• Вероятностные языковые модели •
Лекция 7.
Оценивание качества тематических моделей

Константин Вячеславович Воронцов
k.vorontsov@iai.msu.ru

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

ВМК МГУ • 13 апреля 2026

Содержание

Напоминание. Тематическая модель «мешка термов»

Дано: коллекция текстовых документов D, словарь W;
n_{dw} — частота терма $w \in W$ 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: $\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

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

Напоминание. Тематическая модель локальных контекстов

Дано: последовательность $w_1, \dots, w_n$ термов словаря W;
$C_i \subset \{1, \dots, n\}$ — локальный контекст терма w_i, $1, \dots, n$;
α_{ci} — коэффициент внимания, вес терма 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: $\sum_{i=1}^{n} \ln \sum_{t \in T} \phi_{tw_i} \frac{p(w_i)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} + R(\Phi) \to \max_{\Phi}$

EM-алгоритм (после некоторых насильственных упрощений):

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

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:
  — *внешние критерии* используют дополнительные данные
  — *внутренние критерии* используют Φ, Θ и коллекцию

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

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 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

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 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 $n_{d''w} \to n_{d'w}$ при $n_d \to \infty$
- как связный текст? но начало и конец могут быть о разном

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.

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), 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).

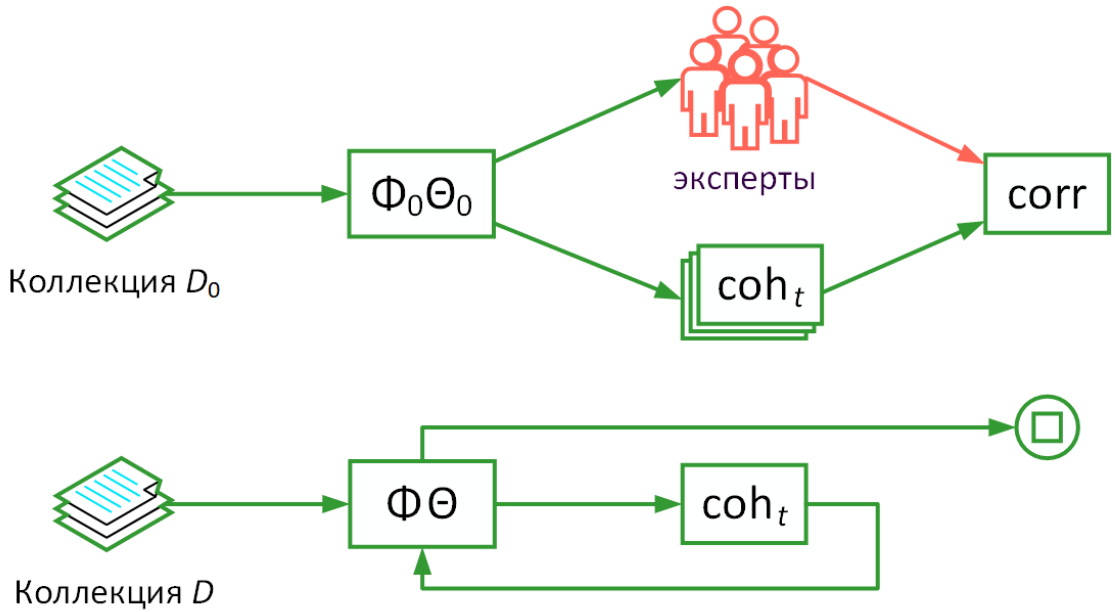
Когерентность модели = средняя когерентность всех тем.

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

Схема калибровочного эксперимента Ньюмана

1. строим по коллекции D_0 тематическую модель $\Phi_0\Theta_0$
2. эксперты оценивают темы (рейтингами или интрузиями)
3. ищем критерий, коррелирующий с оценками экспертов

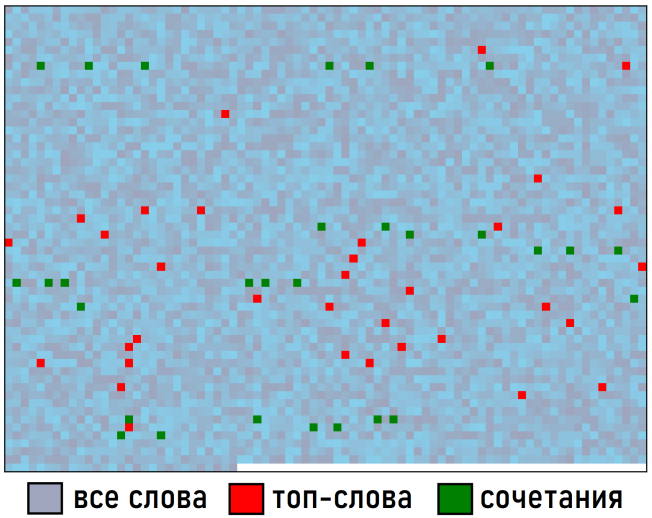
На новой коллекции D используем откалиброванный критерий (когерентность тем coh_t) для оценивания и выбора моделей $\Phi\Theta$



Недостаток когерентности

Обычно берут $k = 10..20$ топовых (самых частых) слов темы, 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:

$$\mathrm{coh}_t = \frac{\sum_{u,v} \mathrm{rel}_t(u,v)\,\mathrm{coh}(u,v)}{\sum_{u,v} \mathrm{rel}_t(u,v)} = \operatorname*{avg}_{u,v}\big(\mathrm{rel}_t(u,v), \mathrm{coh}(u,v)\big)$$

$\mathrm{coh}(u,v)$ — *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

Текстовая когерентность (intra-text coherence)

Средневзвешенная когерентность темы:

$$\mathrm{coh}_t = \frac{\sum_{u,v} \mathrm{rel}_t(u,v)\,\mathrm{coh}(u,v)}{\sum_{u,v} \mathrm{rel}_t(u,v)} = \operatorname*{avg}_{u,v}\big(\mathrm{rel}_t(u,v), \mathrm{coh}(u,v)\big),$$

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 $u = \mathtt{A}$, $v = \mathtt{C}$) — для пар упорядоченных ($uv \neq vu$) и неупорядоченных ($uv = vu$):

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

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

Как проверить адекватность текстовой когерентности

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

Идея: размечать слова в текстах, а не слова в темах

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

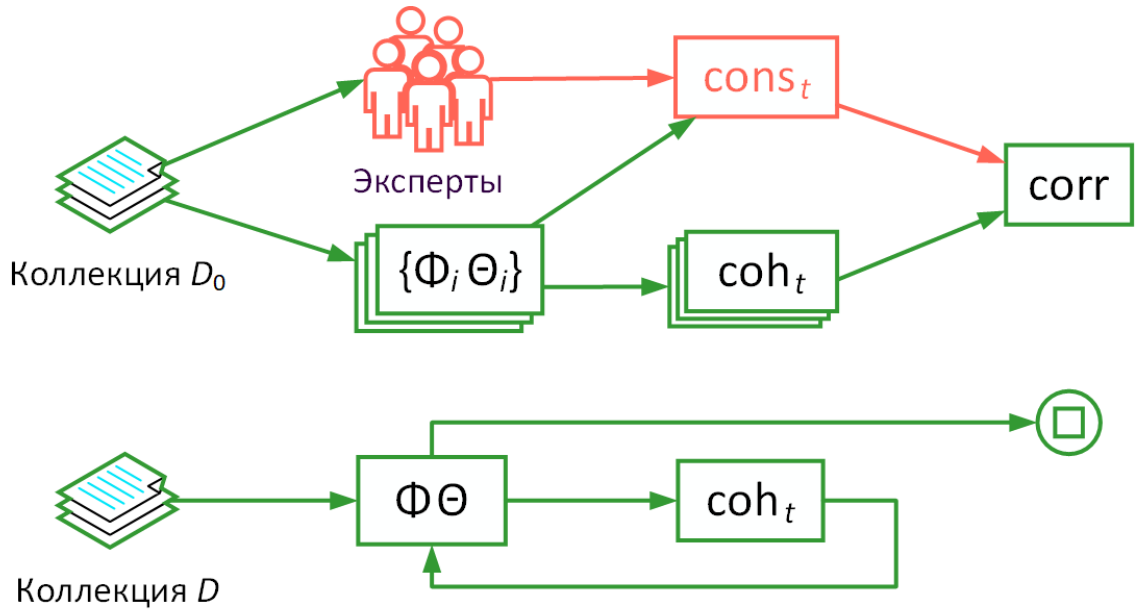
Пример разметки текста: три тематические цепочки

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 текстовой когерентности

1. выбираем из коллекции D_0 фрагменты для разметки
2. эксперты размечают тематические цепочки во фрагментах
3. строим по коллекции D_0 много моделей $\{\Phi_i\Theta_i\}$ с разнообразными темами, как хорошими, так и плохими
4. ищем критерий, коррелирующий с согласованностью cons_t между темами t и размеченными тематическими цепочками



Оценки согласованности тем с размеченными цепочками

Тема интерпретируема, если она согласована с разметкой,
т. е. различает в тексте цепочки, которые различили эксперты.

Тематика цепочки C, по формуле полной вероятности:

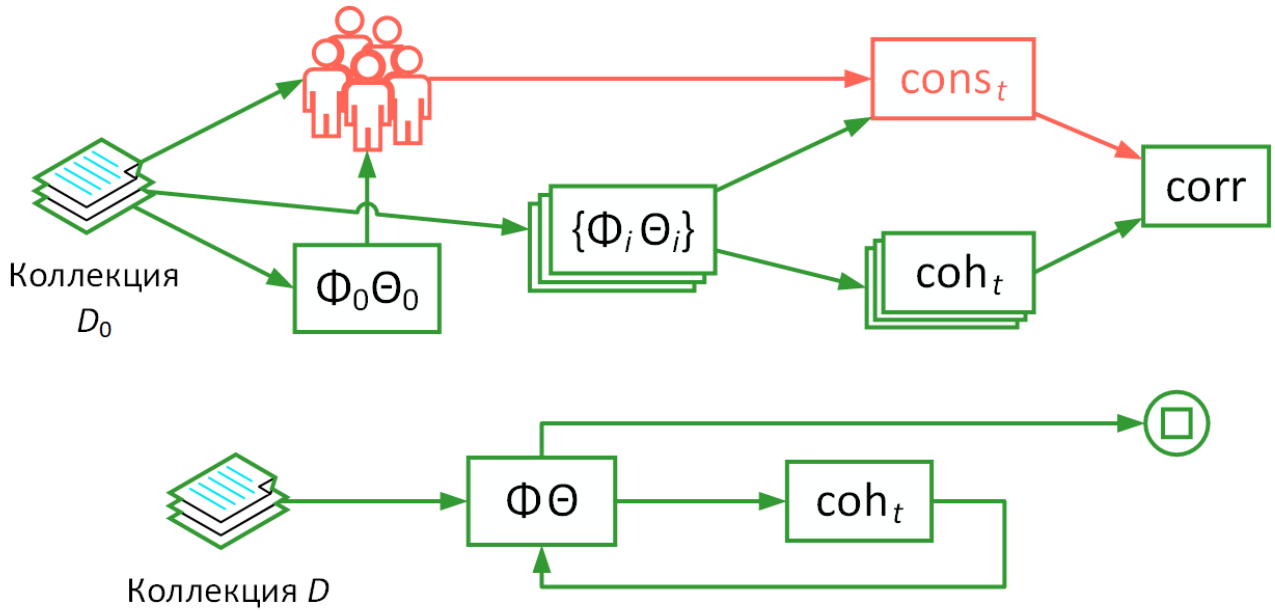
$$p(t|C) = \sum_{w \in C} p(t|w)p(w|C) = \frac{1}{|C|} \sum_{w \in C} p(t|w)$$

вход: Φ_0, множество цепочек $\{C_{dk} : d \in D_0, k \in K_d\}$;
выход: оценки согласованности тем cons_t, $t \in T$;
$S_t := 0$; $N_t := 0$ для всех $t \in T$;
для всех $d \in D_0$: **для всех** размеченных цепочек $k \in K_d$:
    $t := \arg\max_{t \in T} p(t|C_{dk})$ — главная тема цепочки C_{dk};
    $\Delta_t := p(t|C_{dk}) - \max_{j \neq k} p(t|C_{dj})$ — различимость темы t;
    $S_t \mathrel{+}= \Delta_t$; $N_t \mathrel{+}= 1$;
$\text{cons}_t := S_t / N_t$ для всех $t \in T$;

Число тем $|T|$ оптимизируется по максимуму среднего cons_t

Возможно ли (почти) обойтись без экспертов?

1. строим по коллекции D_0 тематическую модель $\Phi_0\Theta_0$
2. находим монотематичные фразы (предложения, абзацы)
3. эксперты проверяют, что темы и фразы интерпретируемые
4. убирая из них стоп-слова, получаем размеченные цепочки C_{dk}
5. строим по коллекции D_0 много моделей $\{\Phi_i\Theta_i\}$
6. ищем критерий, коррелирующий с согласованностью cons_t



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.

Напоминание. 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 $p(w|t)$
- $W_t = \{w\colon p(w|t) > p(w)\}$
- $W_t = \{w\colon p(w|t) > \frac{1}{|W|}\}$ [Кольцов и др., 2014]
- $W_t = \{w\colon p(t|w) > 0.25\}$ [Воронцов, Потапенко, 2014]

Характеристики лексического ядра темы:

- $|W_t|$ — *размер ядра* темы, ориентировочно $|W_t| \sim \frac{|W|}{|T|}$
- $\sum\limits_{w\in W_t} p(w|t)$ — *чистота* темы, из $[0,1]$, лучше больше
- $\frac{1}{|W_t|}\sum\limits_{w\in W_t} p(t|w)$ — *контрастность* темы, $[0,1]$, лучше больше
- $\frac{1}{|W_t|}\sum\limits_{w\in W_t} \log\frac{p(w|t)}{p(w)}$ — 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 (тем выше, чем больше кросс-энтропия; min при распределении $p(t)$ или $\frac{1}{|T|}$, max при вырожденном):

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

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 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:

$$S_d = \mathrm{KL}\left(\hat{p}(w|d) \,\middle\|\, p(w|d)\right) = \sum_w \hat{p}(w|d) \ln \frac{\hat{p}(w|d)}{p(w|d)} =$$

$$= \operatorname*{avg}_w \left(n_{dw}, \ln \frac{\hat{p}(w|d)}{p(w|d)} \right) = \operatorname*{avg}_{w,t} \left(n_{tdw}, \ln \frac{\hat{p}(w|d)}{p(w|d)} \right),$$

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.

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.П.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:

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

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:

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

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

Возможное применение меры семантической неоднородности S_t:

- выявление тем для замены или разбиения на подтемы

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

Семейство средневзвешенных статистик с функцией потерь ℓ_{dw}

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

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

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

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

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

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

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

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

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

$\ln \mathscr{P}_{td} = \text{avg}_{w}\left(n_{tdw}, \ell_{dw}\right)$ — 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_d, 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 $\ell_{dw} = \left[p(w|d) < \frac{\alpha}{n_d}\right]$
  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:

- 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

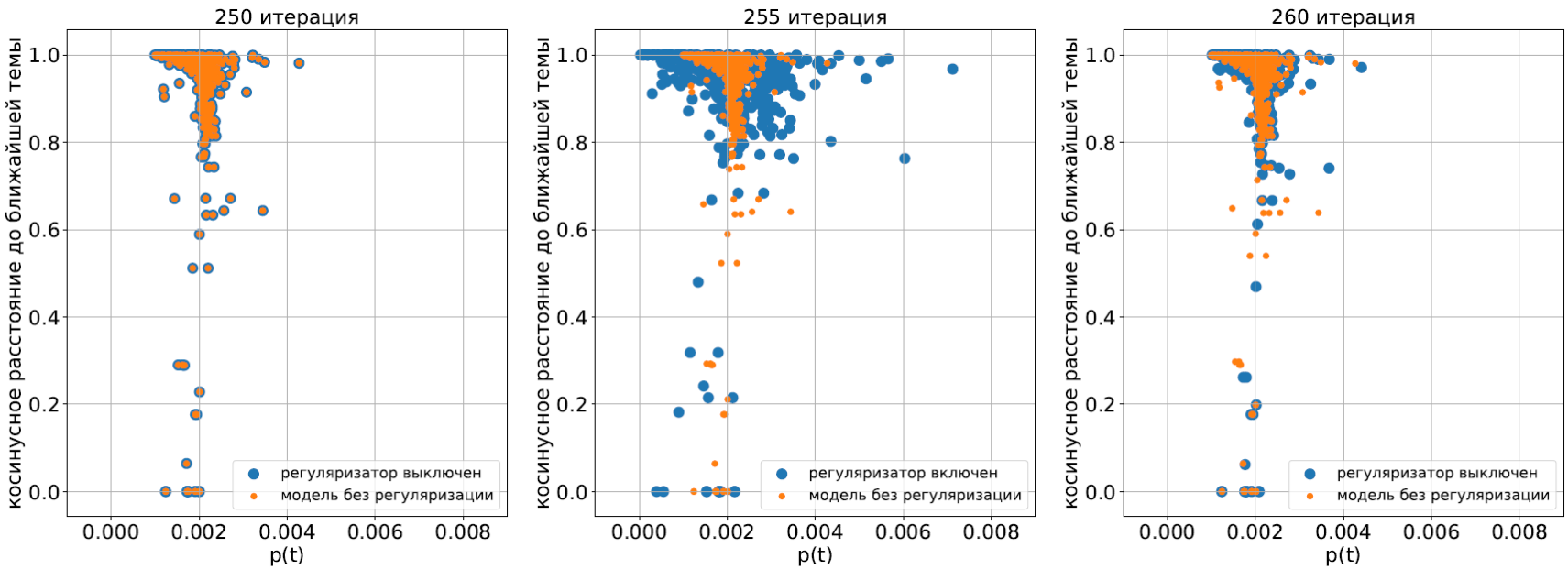
Нормальные значения статистик:

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

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 тем и тем-дубликатов

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

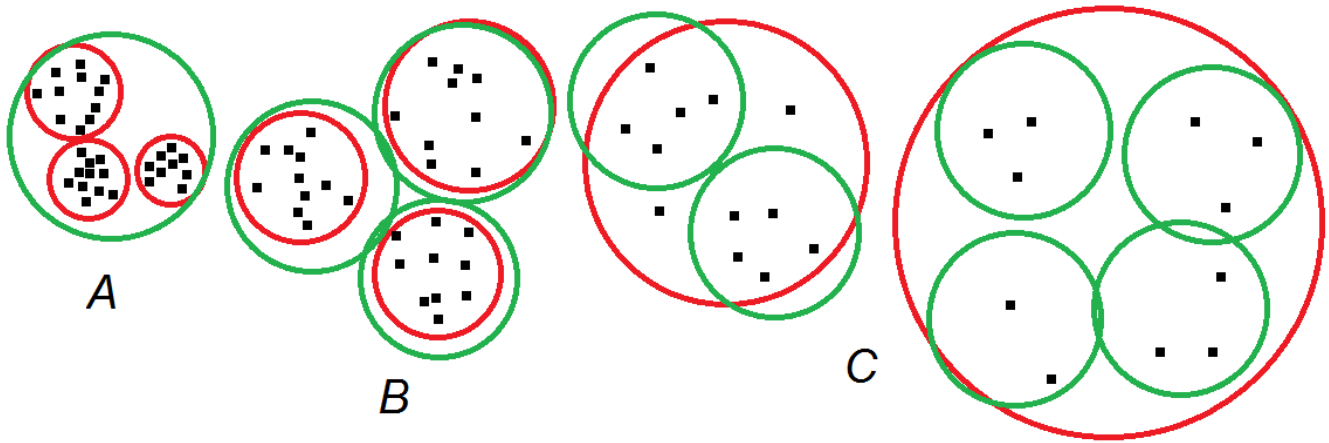
- регуляризатор декоррелирования удаляет дубликаты,
- усиливает разброс тем по их мощности $p(t)$,
- после отключения регуляризатора эти эффекты остаются.

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 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 ($R = 0$)

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \sum_{d,w}\sum_{t\in T} n_{dw}\, p_{tdw}\, Z_t \ln\left(\phi_{wt}\theta_{td}\right) \;\to\; \max_{\Phi,\Theta}$$

при любых Z_t не меняет формулу M-шага для матрицы Φ.

Доказательство. По лемме о максимизации на симплексах:

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}\, p_{tdw}\, Z_t\right) = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}\, p_{tdw}\right)$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw}\, p_{tdw}\, Z_t\right)$$

■

Нормировка Z_t влияет на Θ, но не влияет на Φ и темы $p(w|t)$

E.Veselova, K.Vorontsov. Topic balancing with additive regularization of topic models. 2020

Регуляризатор семантической однородности

Минимизация суммарной семантической неоднородности тем:

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

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:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \qquad \beta_{dw} = \textstyle\sum_t \gamma S_t^{\gamma-1} \frac{p_{tdw}}{p_t}$$

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

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

Резюме

- 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)

Задания по курсу

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

Задания к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Биграммная модель коллекции: $p(w|v) = \xi_{wv}$,
где v — слово, идущее в тексте перед w.
Найти параметры модели ξ_{wv}.

2. Биграммная модель документов: $p(w|v,d) = \xi_{dwv}$.
Найти параметры модели ξ_{dwv}.

Подсказка: применить условия ККТ или основную лемму.

3*. Творческое задание (возможны разные решения).
Предложите модель, разделяющую 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.

Задания к лекции 2

4. Пользуясь основной леммой, докажите, что регуляризатор битермов эквивалентен добавлению псевдодокументов d_u в исходную коллекцию (см. слайд 13)

Прикладная исследовательская задача:
автоматическое выделение научных терминов (ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  метод ATE на основе комбинирования ARTM и TopMine;
  обоснование, что синтаксический анализ не нужен;
  зависимость качества ATE от объёма коллекции
- Критерий:
  качество ATE (Prec, Rec, F1) на размеченных данных

Теоретическое задание к лекции 3

Выведете EM-алгоритм для тематической языковой модели:

5. $p(w|d) = \sum_t \phi_{wt}\theta_{td}$, используя в качестве исходных данных последовательность $(d_i, w_i)_{i=1}^n$ вместо счётчиков n_{dw}.
Докажите эквивалентность обычному EM-алгоритму ARTM.

6. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$, где $p(t)$ фиксировано,
$\phi_{tw} = p(t|w)$, $\theta_{td} = p(t|d)$ — параметры модели.

7. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$, где $p(t)$ фиксировано,
$\phi_{tw} = p(t|w)$ — параметры модели, $\theta_{td} = \sum_w \frac{n_{dw}}{n_d}\phi_{tw}$.

8*. Фиксация $p(t)$ как внешнего параметра упрощает выкладки, но может нарушать условия целостности модели:

$$p(t) = \sum_t \phi_{tw}p(w), \qquad p(t) = \sum_t \theta_{td}p(d).$$

Как обеспечить выполнение этих условий в EM-алгоритме?

Теоретическое задание к лекции 4

9. Докажите, что необходимым условием максимума

$$\sum_{i=1}^{n} \ln \sum_{t \in T} p(w_i, t|i, \Omega) \;\to\; \max_{\Omega}$$

для языковой модели со скрытыми переменными $t \in T$ (не обязательно темами) и параметрами $\Omega = (\omega_{kj})$ — набором неотрицательных нормированных векторов, является система

$$\text{E-шаг:} \quad p(t|w_i, i) = \operatorname*{norm}_{t \in T} p(w_i, t|i, \Omega)$$

$$\text{M-шаг:} \quad \omega_{kj} = \operatorname*{norm}_{k} \left(\sum_{i=1}^{n} \sum_{t \in T} p(t|w_i, i)\, \omega_{kj} \frac{\partial}{\partial \omega_{kj}} \ln p(w_i, t|i, \Omega) \right)$$

10. Выведете отсюда EM-алгоритм для частных случаев:

1) $p(w, t|i, \Omega) = \phi_{wt}\theta_{td_i}$;
2) $p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{w \in d_i} \frac{n_{d_i w}}{n_{d_i}} \phi_{tw}$;
3) $p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c}$.

11. Творческое задание.** Предложите способ ввести обучаемые параметры в тематическую модель внимания.

Исследовательская задача к лекции 4

Реализуйте EM-алгоритм для модели локального контекста (или воспользуйтесь готовой реализацией)

Исследуйте зависимость метрик качества модели

- перплексия: $\mathscr{P} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} p(w|C_i)\right)$
- разреженность, различность, когерентность тем
- дефекты целостности модели:
  $\left\|p(t) - \frac{n_t}{n}\right\|$, $\left\|p(t) - \sum_t \phi_{tw} p(w)\right\|$, $\left\|p(t) - \sum_t \theta_{td} p(d)\right\|$

от номера итерации и от параметров модели:

- $|T|$ — число тем
- L — число проходов
- τ — вес N_{tw} в формуле M-шага, особый случай $\tau = 0$
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, β — баланс левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, секций
- опция «$i \in C_i$ или $i \notin C_i$»

Теоретическое задание к лекции 5

12. Найдите дискретное распределение $P = (p_i)_{i=1}^n$ в задаче $\sum_i n_i \mu(p_i) \to \max$ с гладкой монотонно возрастающей $\mu(p)$.
Отдельно рассмотрите случаи $\mu(p) = p^s$, $s = 1$, $s \to 0$.

13. Выведете EM-алгоритм в случае, когда ln заменён гладкой монотонно возрастающей функцией μ:

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

Подумайте, какие замены логарифма полезны, и почему.

14. Простейшая идея разреживания — обнуление малых вероятностей. Чтобы обосновать эту эвристику, найдите, какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \operatorname*{norm}_w \left(n_{wt} [n_{wt} > \gamma n_t] \right)$$

Подсказка: с учётом подстановки несмещённой оценки ϕ^*_{wt}

Теоретическое задание к лекции 5

Проект «Тематизатор». Аналитик построил модель $\Phi^0\Theta^0$ 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 $t \in T_-$.

15. Предложите регуляризаторы для этого.

16. 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?

17. Предложите способ инициализации Φ для новой модели.

Исследовательская задача к лекции 5

Продолжение исследования по автоматическому выделению научных терминов (Automatic Term Extraction, ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  — оптимальную стратегию регуляризации на основе декоррелирования и сглаживания фоновых тем
  — рекомендации по управлению относительными коэффициентами регуляризации
  — критерий тематичности терминов по расстоянию между распределениями $p(t|w)$ и $p_0(t) = \frac{1}{|T|}$, позволяющий наиболее чётко отличать термины от фоновой лексики
- Критерий:
  — максимум доли терминов в предметных темах
  — минимум доли терминов в фоновых темах

Исследовательская задача к лекции 5

Продолжение исследования модели локального контекста (можно воспользоваться готовой реализацией EM-алгоритма)

Исследуйте устойчивость модели в сравнении с ARTM

- без регуляризации
- с регуляризатором декоррелирования, при различных значениях относительного коэффициента регуляризации

Как на устойчивость модели влияют её параметры:

- $|T|$ — число тем
- L — число проходов
- τ — вес N_{tw} в формуле M-шага, особый случай $\tau = 0$
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, β — баланс левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, секций
- опция «$i \in C_i$ или $i \notin C_i$»

Теоретическое задание к лекции 6

18. 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 М-шага для матрицы Ψ.

19. Предложите способ гарантировать, что если родительская тема t получает только одну дочернюю s, то она переходит в неё целиком и как распределение: $p(w|s) = p(w|t)$, то есть тема t на данном уровне не расщепляется на подтемы.

20. 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 6

Проект «Мастерская знаний». Нужна тематическая модель подборок научных статей и/или поисковой выдачи.

Дано:

- 1000 подборок, в каждой по 1000 аннотаций научных статей, ранжированные по сходству с аннотацией-запросом по эмбедингам модели SciRus (эмбединги тоже даны)

Найти:

- метод согласования тематической модели с эмбедингами
- метод выделения терминов (Automatic Term Extraction)
- метод отбора терминов по тематичности
- метод отсева тематически нерелевантных аннотаций

Критерии:

- согласованность тематической модели с эмбедингами
- интерпретируемость тем
- качество выделения терминов

Теоретическое задание к лекции 7

21. Выведите EM-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 M-шаг эквивалентна введению мультипликативной поправки $(1 + \tau\beta_{dw})$ в критерий log-правдоподобия.

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

23*. Предложите формулу средневзешенных статистик S_* для тематической модели локальных контекстов.
Проверьте, что полученная формула совпадает с введённой на лекции, если контекстом является весь документ.

Исследовательская задача к лекции 7

Исследование EM-алгоритма для модели локального контекста

- Оценивание внутритекстовой когерентности
  - реализуйте вычисление средневзвешенной когерентности
  - подберите наилучшее сочетание эвристик rel и coh в калибровочном эксперименте без экспертной разметки
  - какие параметры модели локального контекста улучшают внутритекстовую когерентность?
  - насколько это улучшение устойчиво для разных коллекций?
- Оценивание средневзвешенных статистик
  - реализуйте вычисление S_t, S_{wt}
  - как зависит вид распределения $\{S_t\}$ от числа тем?
  - есть ли корреляция между S_t и когерентностью coh_t?
  - предложите способ разделения темы с большим S_t на подтемы и их инициализацию термами с большими S_{wt}
- Оценивание несбалансированности тем
  - реализуйте генератор коллекций с заданным дисбалансом тем
  - как дисбаланс влияет на число разделённых и слитых тем?
  - модели локального контекста лишены этой проблемы?
  - уменьшает ли регуляризатор семантической однородности число разделённых и слитых тем?

Примеры датасетов для практических заданий по курсу

1. Открытые датасеты (английский): 20NG, NIPS, KOS
2. Ранжированные результаты поиска научных статей (по данным eLibrary, arXiv, PubMed)
3. Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
4. Техноблоги: Хабр (русский), TechCrunch (английский)
5. Данные социальных сетей: VK, Twitter, Telegram,...
6. Статьи по Complexity Sciences (для хронокарты науки)

- Википедия
- Викиновости (1.5M статей, проект закрыт 30/03/2026)
- 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

- «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: кол./кач. анализ предметной области,
  — реализация данного сценария как модуля в среде Orange

- «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: помочь в понимании предметной области

Открытые проблемы тематического моделирования

1. Тематические модели внимания последовательного текста
2. Проблема несбалансированности тем в коллекции
3. Измерение интерпретируемости тем (когерентность)
4. Обеспечение 100%-й интерпретируемости тем
5. Автоматическое именование и суммаризация тем
6. Калибровка моделей тематической фильтрации
7. Согласование тем с предобученными эбмедингами LLM
8. Статистические оценки состоятельности тем
9. Обнаружение новых тем или трендов в потоке текстов
10. Обеспечение устойчивости и полноты множества тем
11. Автоматический подбор гиперпараметров, AutoML
12. Гиперграфовые тематические модели для RecSys