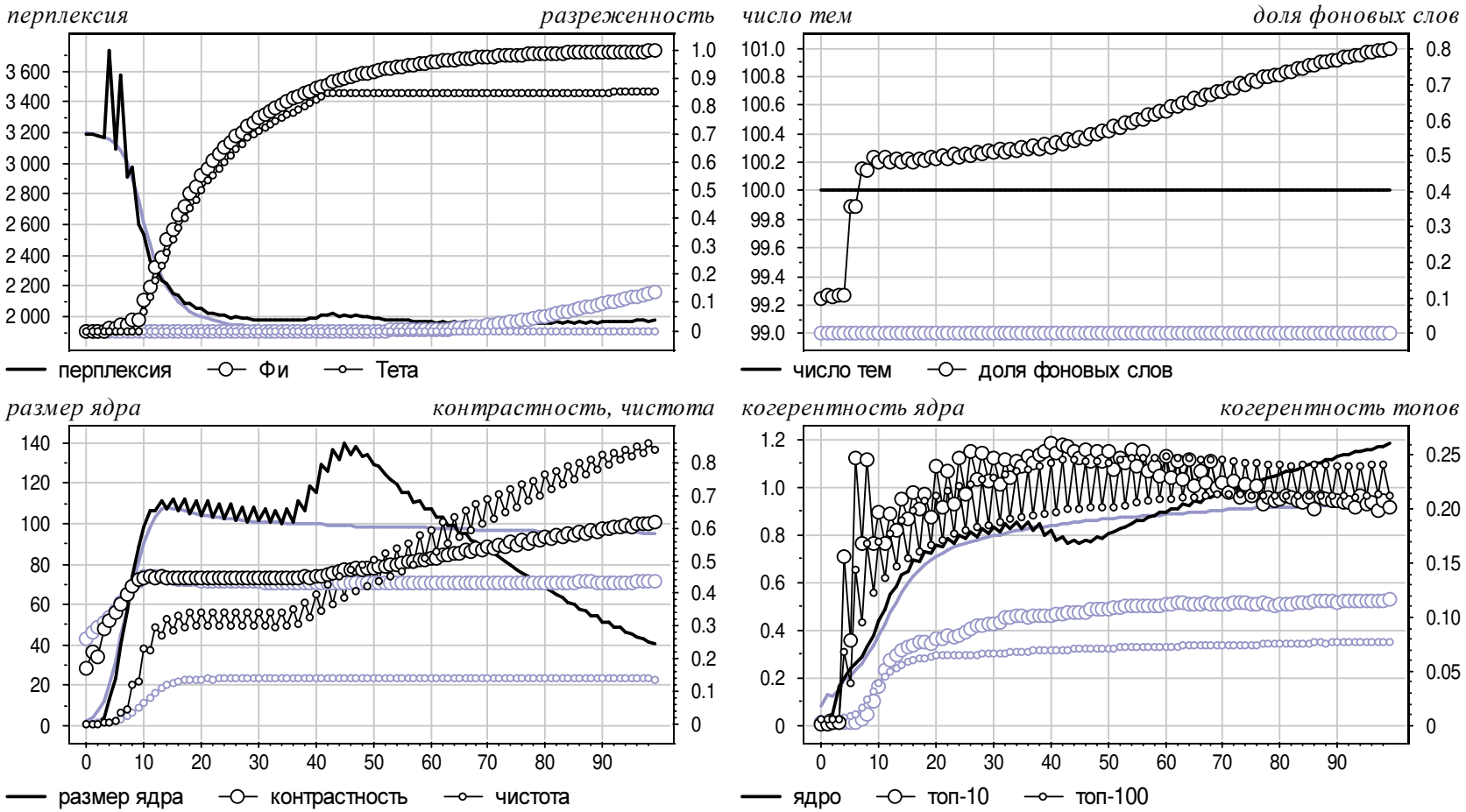
$$\theta_{td} = \underset{t}{\text{norm}}\Big(n_{td} + \tau_4 \underbrace{\alpha_t[t \in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \tau_5 \underbrace{\alpha_t[t \in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} - \tau_6 \underbrace{\frac{n_d}{n_t}\theta_{td}}_{\substack{\text{удаление}\\ \text{малых тем}}} \Big)$$

Данные: статьи NIPS (Neural Information Processing System)
$|D| = 1566$ статей, $n = 2.3\,\text{M}$, $|W| = 13\,\text{K}$,
контрольная коллекция: $|D'| = 174$.

Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST'2014.

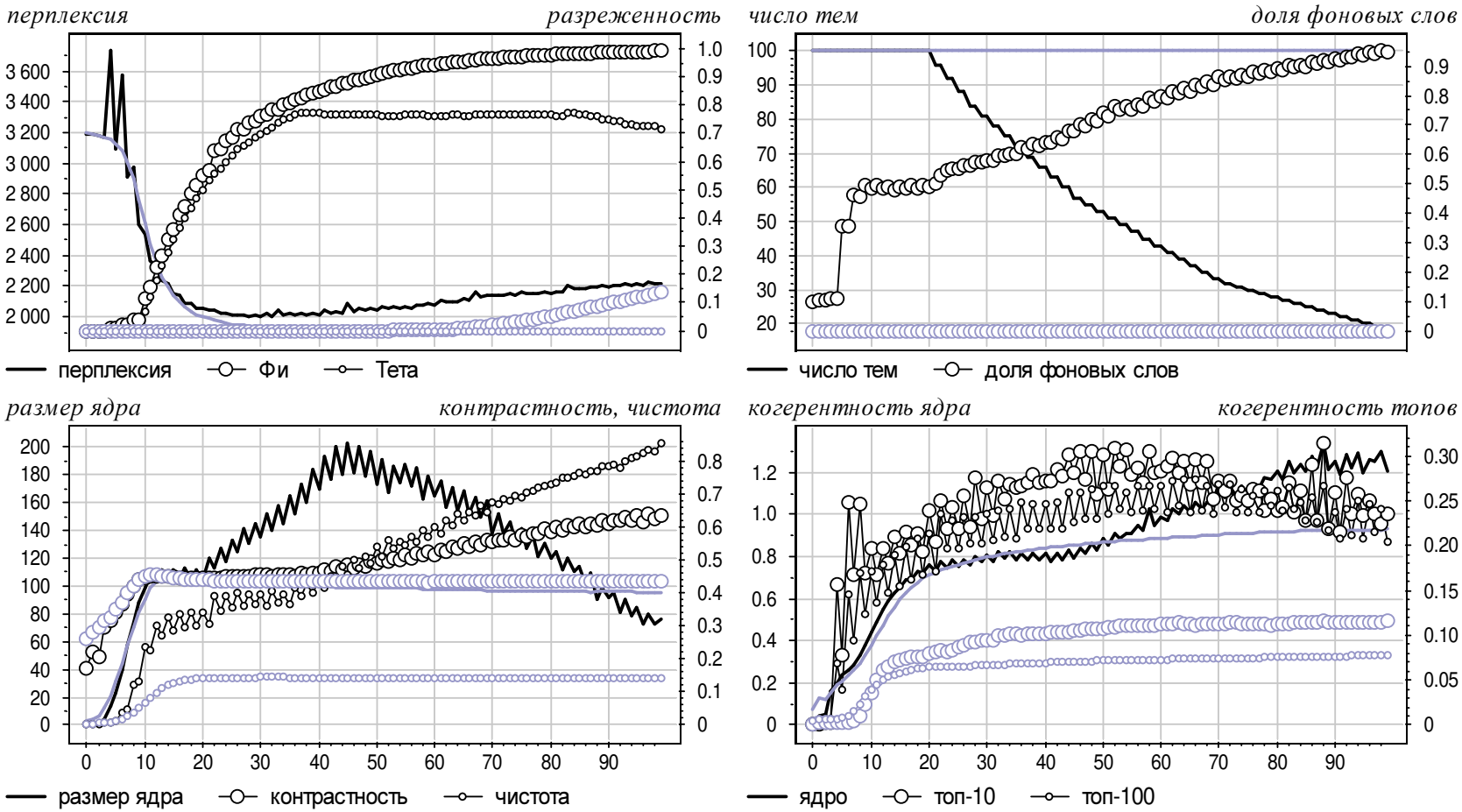
Разреживание, сглаживание, декоррелирование

Зависимости критериев качества от итераций EM-алгоритма (серый — PLSA, чёрный — ARTM)



Те же регуляризаторы, плюс отбор тем

Зависимости критериев качества от итераций EM-алгоритма (серый — PLSA, чёрный — ARTM)



Выводы по результатам экспериментов

Одновременное улучшение многих критериев качества
при незначительной деградации *перплексии* (правдоподобия):

- *разреженность* выросла от 0 до 95%–98%
- *когерентность тем* выросла от 0.1 до 0.3
- *чистота тем* выросла от 0.15 до 0.8
- *контрастность тем* выросла от 0.4 до 0.6

Рекомендации по выбору *траектории регуляризации*:

- разреживание включать постепенно после 10-20 итераций
- сглаживание включать сразу
- декоррелирование включать сразу и как можно сильнее
- отбор тем включать постепенно,
- не совмещая с декоррелированием на одной итерации

Постановка задачи планирования экспериментов (AutoML)

$\Lambda = (\lambda_1, \ldots, \lambda_K)$ — вектор гиперпараметров ($|T|$, τ_i, τ_m и др.)
$Q = (q_1, \ldots, q_M)$ — вектор критериев качества модели

Гипотезы:

1. существует регрессионная зависимость $Q(\Lambda, D)$
2. Q_0 — область допустимых решений в пространстве Q
3. Q_* — область оптимальных решений в пространстве Q
4. $Q(\Lambda, D') \approx Q(\Lambda, D)$ — устойчивость к подвыборкам $D' \subset D$

Задача: построить итерационный процесс $\Lambda_{k+1} := f(\Lambda_k)$, который за минимальное время приводит к точке (Λ, Q), не выходящей за пределы Q_0 и достаточно близкой к Q_*

А.Кузьмин. Адаптивный выбор траектории регуляризации. МФТИ, 2017.

М.Ходорченко. Эволюционные методы оптимизации для автоматической настройки гиперпараметров тематических моделей с аддитивной регуляризацией. 2022.

Гипотеза условной независимости

$$\left.\begin{aligned} p(w|d,t) &= p(w|t) \\ p(d|w,t) &= p(d|t) \\ p(w,d|t) &= p(w|t)\,p(d|t) \end{aligned}\right\} \text{три эквивалентных представления}$$

Гипотеза *семантической однородности темы t*
— в теме t термы и документы порождаются независимо:

$$H_0(t): \quad \hat{p}(w,d|t) \sim p(w|t)\,p(d|t)$$

Гипотеза *согласованности документа d с темой t*
— термы темы t порождаются независимо от документов:

$$H_0(t,d): \quad \hat{p}(w|d,t) \sim p(w|t)$$

Гипотеза *согласованности терма w с темой t*
— тема t распределена по документам независимо от термов:

$$H_0(t,w): \quad \hat{p}(d|w,t) \sim p(d|t)$$

Мера семантической неоднородности темы t в коллекции

Статистика для проверки гипотезы $H_0(t)$:

$$S_t = \mathrm{KL}\big(\hat{p}(w,d|t) \,\big\|\, p(w|t)p(d|t)\big) = \sum_{d,w} \hat{p}(w,d|t) \ln \frac{\hat{p}(w,d|t)}{p(w|t)p(d|t)}$$

По определению условной вероятности и формуле E-шага:

$$\frac{\hat{p}(w,d|t)}{p(w|t)\,p(d|t)} = \frac{p(t|d,w)\,\hat{p}(w|d)\frac{p(d)}{p(t)}}{p(w|t)\,p(t|d)\frac{p(d)}{p(t)}} = \frac{p_{tdw}}{\phi_{wt}\theta_{td}}\hat{p}(w|d) = \frac{\hat{p}(w|d)}{p(w|d)}$$

$$S_t = \sum_{d\in D}\sum_{w\in d} \frac{n_{tdw}}{n_t} \ln \frac{\hat{p}(w|d)}{p(w|d)} = \operatorname*{avg}_{d,w}\left(n_{tdw}, \ln \frac{\hat{p}(w|d)}{p(w|d)}\right),$$

где $\operatorname*{avg}_{i\in I}(\gamma_i, x_i) = \dfrac{\sum_{i\in I}\gamma_i x_i}{\sum_{i\in I}\gamma_i}$ — *средневзвешенное x_i с весами γ_i*

Mimno D., Blei D. Bayesian checking for topic models. 2011.

Мера несогласованности документа d с темой t

Статистика для проверки гипотезы $H_0(d,t)$:

$$S_{td} = \mathrm{KL}\big(\hat{p}(w|d,t) \,\big\|\, p(w|t)\big) = \sum_{w\in d} \hat{p}(w|d,t) \ln \frac{\hat{p}(w|d,t)}{p(w|t)}$$

По определению условной вероятности и формуле E-шага:

$$\frac{\hat{p}(w|d,t)}{p(w|t)} = \frac{p(t|d,w)\,\hat{p}(w|d)\,\cancel{p(d)}}{p(w|t)\,p(t|d)\,\cancel{p(d)}} = \frac{p_{tdw}}{\phi_{wt}\theta_{td}}\hat{p}(w|d) = \frac{\hat{p}(w|d)}{p(w|d)}$$

$$S_{td} = \sum_{w\in d} \frac{n_{tdw}}{n_{td}} \ln \frac{\hat{p}(w|d)}{p(w|d)} = \operatorname*{avg}_{w\in d}\left(n_{tdw}, \ln \frac{\hat{p}(w|d)}{p(w|d)}\right)$$

Возможные применения меры несогласованности S_{td}:

- выделение документов, наиболее релевантных теме
- выявление нетематизируемых «грязных» документов
- ранняя остановка итераций по документу

Мера несогласованности терма w с темой t

Статистика для проверки гипотезы $H_0(w,t)$:

$$S_{wt} = \mathrm{KL}\big(\hat{p}(d|w,t) \,\big\|\, p(d|t)\big) = \sum_{d\in D} \hat{p}(d|w,t) \ln \frac{\hat{p}(d|w,t)}{p(d|t)}$$

По определению условной вероятности и формуле E-шага:

$$\frac{\hat{p}(d|w,t)}{p(d|t)} = \frac{p(t|d,w)\,\hat{p}(w|d)\,\cancel{p(d)}}{p(w|t)\,\cancel{p(t)}\,p(t|d)\frac{\cancel{p(d)}}{\cancel{p(t)}}} = \frac{p_{tdw}}{\phi_{wt}\theta_{td}}\hat{p}(w|d) = \frac{\hat{p}(w|d)}{p(w|d)}$$

$$S_{wt} = \sum_{d\in D} \frac{n_{tdw}}{n_{wt}} \ln \frac{\hat{p}(w|d)}{p(w|d)} = \operatorname*{avg}_{d\in D}\left(n_{tdw}, \ln \frac{\hat{p}(w|d)}{p(w|d)}\right)$$

Возможные применения меры несогласованности S_{wt}:

- выделение семантического ядра темы
- выделение термов общеупотребительной лексики
- формирование начальных приближений новых тем

Средневзвешенные статистики с произвольной функцией потерь

При $\ell(d,w) = \ln \frac{\hat{p}(w|d)}{p(w|d)}$ — рассмотренные выше *KL-статистики*:

$S_t = \operatorname{avg}_{d,w}\big(n_{tdw}, \ell(d,w)\big)$ — неоднородность темы в коллекции

$S_{td} = \operatorname{avg}_{w\in d}\big(n_{tdw}, \ell(d,w)\big)$ — несогласованность документа с темой

$S_{wt} = \operatorname{avg}_{d\in D}\big(n_{tdw}, \ell(d,w)\big)$ — несогласованность терма с темой

При $\ell(d,w) = \ln \frac{1}{p(w|d)}$ — *перплексия* (чем меньше, тем лучше):

$\ln \mathscr{P} = \operatorname{avg}_{d,w,t}\big(n_{tdw}, \ell(d,w)\big) = \operatorname{avg}_{d,w}\big(n_{dw}, \ell(d,w)\big)$ — коллекции

$\ln \mathscr{P}_d = \operatorname{avg}_{w,t}\big(n_{tdw}, \ell(d,w)\big) = \operatorname{avg}_{w\in d}\big(n_{dw}, \ell(d,w)\big)$ — документа

$\ln \mathscr{P}_t = \operatorname{avg}_{d,w}\big(n_{tdw}, \ell(d,w)\big)$ — темы t

$\ln \mathscr{P}_{td} = \operatorname{avg}_{w\in d}\big(n_{tdw}, \ell(d,w)\big)$ — темы t в документе d

Функции потерь, ослабляющие мощность стат. критерия

Условная независимость — избыточно сильное предположение:

- в каждом документе может использоваться лишь часть аспектов темы и, соответственно, лишь часть слов темы
- явление *повторяемости слов* (word burstiness):
  если слово встретилось в тексте один раз,
  то оно с большой вероятностью встретится ещё

Статистики S_t, S_{td}, S_{wt}, толерантные к повторяемости слов:

- игнорирование частот термов: замена $n_{dw} \to 1$, $n_{tdw} \to p_{tdw}$
- бинарная функция потерь $\ell(d,w) = \left[p(w|d) < \frac{\alpha}{n_d}\right]$
  с параметром $\alpha \approx 1$

Тогда средневзвешенные статистики $S_t, S_{td}, S_{wt} \in [0,1]$ выражают долю термов темы t, для которых модель предсказывает слишком малую вероятность.

Doyle G., Elkan C. Accounting for burstiness in topic models. 2009.

Применения оценок семантической однородности

Аномально высокие значения статистик:

- Определение перемешанных тем для расщепления
- Определение общеупотребительных слов в темах
- Определение плохо тематизируемых документов
- Распознавание наличия новой темы в документе
- Выделение термов для инициализации новой темы

Аномально низкие значения статистик:

- Выделение термов лексического ядра темы
- Выделение наиболее тематичных фраз/документов темы
- Выделение термов шаблонных фраз в темах

Нормальные значения статистик:

- Определение числа тем в коллекции
- Подрезание многоуровневой тематической иерархии
- Моделирование тематически несбалансированных коллекций

Регуляризатор семантической однородности

Минимизация суммарной семантической неоднородности тем:

$$\sum_{t\in T} S_t = \sum_{d\in D}\sum_{w\in d}\left(\sum_{t\in T}\frac{n_{tdw}}{n_t}\right)\ln\frac{\hat{p}(w|d)}{p(w|d)} \;\to\; \min_{\Phi,\Theta}$$

Регуляризатор в сумме с log-правдоподобием, $\beta_{dw} = \sum_t \frac{p_{tdw}}{p_t}$ (увеличение веса β_{dw} для термов из редких тем):

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\left(1+\tau\beta_{dw}\right)\ln\sum_{t\in T}\phi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

Модифицированный EM-алгоритм

$$p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \qquad \beta_{dw} = \sum_t \frac{p_{tdw}}{p_t}$$

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_d \tilde{n}_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \qquad \tilde{n}_{dw} = n_{dw}(1+\tau\beta_{dw})$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_w \tilde{n}_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \qquad p_t = \frac{1}{n}\sum_{dw} n_{dw} p_{tdw}$$

Резюме

- Построение BTM — задача многокритериальная:
  много регуляризаторов, много критериев (метрик) качества
- ARTM позволяет улучшать сразу несколько критериев,
  ценой незначительного ухудшения перплексии
- Сглаживание + разреживание + декоррелирование
  — часто используемая комбинация регуляризаторов

Открытые проблемы

- Подобрать стратегию регуляризации для наилучшей согласованности модели с тематическими цепочками
- Подобрать лучшую формулу внутритекстовой когерентности (в новом дизайне эксперимента)
- Автоматизировать подбор коэффициентов регуляризации по заданному приоритетному списку критериев
- Проверить, решает ли регуляризатор семантической однородности проблему несбалансированности тем