Вероятностные тематические модели
Лекция 6. Оценивание качества тематических моделей

К. В. Воронцов
k.vorontsov@iai.msu.ru

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

ВМК МГУ • 31 марта 2025

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 аспектов качества модели интерпретируемые, не всегда гладкие, их много разных

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

Напоминание. 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?

Напоминание. 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.

Напоминание. Эксперимент по поиску меры интерпретируемости

Измерялась ранговая корреляция Спирмена экспертных оценок с каждой из 15 мер 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).

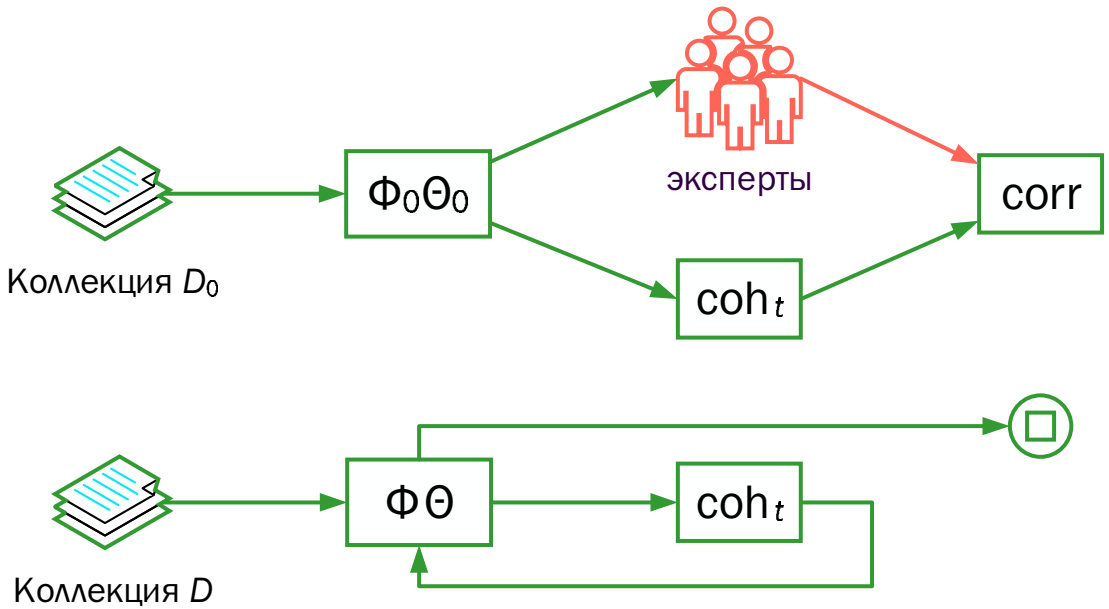
Когерентность модели = средняя когерентность всех тем.

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

Схема калибровочного 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

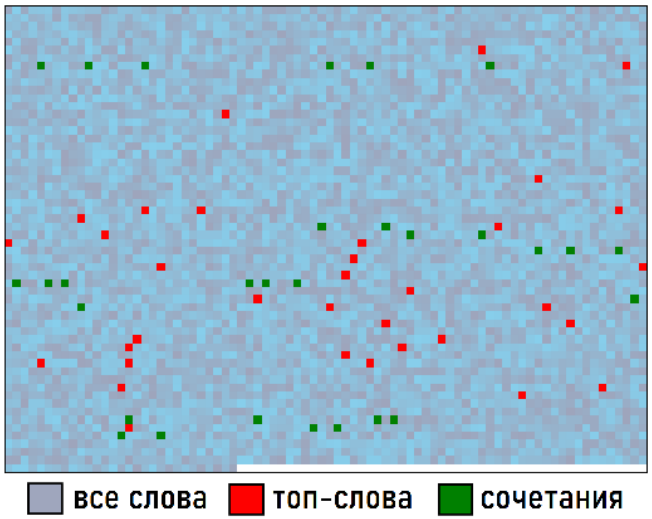
На новой коллекции D используем откалиброванный критерий (когерентность тем coh_t) для оценивания и выбора моделей $\Phi\Theta$



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 $(u,v) \in W^2$ 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:
  $\mathrm{rel}_t(u,v) = \sqrt{\phi_{ut}\phi_{vt}}$ или $[\phi_{ut}\phi_{vt} \geqslant \varepsilon]$
- поэкспериментировать 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 (intra-text coherence)

Средневзвешенная когерентность темы:

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

где суммирование по парам слов (u,v) в общих *контекстах*, например, в одном предложении или на расстоянии ± 10 слов.

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

Возможные модификации:

- $\mathrm{rel}_t(u,v) = \sqrt{p(t|d,u)\,p(t|d,v)}$ после E-шага
- перейти в coh от документных частот N_* к терм-парным m_*:
  $\mathrm{coh}(u,v) = \frac{m}{m_u m_v}$, $m_w = \sum_{u,v} [u=w] + [v=w]$, $m = \sum_w m_w$,

---

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

Что такое терм-парные частоты (term-pair frequency)

Пример: словарь $W = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}\}$, ширина окна $h = 5$
текст: «$\boxed{\mathtt{A\ B\ C\ B\ A}}\ \mathtt{D\ A\ C\ C\ B\ D\ A}$», длина текста $n = 12$
число пар термов во всех окнах: $m = (n - h + 1)(h - 1) = 32$

частоты m_{uv} пар термов

A A	2
A B	3
A C	5
A D	2
B A	3
B B	1
B C	2
B D	2
C A	3
C B	2
C C	1
C D	2
D A	1
D B	1
D C	2
D D	0
	32

частоты n_w термов

A	4
B	3
C	3
D	2
	12

частоты m_w термов-в-парах: левые, правые, двусторонние

A∗	12	∗A	9	A	21
B∗	8	∗B	7	B	15
C∗	8	∗C	10	C	18
D∗	4	∗D	6	D	10
	32		32		64

Два варианта расчёта отношения вероятностей пары (u, v) к вероятностям термов-в-паре (на примере $u = \mathtt{A}$, $v = \mathtt{C}$):

$$\frac{P_{uv}}{P_{u*} \cdot P_{*v}} = \frac{P(\mathtt{AC})}{P(\mathtt{A}*) \cdot P(*\mathtt{C})} = \frac{\frac{5}{32}}{\frac{12}{32} \cdot \frac{10}{32}} = 1.33$$

$$\frac{P_{uv,vu}}{P_{u*,*u} \cdot P_{v*,*v}} = \frac{P(\mathtt{AC} \cup \mathtt{CA})}{P(\mathtt{A}) \cdot P(\mathtt{C})} = \frac{\frac{5+3}{64}}{\frac{12+9}{64} \cdot \frac{8+10}{64}} = 1.35$$

Как проверить адекватность внутритекстовой когерентности

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

Идея:

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

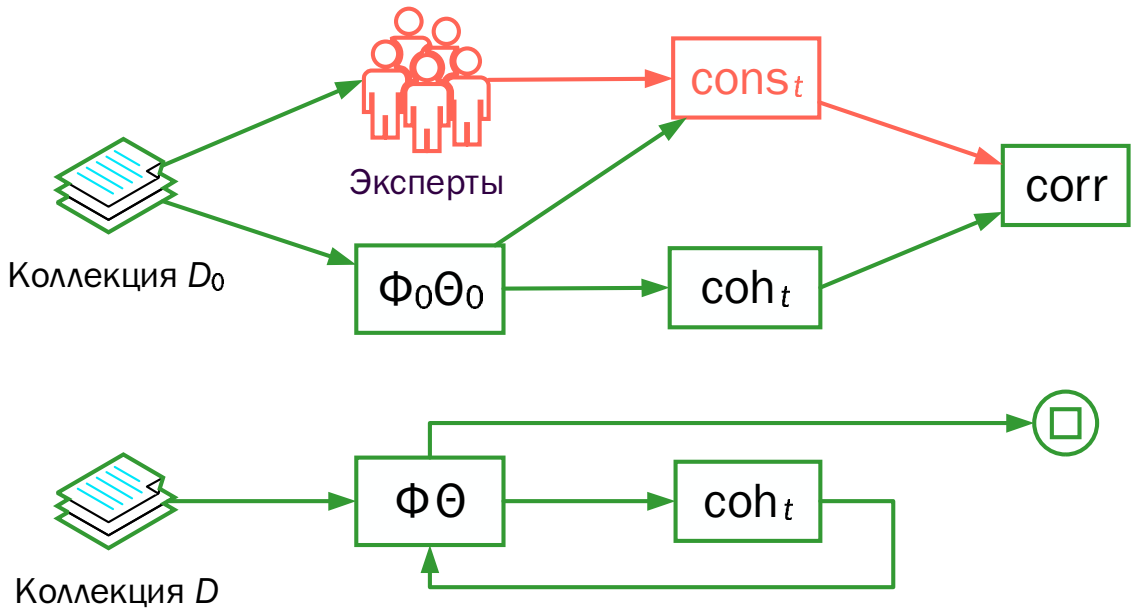
Пример разметки:

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 для 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:

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

Цепочки фрагмента должны относиться к различным темам:

$$\frac{\sum_d \#\{t \colon C_{di} \in C(t)\}}{\sum_d \#\{C_{di}\}} \approx 1 \quad (\leqslant 1)$$

Внутритекстовая когерентность: преимущества и проблемы

Преимущества: теперь

- темы оцениваются по полному массиву текста
- вместо документов учитываются локальные контексты
- разметка отделена от построения тематической модели

Открытые проблемы:

- подобрать оптимальные формулы для $\mathrm{rel}_t(u,v)$ и $\mathrm{coh}(u,v)$
- исследовать варианты вычисления cons_t и $C(t)$
- не вырождаются ли распределения cons_t по темам?
- возможно ли определять оптимальное число тем T по максимуму средней согласованности cons?
- как согласовывать тематические модели с цепочками?
- собрать больше тематических цепочек в разных доменах

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):

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

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

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

Естественная разреженность матриц Φ и Θ в экспериментах:

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

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

Напоминание. Лексическое ядро, чистота, 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

Гипотеза условной независимости

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

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:

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

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

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

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

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

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

Напоминание. Разреживающий регуляризатор 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.

Эксперименты с отбором тем на синтетических данных

Коллекция статей NIPS (Neural Information Processing System)

- $|D| = 1566$ обучающих документов; $|D'| = 174$ тестовых
- $|W| = 13\,\text{K}$ — мощность словаря

Синтетическая коллекция:

- строим PLSA за 500 итераций, $|T_0| = 50$ тем на NIPS
- генерируем коллекцию (n^0_{dw}) из полученных Φ и Θ:

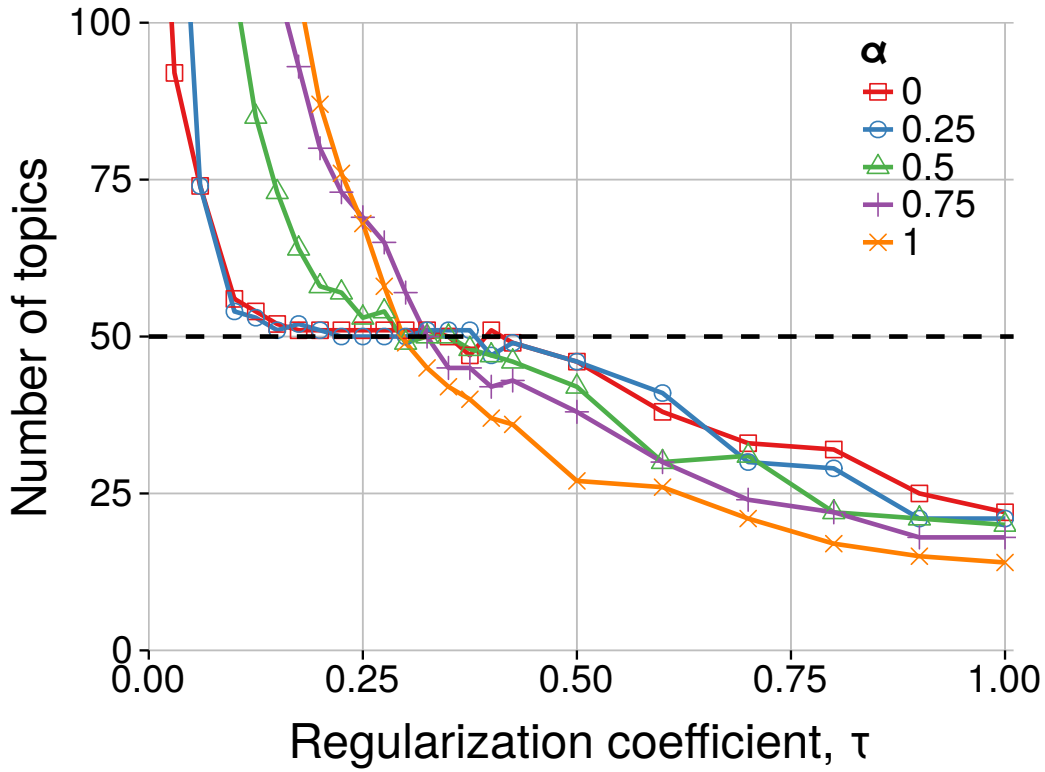
$$n^0_{dw} = n_d \sum_{t \in T_0} \phi_{wt}\theta_{td}$$

Параметрическое семейство полусинтетических данных:

- n^α_{dw} — смесь синтетических данных n^0_{dw} и реальных n_{dw}:

$$n^\alpha_{dw} = \alpha n_{dw} + (1-\alpha) n^0_{dw}$$

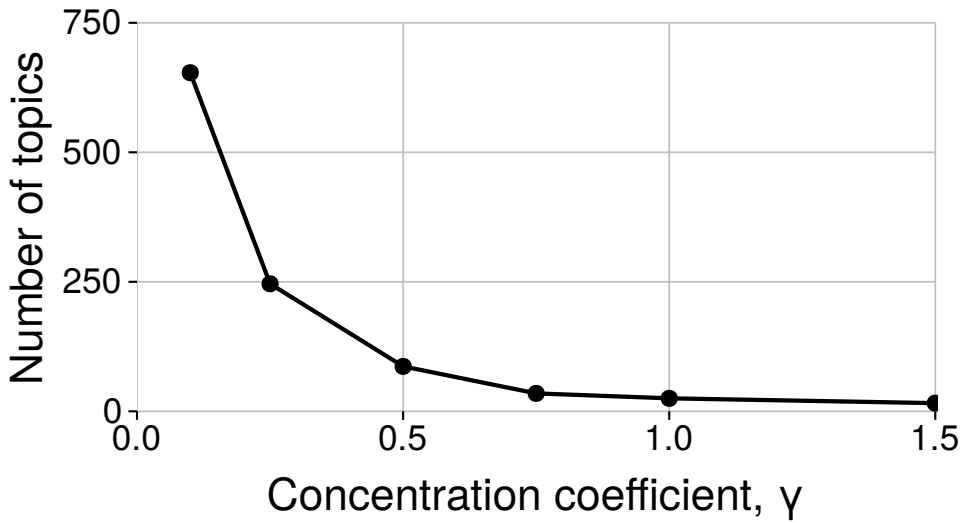
Попытка определения числа тем



- на синтетических данных надёжно находим $|T| = 50$
- причём в широком интервале значений коэффициента τ
- однако на реальных данных чёткого интервала нет

Сравнение с байесовской тематической моделью HDP

HDP, Hierarchical Dirichlet Process [Tech et.al, 2006] — «state-of-the-art» байесовский подход к определению числа тем

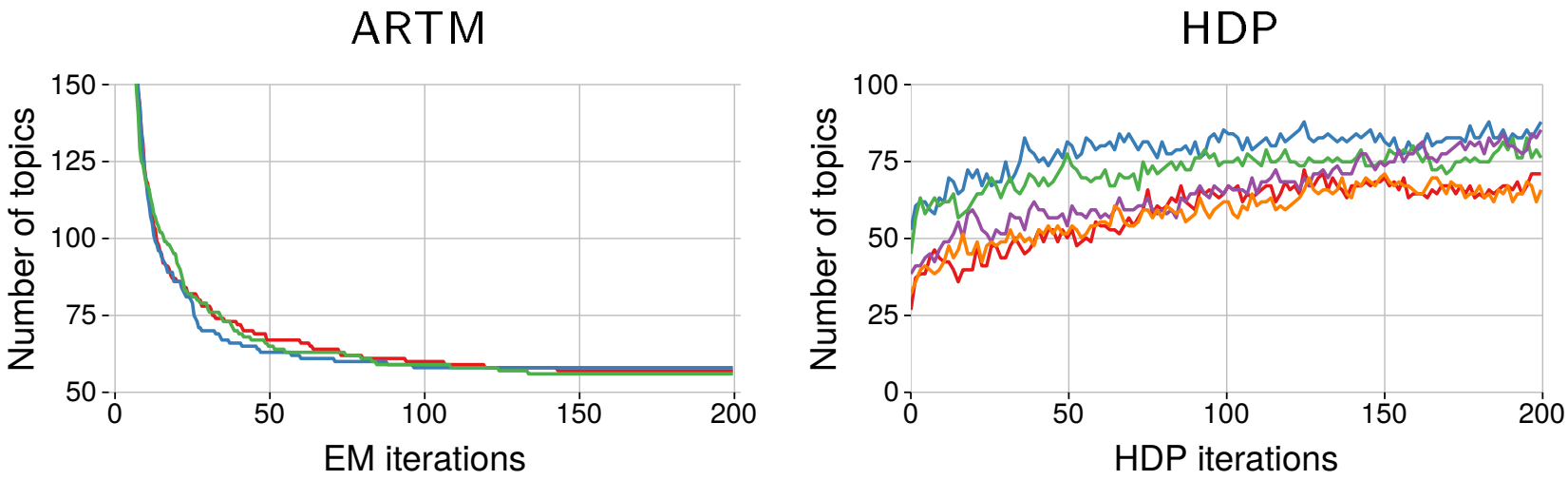


- *Коэффициент концентрации* γ в HDP влияет на $|T|$ так же сильно, как выбор коэффициента τ в ARTM.

Y.W.Teh, M.Jordan, M.Beal, D.Blei. Hierarchical Dirichlet Processes. 2005.

Сравнение ARTM и HDP по устойчивости

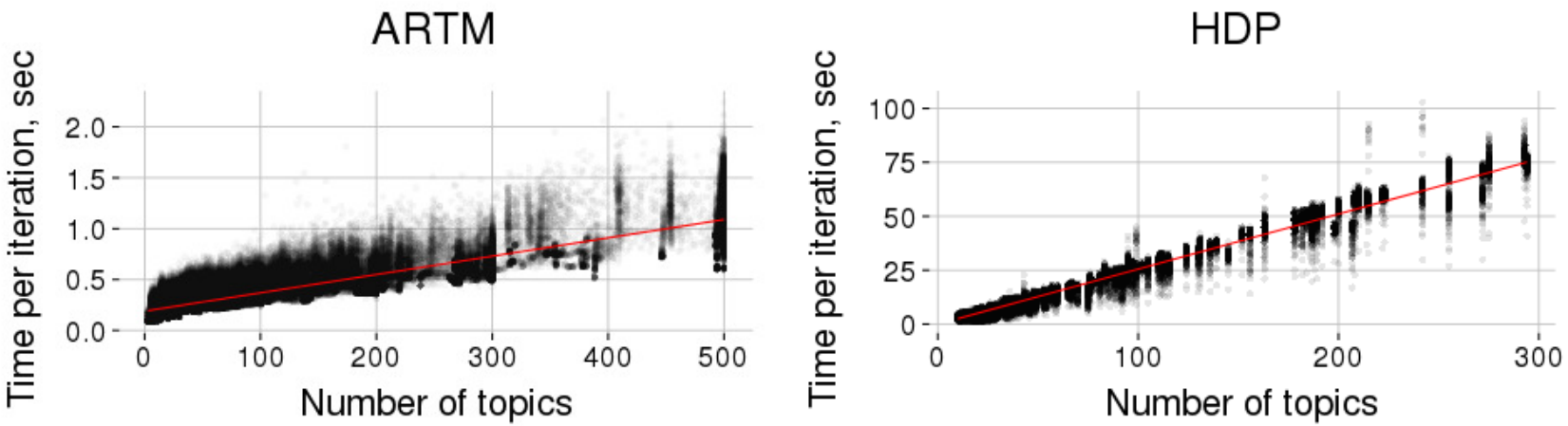
Запуск ARTM и HDP много раз из случайных инициализаций:



- HDP менее устойчив, причём в двух смыслах:
  - число тем сильнее флуктуирует от итерации к итерации;
  - результаты нескольких запусков различаются сильнее.
- «Рекомендуемые» значения параметров γ в HDP и τ в ARTM дают примерно равное число тем $|T| \approx 60$

Сравнение ARTM и HDP по времени вычислений

Сравнение времени одного прохода коллекции (sec)

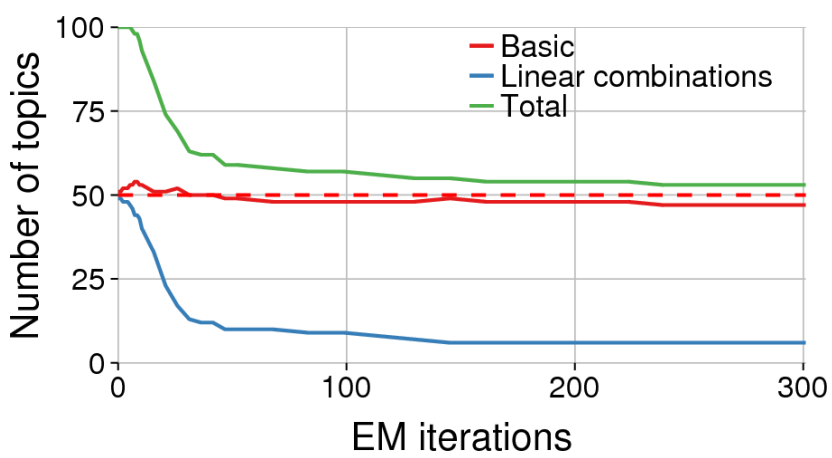


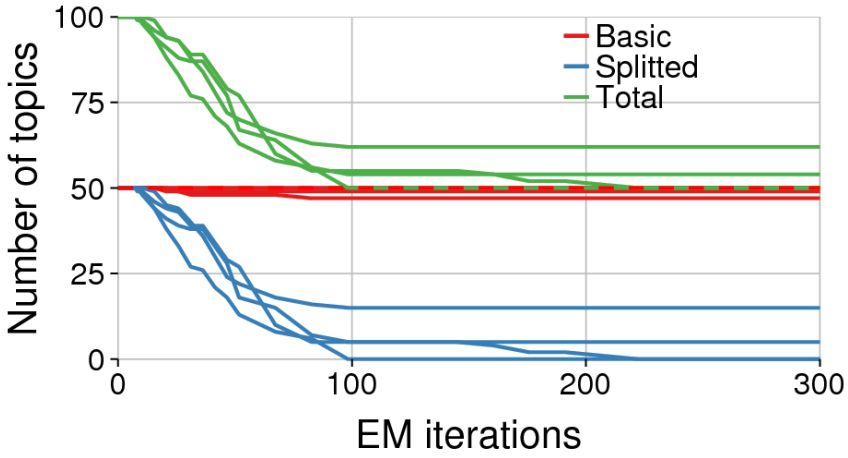
- ARTM в 100 раз быстрее!

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.

Удаление линейно зависимых и расщеплённых тем

Добавили 50 линейных комбинаций тем в модельную Φ.
Расщепили 50 тем, каждую на две подтемы в модельной Φ.





- Удаляются линейно зависимые и расщеплённые темы
- Остаются наиболее различные темы исходной модели.

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.

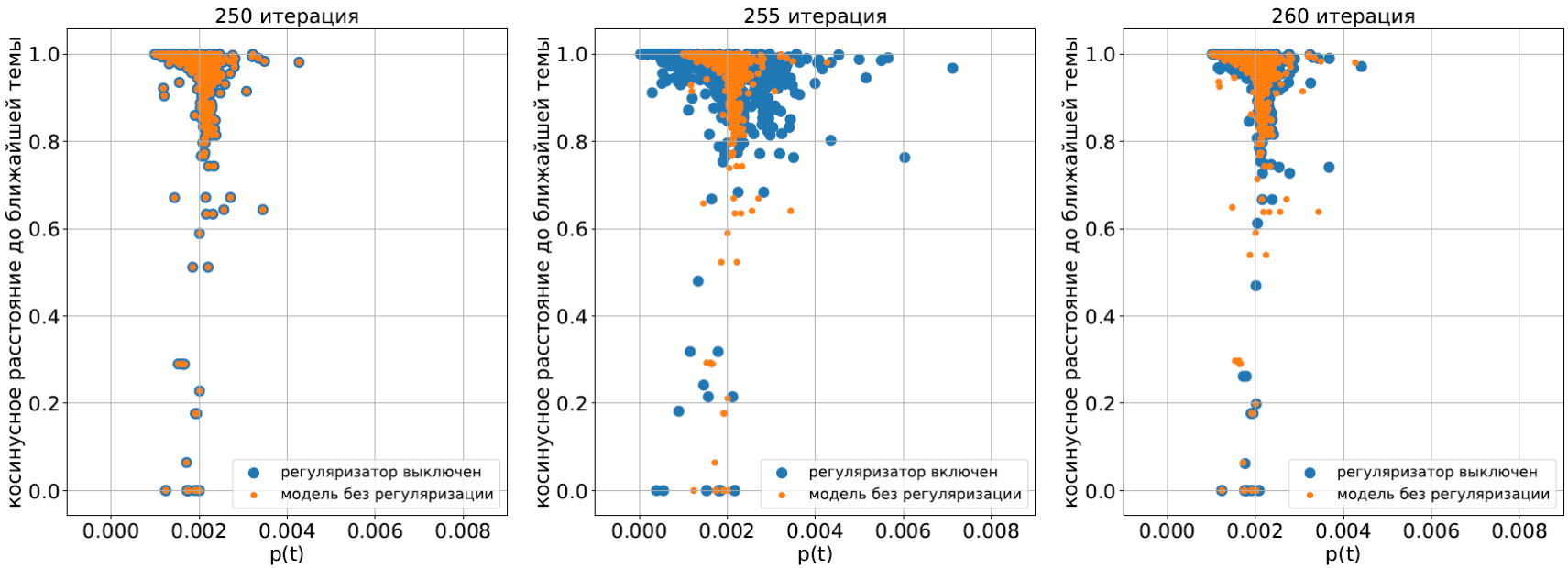
Выводы по результатам экспериментов

- Регуляризатор отбора тем удаляет незначимые темы и определяет оптимальное число тем, если оно существует
- Увы, в реальных данных его не существует!
  Оно задаётся исходя из целей моделирования.
- Значит, надо иерархически дробить темы на подтемы,
  и пусть пользователь выбирает нужную ему детализацию
- Есть простой метод для удаления лишних тем, но
  как обнаруживать новые темы в потоке или в батчах
  и добавлять их в ARTM — пока **открытая проблема**
- Регуляризатор отбора тем имеет полезный побочный эффект, удаляя линейно зависимые и расщеплённые темы
- Почему это происходит — **открытая проблема**

Проблема малых тем и тем-дубликатов

Эксперимент на коллекции postnauka.ru, $|T| = 500$

- Регуляризатор отбора тем плохо устраняет дубликаты,
- усиливает разброс тем по их мощности $p(t)$,
- который исчезает после отключения регуляризатора.
- Матричное разложение само не производит малые темы.

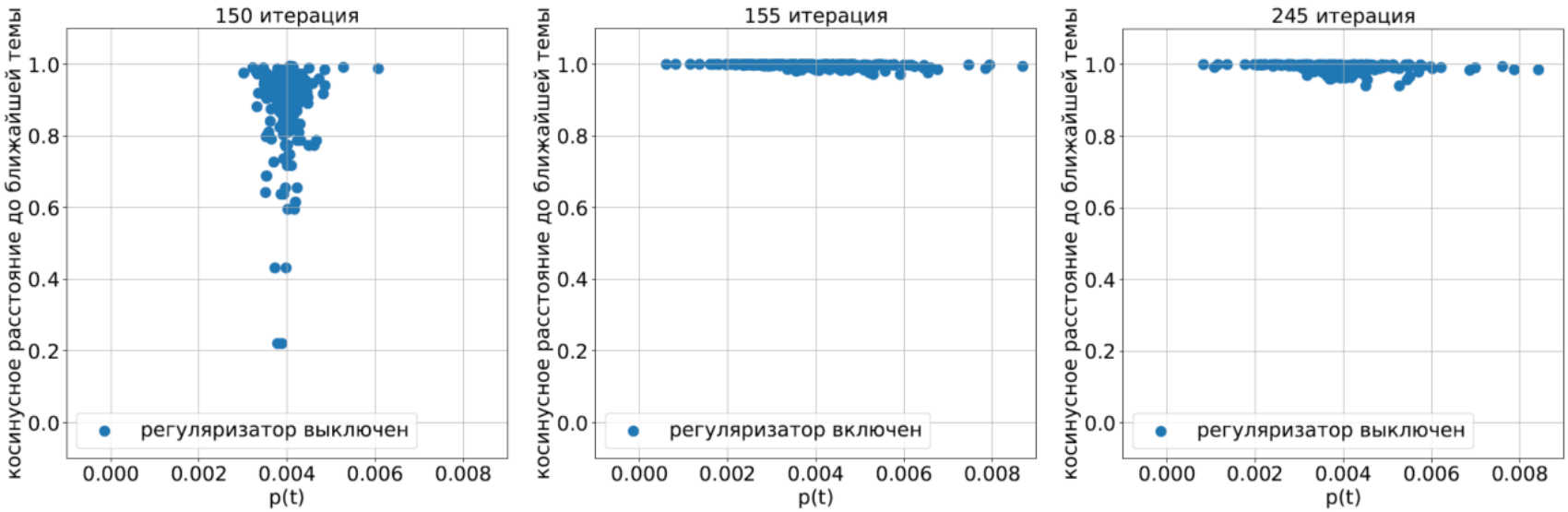


Г.Фоминская. Выявление тем-дубликатов в тематических моделях. Курсовая работа, ВМК МГУ, 2018.

Проблема малых тем и тем-дубликатов

Эксперимент на коллекции postnauka.ru, $|T| = 250$

- Регуляризатор декоррелирования удаляет дубликаты,
- усиливает разброс тем по их мощности $p(t)$;
- после отключения регуляризатора эти эффекты остаются.

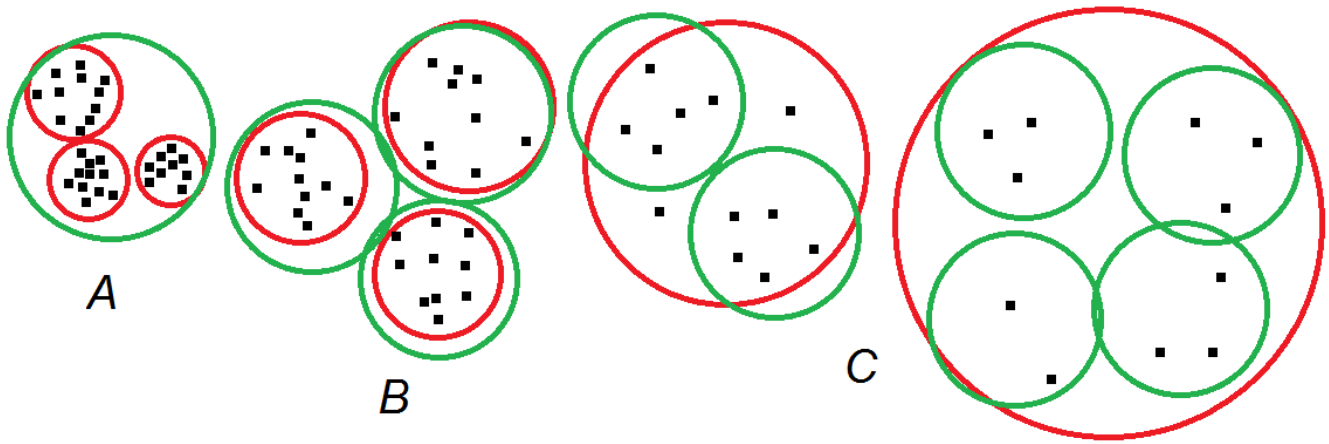


Г.Фоминская. Выявление тем-дубликатов в тематических моделях. Курсовая работа, ВМК МГУ, 2018.

Проблема расщепления и слияния тем

Тема — кластер на единичном симплексе размерности $|W| - 1$ с центром $p(w|t)$ и точками $p(w|t,d)$, $d \in D: \theta_{td} > 0$

- Тематические модели стремятся выравнивать темы по их мощности (красные кластеры).
- Это приводит к появлению тем-дубликатов (A) и семантически разнородных тем (C).
- Выравнивание тем по *радиусу семантической однородности* (зелёные кластеры) должно решать обе проблемы.



Резюме

- Построение BTM — задача многокритериальная:
  много регуляризаторов, много критериев качества
- ARTM позволяет улучшать сразу несколько критериев,
  ценой незначительного ухудшения правдоподобия
- Оптимизация числа тем — только по внешнему критерию
- Новая улучшенная мера интерпретируемости тем
  — внутритекстовая когерентность
- Новое семейство средневзвешенных статистик для
  проверки статистических гипотез условной независимости
- Новый регуляризатор семантической однородности
  — решает ли проблему несбалансированности тем?
- Тематическая несбалансированность текстовой коллекции
  — основная причина плохой интерпретируемости тем
  (слияния мелких тем и дублирования крупных)

Задания по курсу

Задача-минимум: научиться решать задачи NLP
с использованием тематического моделирования в BigARTM

Задача-максимум: сделать полезное мини-исследование

виды деятельности	оценка
теоретические задания	$\sum_i X_i$
решение прикладной задачи	$5X$
обзор по NeuralTM	$5X$
интеграция ARTM в pyTorch	$5X$
участие в одном из проектов	$10X$
работа над открытой проблемой	$10X$

где X — оценка за вид деятельности по 5-балльной шкале.

Итоговая оценка: $\min\big(5, \lfloor \text{score}/10 \rfloor\big)$ по 5-балльной шкале.

Теоретическое задание к лекции 2

Упражнения на принцип максимума правдоподобия:

1. Униграммная модель документов: $p(w|d) = \xi_{dw}$
Найти параметры модели ξ_{dw}.

2. Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d
Найти параметры модели ξ_w.

Подсказка: применить условия ККТ или основную лемму.

3. (более творческое задание)
Предложите модель, определяющую роли слов в текстах:
— тематические слова
— специфичные слова документа (шум)
— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т}, \text{ш}, \text{ф}\}$.
Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.
Подсказка 3: можно использовать документную частоту слов.

Теоретическое задание к лекции 3

4. Заменим ln гладкой монотонно возрастающей функцией μ:

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

Как изменится EM-алгоритм? Возможно ли подобрать функцию μ так, чтобы сократился объём вычислений?

5. Заменим ln гладкой монотонно возрастающей функцией μ в регуляризаторе сглаживания–разреживания (модель LDA):

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

Как изменится M-шаг и воздействие регуляризатора на модель?

6. Какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \operatorname*{norm}_{w}\big(n_{wt}[n_{wt} > \gamma n_t]\big)$$

Теоретическое задание к лекции 3

Аналитик построил тематическую модель Φ^0, Θ^0
и отметил среди столбцов матрицы Φ^0 темы двух типов:
удачные $T_+ \subset T$ и неудачные $T_- \subset T$.

Теперь он хочет построить модель ещё раз так, чтобы

- удачные темы остались в матрице Φ;
- остальные темы построились по-другому и были не похожи на каждую из неудачных тем $t \in T_-$.

7. Предложите регуляризаторы для этого.

8. Не получится ли так, что новые темы будут отдаляться от суммы неудачных тем $\sum_{t \in T_-} \phi^0_{wt}$ вместо того, чтобы отдаляться от каждой из неудачных тем по отдельности? Почему это плохо и как этого избежать?

9. Предложите способ инициализации Φ для новой модели.

Теоретическое задание к лекции 4

10. Выведите EM-алгоритм с локализованным E-шагом (слайд 13) для локализованной тематической модели. Какие переменные удобнее оставить в модели, ϕ_{wt} или ϕ'_{tw}?

11. Предложите параметризацию для тематической модели внимания (слайд 21) Используя «основную лемму», получите уравнения для новых параметров модели.

Исследовательское задание к лекции 4

Открытая проблема. Продолжить исследование Ильи Ирхина:

- Освоить код: https://github.com/ilirhin/python_artm
- Реализовать локализованный E-шаг

Исследовать зависимость метрик качества от параметров (перплексия, разреженность, различность, когерентность):

- L — число проходов
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — асимметричность левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, глав
- β — баланса левого и правого контекста
- α, δ — параметры онлайнового EM-алгоритма
- опция «подставлять p_{ti}/n_t вместо $\phi_{w_i t}$ на E-шаге»
- опция «исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}$, $\overleftarrow{\theta}_{ti}$»

Теоретическое задание к лекции 5

12. Для иерархической тематической модели с рег. $R(\Phi, \Psi)$ предложите способ разреживания матрицы связей $\Psi = \big(p(s|t)\big)$, гарантирующий, что
1) у каждой родительской темы будет хотя бы одна дочерняя;
2) у каждой дочерней темы будет хотя бы одна родительская.

Подсказка: можно придумывать критерий регуляризации, а можно — формулу M-шага для матрицы Ψ.

13. Предложите способ гарантировать, что если родительская тема t получает только одну дочернюю s, то она переходит в неё целиком и как распределение: $p(w|s) = p(w|t)$.

14. Предложите способ согласования вероятностных смесей $p(w|t) \approx \sum_{s \in S} p(w|s)p(s|t)$ и $p(t|d) \approx \sum_{s \in S} p(t|s)p(s|d)$ с учётом тождества $p(s|t)p(t) = p(t|s)p(s)$.

Исследовательское задание к лекции 5

Участие в проекте «Мастерская знаний»

Дано:

- подборки, сгенерированные SciRus по одной статье
- асессорская разметка статей подборки по релевантности
- несколько вариантов токенизации
  — в том числе с автоматическим выделением терминов

Найти:

- тематическую модель
- модель ранжирования подборки по релевантности
- оптимальные: токенизацию, число тем, регуляризаторы
- распределение терминов по тематичности

Критерий:

- качество ранжирования
- (визуально) интерпретируемость тем
  — в том числе автоматического именования тем

Исследовательские задания к лекции 6

- Проблема несбалансированности тем
  - генераторы синтетических несбалансированных коллекций
  - модели локального контекста лишены этой проблемы?
  - регуляризаторы декоррелирования + семантической однородности
- Семейство средневзвешенных статистик
  - генераторы синтетических коллекций, удовлетворяющих гипотезе условной независимости
  - как (и нужно ли) определять пороги для построения статистических тестов условной независимости?
  - как ослабить проверку гипотезы условной независимости в модели локального контекста?
  - как перестраивать несогласованные темы?
- Критерий внутритекстовой когерентности
  - найти лучший вариант критерия с помощью калибровки по размеченным тематическим цепочкам
  - вычисление критерия должно естественным образом встраиваться в модель локального контекста

Примеры датасетов для практических заданий по курсу

- Открытые датасеты (английский): 20 newsgroups, NIPS, KOS
- Научные статьи: eLibrary, Semantic Scholar, arXiv, PubMed
- Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
- TechCrunch (английский)
- Данные социальных сетей: VK, Twitter, Telegram,...
- Википедия
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

---

http://bigartm.org
http://drive.google.com/drive/folders/1PPnw6aZOJAJoLRYuwdGm437RssV-XQx0

Проекты

- «Мастерская знаний» для научного поиска
  — пользователь строит тематические подборки статей,
  — поисковая выдача формируется моделью SciRus.
  — задача: показать пользователю тематику подборки
  — понадобится автоматическое выделение терминов,
  — выделение тематических фраз из документов,
  — автоматическое именование и суммаризация тем
  — конечная цель: ускорить понимание предметной области
- «Тематизатор» для социо-гуманитарных исследований
  — пользователь задаёт грубый фильтр текстового потока
  — задача: «классифицировать иголки в стоге сена»,
  — разделив темы на информативные и мусорные,
  — выделив аспекты и тональности в каждой теме
  — конечная цель: q&q аналитика проблемной среды

Открытые проблемы тематического моделирования

1. Проблема несбалансированности тем в коллекции
2. Обеспечение 100%-й интерпретируемости тем
3. Тематические модели внимания последовательного текста
4. Обнаружение новых тем или трендов в потоке текстов
5. Автоматическое именование и аннотирование тем
6. Обзор подходов в нейросетевых тематических моделях
7. Обеспечение полноты и устойчивости множества тем
8. Автоматический подбор гиперпараметров, AutoML
9. Оптимизация гиперпараметров в потоковом режиме
10. Проблема несбалансированности текстов по длине
11. Бережное слияние моделей нескольких коллекций
12. Гиперграфовые тематические модели в RecSys