

Вероятностные тематические модели

Лекция 3. Оценивание качества тематических моделей

Константин Вячеславович Воронцов

k.v.vorontsov@phystech.edu

Этот курс доступен на странице вики-ресурса

<http://www.MachineLearning.ru/wiki>

«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

МФТИ.ФПМИ.ИС.ИАД • 2026-10-08

1 Измерение качества тематических моделей

- Правдоподобие и перплексия
- Интерпретируемость и когерентность
- Разреженность и различность тем

2 Проверка гипотезы условной независимости

- Проверка статистических гипотез
- Статистики на основе KL-дивергенции
- Обобщение средневзвешенных статистик

3 Проблема несбалансированности тем

- Влияние регуляризаторов на балансировку тем
- Балансировки тем с помощью нормировки
- Регуляризатор семантической однородности

Напоминание. ARTM: тематическая модель «мешка термов»

Дано: коллекция текстовых документов D , словарь W ;
 n_{dw} — частота термина $w \in W$ в документе $d \in D$.

Найти: вероятностную языковую модель $p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$
 с параметрами $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$

Критерий: $\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\text{Е-шаг: } \begin{cases} p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}) \end{cases}$$

$$\text{М-шаг: } \begin{cases} \phi_{wt} = \text{norm}_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \text{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum_{w \in W} n_{dw} p_{tdw} \end{cases}$$

Напоминание. Attentive ARTM: модель локальных контекстов

Дано: последовательность $w_1, \dots, w_n$ термов словаря W ;
 $C_i \subset \{1, \dots, n\}$ — локальный контекст термина w_i , $i \in \{1, \dots, n\}$;
 α_{ci} — коэффициент внимания, вес термина w_c из C_i для w_i .

Найти: вер. языковую модель $p(w|C_i) = \sum_{t \in T} \phi_{tw} \frac{p(w)}{p(t)} p(t|C_i)$
 с параметрами $\phi_{tw} = p(t|w)$

Критерий: $\sum_{i=1}^n \ln \sum_{t \in T} \phi_{tw_i} \frac{p(w_i)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} + R(\Phi) \rightarrow \max_{\Phi}$

ЕМ-алгоритм (после некоторых насильственных упрощений):

$$\begin{aligned} \text{Е-шаг: } & \begin{cases} p_{ti} = \text{norm}_{t \in T} \left(\frac{\phi_{tw_i}}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} \right), & p(t) = \sum_{w \in W} \phi_{tw} p(w) \end{cases} \\ \text{М-шаг: } & \begin{cases} \phi_{tw} = \text{norm}_{t \in T} \left(n_{tw} + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} \right), & n_{tw} = \sum_{i=1}^n p_{ti} [w_i = w] \end{cases} \end{aligned}$$

Проблематика оценивания качества тематических моделей

- языковая модель оптимизируется для предсказания слов в тексте, но используется для других целей
- цели трудно формализуемые, задача многокритериальная
- качество тем часто не устраивает прикладных экспертов
- для оптимизации — гладкие критерии регуляризации
- для измерения разных аспектов качества — критерии интерпретируемые, часто не гладкие, разнообразные
- два основных подхода к оцениванию качества:
 - *внутренние критерии* используют Φ , Θ и коллекцию
 - *внешние критерии* используют дополнительные данные

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.

Anna Shadrova. Topic models do not model topics: epistemological remarks and steps towards best practices. 2021

Критерии (метрики, меры) качества тематических моделей

Внутренние критерии используют только Φ , Θ и коллекцию

- 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
- статистические тесты согласия, условной независимости

Внешние критерии используют внешние данные

- качество решения прикладной задачи: ранжирования, поиска, категоризации, суммаризации, сегментации и т.п.
- экспертные оценки полезности (интерпретируемости) тем

Правдоподобие и перплексия (perplexity)

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

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

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

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

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):

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

$d = d' \sqcup d''$ — случайное разбиение на две части равной длины;
параметры ϕ_{wt} оцениваются по обучающей коллекции D ;
параметры θ_{td} оцениваются по первой половине d' ;
перплексия $\mathcal{P}(D')$ вычисляется по вторым половинам d'' .

Проблема: как разбивать документ на две половины?

- как мешок слов? но тогда $n_{d''w} \rightarrow n_{d'w}$ при $n_d \rightarrow \infty$
- как связный текст? но начало и конец могут быть о разным

Напоминание. Когерентность Ньюмана

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), где:

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.

Обобщение — семейство средневзвешенных когерентностей

Средневзвешенная когерентность темы:

$$\text{coh}_t = \frac{\sum_{u,v} \text{rel}_t(u, v) \text{coh}(u, v)}{\sum_{u,v} \text{rel}_t(u, v)} = \text{avg}_{u,v}(\text{rel}_t(u, v), \text{coh}(u, v))$$

$\text{coh}(u, v)$ — сочетаемость пары слов $(u, v) \in W^2$ в текстах

$\text{rel}_t(u, v)$ — релевантность слов u и v теме t , в частности

$\text{rel}_t(u, v) = [\phi_{ut}, \phi_{vt} > \text{top}_k \phi_{wt}]$ — когерентность Ньюмана

Текстовая когерентность (intra-text coherence):

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 аккумуляруются суммы в числителе и в знаменателе.

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

Текстовая когерентность: преимущества и проблемы

Преимущества: теперь

- когерентность оценивается по полному массиву текста
- разметка отделена от построения тематической модели
- участие экспертов можно минимизировать или исключить

Открытые проблемы:

- провести калибровку для текстовой когерентности:
подобрать наилучшее сочетание эвристик: rel, coh, окно
- как полностью отказаться от экспертных оценок? [2025]

Michael Röder et al. Exploring the space of topic coherence measures. 2015.

Alexander Hoyle et al. Is automated topic model evaluation broken? The incoherence of coherence. 2021.

Dominik Stammbach et al. Revisiting automated topic model evaluation with large language models. 2023.

Hamed Rahimi et al. Contextualized Topic Coherence Metrics. 2024.

Zhiyin Tan, Jennifer D'Souza. Bridging the evaluation gap. Leveraging large language models for topic model evaluation. 2025.

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):

$$WR_t = \frac{1}{|W|} \sum_{w \in W} [\phi_{wt} > \frac{1}{|W|}] \quad WR = \frac{1}{|T|} \sum_{t \in T} WR_t$$

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

$$DR_d = \frac{1}{|T|} \sum_{t \in T} [\theta_{td} > \frac{1}{|T|}] \quad DR = \frac{1}{|D|} \sum_{d \in D} DR_d$$

Естественная разреженность матриц Φ и Θ в экспериментах:

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

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

Лексическое ядро, чистота, 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 между всеми парами тем — плохо!

Среднее расстояние от темы t до ближайшей к ней темы:

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

Расстояния между вероятностными распределениями (от 0 до 1):

- $\rho(\phi_t, \phi_s) = 1 - \frac{\sum_w \phi_{ws} \phi_{wt}}{(\sum_w \phi_{ws}^2)^{1/2} (\sum_w \phi_{wt}^2)^{1/2}}$ — косинусное
- $\rho(\phi_t, \phi_s) = 1 - |W_t \cap W_s| : |W_t \cup W_s|$ — Жаккара
- $\rho^2(\phi_t, \phi_s) = \frac{1}{2} \sum_w (\sqrt{\phi_{ws}} - \sqrt{\phi_{wt}})^2$ — 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 термина (тем выше, чем больше кросс-энтропия; $\min$ при распределении $p(t)$ или $\frac{1}{|T|}$, $\max$ при вырожденном):

$$H(w) = - \sum_{t \in T} p(t) \ln p(t|w) \quad \text{или} \quad \sum_{t \in T} p(t|w)^\gamma, \quad \gamma > 1$$

Доля нетематических термов:

- $\frac{1}{|W|} \sum_w [H(w) < H_0]$ — в словаре W
- $\frac{1}{n_d} \sum_w n_{dw} [H(w) < H_0]$ — в документе d
- $\frac{1}{n} \sum_d \sum_w n_{dw} [H(w) < H_0]$ — 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 о согласии дискретных распределений

Гипотеза: эмпирическое $\hat{p}(w|d)$ порождается моделью $p(w|d)$

$$H_0(d) : \quad \hat{p}(w|d) = \frac{n_{dw}}{n_d} \quad \sim \quad p(w|d) = \sum_t \phi_{wt} \theta_{td}.$$

Статистика для проверки этой гипотезы:

$$\begin{aligned} S_d &= \text{KL}(\hat{p}(w|d) \parallel p(w|d)) = \sum_w \hat{p}(w|d) \ln \frac{\hat{p}(w|d)}{p(w|d)} = \\ &= \text{avg}_w \left(n_{dw}, \ln \frac{\hat{p}(w|d)}{p(w|d)} \right) = \text{avg}_{w,t} \left(n_{tdw}, \ln \frac{\hat{p}(w|d)}{p(w|d)} \right), \end{aligned}$$

где $\text{avg}_{i \in I}(\gamma_i, x_i) = \frac{\sum_{i \in I} \gamma_i x_i}{\sum_{i \in I} \gamma_i}$ — средневзвешенное x_i с весами γ_i .

Проблема: для разреженных $p(w|d)$ нет асимптотики $S_d \sim \chi^2$

N.Cressie, T.R.C.Read. Multinomial goodness-of-fit tests, 1984.

В.Р.Целых, К.В.Воронцов. Критерии согласия для разреженных дискретных распределений и их применение в тематическом моделировании // JMLDA, 2012.

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:

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

Гипотеза согласованности документа d с темой t

— термы темы t порождаются независимо от документов:

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

Гипотеза согласованности термина w с темой t

— тема t распределена по документам независимо от термов:

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

Мера семантической неоднородности темы t в коллекции

Статистика для проверки гипотезы $H_0(t)$:

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

По определению условной вероятности и формуле Е-шага:

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

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

Возможное применение меры семантической неоднородности S_t :

- выявление тем для замены или разбиения на подтемы

Мера несогласованности документа d с темой t

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

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

По определению условной вероятности и формуле Е-шага:

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

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

Возможные применения меры несогласованности S_{td} :

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

Мера несогласованности термина w с темой t

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

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

По определению условной вероятности и формуле Е-шага:

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

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

Возможные применения меры несогласованности S_{wt} :

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

Семейство средневзвешенных статистик с функцией потерь ℓ_{dw}

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

$S_d = \text{avg}_{w,t}(n_{tdw}, \ell_{dw})$ — несогласованность документа

$S_t = \text{avg}_{d,w}(n_{tdw}, \ell_{dw})$ — неоднородность темы в коллекции

$S_{td} = \text{avg}_w(n_{tdw}, \ell_{dw})$ — несогласованность документа с темой

$S_{wt} = \text{avg}_d(n_{tdw}, \ell_{dw})$ — несогласованность терма с темой

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

$\ln \mathcal{P} = \text{avg}_{d,w,t}(n_{tdw}, \ell_{dw}) = \text{avg}_{d,w}(n_{dw}, \ell_{dw})$ — коллекции

$\ln \mathcal{P}_d = \text{avg}_{w,t}(n_{tdw}, \ell_{dw}) = \text{avg}_w(n_{dw}, \ell_{dw})$ — документа

$\ln \mathcal{P}_t = \text{avg}_{d,w}(n_{tdw}, \ell_{dw})$ — темы t (новая возможность!)

$\ln \mathcal{P}_{td} = \text{avg}_w(n_{tdw}, \ell_{dw})$ — 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 i , 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_d, S_t, S_{td}, S_{wt} , толерантные к повторяемости слов:

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

Тогда средневзвешенные статистики $S_d, S_t, S_{td}, S_{wt} \in [0, 1]$
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:

- S_d : определение плохо тематизируемых документов
- S_t : определение неоднородных тем для расщепления
- S_{wt} : определение общеупотребительных слов в темах
- S_{td} : распознавание присутствия новых тем в документе
- S_{wt}, ℓ_{dw} : выделение термов для инициализации новой темы

Аномально низкие значения статистик:

- S_t : выделение тем для интерпретации и анализа
- S_{wt} : выделение термов лексического ядра темы
- S_{td} : выделение тематичных документов/фраз темы

Нормальные значения статистик:

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

Регуляризатор декоррелирования тем

Цели: усилить различность тем; выделить в каждой теме лексическое ядро, отличающее её от других тем; способствовать переходу общей лексики в фоновые темы

Критерий: минимум ковариаций между вектор-столбцами ϕ_t

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

Подставляем в формулы М-шага, получаем поправку, которая разреживает и контрастирует строки матрицы Φ (малые вероятности ϕ_{wt} в строке становятся ещё меньше):

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

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:

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

Подставляем в M-шаг R , затем несмещённую оценку $\theta_{td}^* = \frac{n_d}{n_t}$:

$$\theta_{td} = \text{norm}_{t \in T} \left(n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \right) = \text{norm}_{t \in T} \left(n_{td} \left(1 - \frac{\tau}{n_t} \right) \right)$$

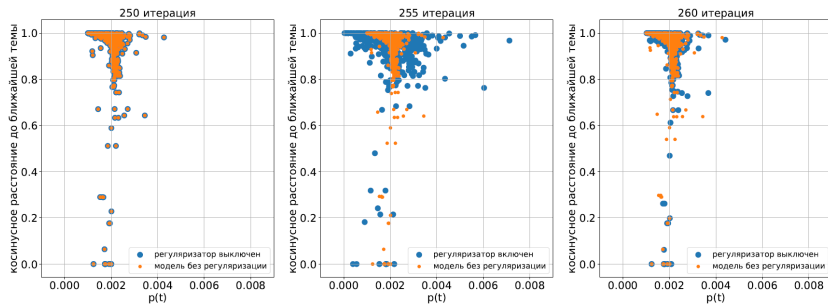
Эффект: обнуляются строки матрицы Θ с малыми n_t ,
заодно удаляются зависимые и расщеплённые темы

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

Проблема малых мусорных тем и тем-дубликатов

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

- **регуляризатор отбора тем** плохо устраняет дубликаты;
- усиливает разброс тем по их мощности $p(t)$,
но после отключения регуляризатора он исчезает;
- матричное разложение само не производит мелкие темы

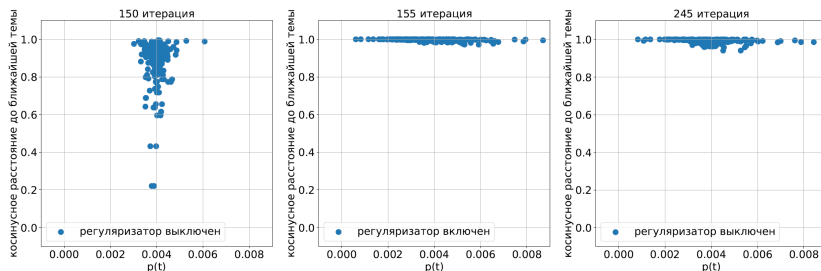


Г.Фоминская. Выявление тем-дубликатов в тематических моделях. Курсовая работа, ВМК МГУ, 2018.

Проблема малых мусорных 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.

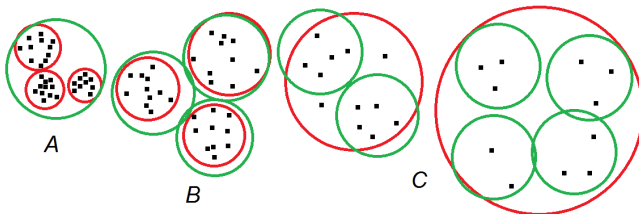


Г.Фоминская. Выявление тем-дубликатов в тематических моделях. Курсовая работа, ВМК МГУ, 2018.

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 выравнять темы по их мощности (**красные кластеры**)
- Это приводит к разделению крупных тем на дубликаты (A) и слиянию мелких тем в неоднородные «мусорные» (C)
- Выравнивание тем по *радиусу семантической однородности* (**зелёные кластеры**) должно решать обе проблемы



Эвристика балансировки тем с помощью нормирования

Теорема. Нормировка в EM-алгоритме для PLSA ($R = 0$)

$$\begin{aligned} \text{Е-шаг: } & \left\{ p_{tdw} \equiv p(t|d, w) = \text{norm}_{t \in T}(\phi_{wt}\theta_{td}) \right. \\ \text{М-шаг: } & \left\{ \sum_{d,w} \sum_{t \in T} n_{dw} p_{tdw} \mathbf{Z}_t \ln(\phi_{wt}\theta_{td}) \rightarrow \max_{\Phi, \Theta} \right. \end{aligned}$$

при любых $\mathbf{Z}_t$ не меняет формулу М-шага для матрицы Φ .

Доказательство. По лемме о максимизации на симплексах:

$$\begin{aligned} \phi_{wt} &= \text{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} \mathbf{Z}_t \right) = \text{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} \right) \\ \theta_{td} &= \text{norm}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} \mathbf{Z}_t \right) \end{aligned}$$

Нормировка $\mathbf{Z}_t$ влияет на Θ , но не влияет на Φ и темы $p(w|t)$

Регуляризатор семантической однородности

Минимизация суммарной семантической неоднородности тем:

$$R(\Phi, \Theta) = - \sum_{t \in T} S_t^\gamma = - \sum_{t \in T} \left(\sum_{d \in D} \sum_{w \in d} \frac{n_{tdw}}{n_t} \ln \frac{\hat{p}(w|d)}{p(w|d)} \right)^\gamma \rightarrow \max_{\Phi, \Theta}$$

Этот регуляризатор эквивалентен поправке log-правдоподобия $\beta_{dw} = \sum_t \gamma S_t^{\gamma-1} \frac{p_{tdw}}{p_t}$, повышающей вес термов из редких тем:

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\begin{aligned} p_{tdw} &= \text{norm}_{t \in T} (\phi_{wt} \theta_{td}) & \beta_{dw} &= \sum_t \gamma S_t^{\gamma-1} \frac{p_{tdw}}{p_t} \\ \phi_{wt} &= \text{norm}_{w \in W} \left(\sum_d \tilde{n}_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) & \tilde{n}_{dw} &= n_{dw} (1 + \tau \beta_{dw}) \\ \theta_{td} &= \text{norm}_{t \in T} \left(\sum_w \tilde{n}_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) & p_t &= \frac{1}{n} \sum_{dw} n_{dw} p_{tdw} \end{aligned}$$

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

Задача-минимум: научиться решать задачи анализа текстов с использованием тематического моделирования

Задача-максимум: получить новый научный результат

виды деятельности	оценка
теоретическая задача	X
теоретическая задача*	2X
теоретическая задача**	3X
решение прикладной задачи	10X
обзор по последним PTM/NTM	10X
участие в проекте	20X
работа над открытой проблемой	25X

где X — оценка за вид деятельности по 5-балльной шкале.
score — суммарная оценка по всем видам деятельности.

Итоговая оценка: $\min(5, \lfloor \text{score}/20 \rfloor)$ по 5-балльной шкале.

Теоретическое задание к лекции 1

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*. Творческое задание (возможны разные решения)

Предложите модель, определяющую роли слов в текстах:

- тематические слова
- специфичные слова документа (шум)
- слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т, ш, ф}\}$.

Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.

Подсказка 3: можно использовать документную частоту слов.

4. Пользуясь основной леммой, докажите, что регуляризатор бигермов эквивалентен добавлению псевдодокументов d_u в исходную коллекцию (см. слайды 27–28)

Исследовательская задача / открытая проблема:

оценивание (внутри)текстовой когерентности

- реализуйте вычисление средневзвешенной когерентности
- предложите способ разметки с помощью LLM
- подберите наилучшее сочетание эвристик rel и coh в калибровочном эксперименте без экспертной разметки
- какие регуляризаторы улучшают текстовую когерентность?
- воспроизводимо ли это улучшение на разных коллекциях?

5. Выведите формулы для вычисления β_{wk} (слайд 30)

при $\alpha_{ci} = \beta_{w_i, c-i}$

6.** Выведите формулы для вычисления β_{wk} , γ_w (слайд 38)

при $\alpha_{ci} = \text{norm}_{c \in C_i}(\gamma_{w_c} \beta_{w_c, i-c})$

Исследовательская задача / открытая проблема:

предложите свою параметризацию коэффициентов внимания, которая может привести к увеличению оптимальной ширины контекста и повышению качества тематической модели.

- возьмите за основу код Ильи Дьякова и Матвея Карпова
- добавьте свою модель в готовый сценарий экспериментов по сравнению моделей на датасетах
- предложите и реализуйте серию улучшений своей модели

7. Выведите ЕМ-алгоритм с регуляризатором семантической однородности, предполагая, что n_{tdw} и n_t — константы (внешние параметры, не зависящие от Φ, Θ).

Докажите, что подстановка этого регуляризатора в М-шаг эквивалентна введению мультипликативной поправки $(1 + \tau\beta_{dw})$ в критерий log-правдоподобия.

8.** Выведите ЕМ-алгоритм с регуляризатором семантической однородности, предполагая, что n_{tdw} и n_t выражаются через параметры модели Φ, Θ .

9.** Предложите формулу средневзешенных статистик S_* для тематической модели локальных контекстов.

Для проверки каких статгипотез подходят эти S_* ?

Проверьте, что полученная формула совпадает с введённой на лекции, если контекстом является весь документ.

Проблемы воспроизводимости и несбалансированности тем

- ❶ Генерация коллекций с заданным дисбалансом тем:
 - подбор нескольких сбалансированных коллекций (20NewsGroups? Фильтрация Википедии / Pubmed?)
 - построение сбалансированной ТМ с заданным $|T|$
 - генерация синтетических документов по заданным темам
 - введение скалярной меры *степени дисбаланса тем*
- ❷ Реализация CSMD метрик соответствия тем в двух ТМ:
 - Correspond — число совпавших тем
 - Split — число расщепившихся тем
 - Merge — число слившихся тем
 - Disappear — число несопоставленных (исчезнувших) тем
- ❸ Эксперименты с несбалансированными коллекциями:
 - зависимость CSMD метрик от $|T|$ и *степени дисбаланса тем*
 - для модели локального контекста
 - для регуляризаторов декоррелирования тем
 - для регуляризаторов когерентности
 - для регуляризатора семантической однородности