

# Вероятностные тематические модели

## Лекция 5. Оценивание качества тематических моделей

К. В. Воронцов  
vokov@forecsys.ru

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

ВМК МГУ • 21 марта 2023

- 1 Измерение качества тематических моделей**
  - Правдоподобие и перплексия
  - Интерпретируемость и когерентность
  - Разреженность и различность
- 2 Многокритериальное оценивание моделей**
  - Разреживание, сглаживание, декоррелирование
  - Эксперименты с комбинированием регуляризаторов
  - Многокритериальная оптимизация гиперпараметров
- 3 Проверка гипотезы условной независимости**
  - Статистики на основе KL-дивергенции и их обобщения
  - Применения оценок семантической однородности
  - Регуляризатор семантической однородности

## 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:

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

Задача ВТМ по природе своей многокритериальная:

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

## 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  $\Phi$  и  $\Theta$

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

## Правдоподобие и перплексия (perplexity)

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

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

*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 тестового документа на две половины равной длины;

параметры  $\phi_{wt}$  оцениваются по обучающей коллекции  $D$ ;

параметры  $\theta_{td}$  оцениваются по первой половине  $d'$ ;

перплексия вычисляется по второй половине  $d''$ .

**Проблема:** как разбивать документ на две половины?

## Измерение интерпретируемости тем

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

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

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

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

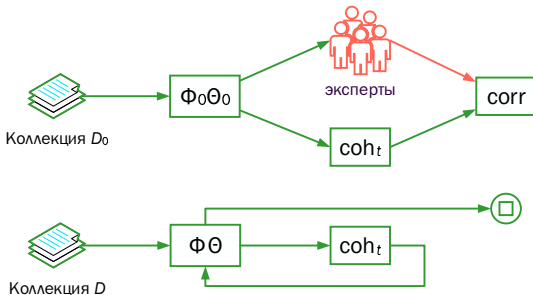
---

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

## Схема эксперимента Ньюмана

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

На новой коллекции  $D$  используем откалиброванный критерий (когерентность тем  $\text{coh}_t$ ) для оценивания и выбора моделей  $\Phi\Theta$





## Эксперимент. Связь когерентности и интерпретируемости

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

PMI — лучшая метрика.

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

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	-0.20	0.19
	LCH	-0.31	-0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	-0.08	0.27
	WuP	0.41	0.26
	Wikipedia	RACO	0.62
MiW		0.68	0.70
DOC SIM		0.59	0.60
<b>PMI</b>		<b>0.74</b>	<b>0.77</b>
Google	TITLES	0.51	
	LOGHITS	-0.19	
Gold-standard	IAA	<b>0.82</b>	<b>0.78</b>

**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  $\phi_{wt}$ ,

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

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

$P_u$  — доля документов, в которых  $u$  встретился хотя бы 1 раз,  $P_{uv}, P_u$  можно вычислять по другой коллекции (Википедии).

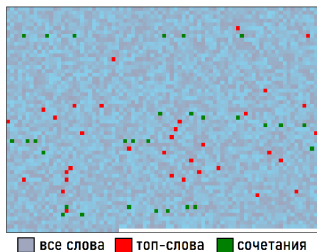
Когерентность модели = средняя когерентность всех тем.

---

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

## Недостаток когерентности

Обычно берут  $k = 10..20$  топовых (наиболее частотных) слов, но они занимают лишь 1–2% текста совместно по всем темам, а пары с большим  $N_{uv}$  образуются из топовых слов ещё реже! Более 99% текста игнорируется оценкой когерентности модели, и «золотой стандарт» Ньюмана страдает тем же недостатком!



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

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:

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

$\text{coh}(u, v)$  — сочетаемость пары слов  $u, v$  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

### Возможные модификации:

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

**Проблема:** большой объем вычислений по всем парам слов

## Внутритекстовая когерентность

### Средневзвешенная когерентность темы:

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

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

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

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

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

---

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

## Как проверить адекватность внутритекстовой когерентности

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

### Идея:

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

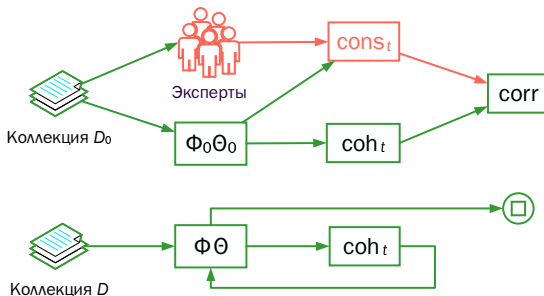
### Пример разметки:

транспорт    психология    общенаучная лексика    общеупотребительная лексика

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

## Схема калибровки внутритекстовой когерентности

- 1 выбираем коллекцию и фрагменты для разметки
- 2 эксперты размечают тематические цепочки во фрагментах
- 3 строим тематическую модель (можно 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  $t$ :

$$C(t) = \{C_{di} : t = \arg \max_t p(t|C_{di})\}$$

*Мера согласованности темы с размеченными цепочками:*

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

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

$$\operatorname{diff} = \frac{\sum_d \#\{t : C_{di} \in C(t)\}}{\sum_d |C_d|}$$



## Критерии разреженности матриц $\Phi$ и $\Theta$

*Разреженность* — доля нулевых элементов в  $\Phi$  и  $\Theta$

Однако  $\phi_{wt}$  и  $\theta_{td}$  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  $\Phi$  и  $\Theta$  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 К  $\rightarrow$  8 К

---

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

## Лексическое ядро, чистота и контрастность темы

Лексическое ядро  $W_t$  темы  $t$ , варианты определения:

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

Характеристики лексического ядра темы:

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

---

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

## Критерии различности тем

Среднее расстояние от темы  $t$  до ближайшей к ней темы

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

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

Дивергенции — несимметричные меры «вложенности»  $\phi_t$  в  $\phi_s$ :

- $\rho(\phi_t, \phi_s) = \sum_w \phi_{wt} \ln\left(\frac{\phi_{wt}}{\phi_{ws}}\right)$  — 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* (чем выше кросс-энтропия, тем тематичнее):

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

*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  $D$

## Напоминание. Регуляризаторы сглаживания и разреживания

Общий вид регуляризаторов сглаживания и разреживания:

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

где  $\beta_0 > 0$ ,  $\alpha_0 > 0$  — коэффициенты регуляризации,

$\beta_{wt}$ ,  $\alpha_{td}$  — параметры, задаваемые пользователем:

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

**Возможные применения** сглаживания и разреживания:

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

## Напоминание. Регуляризатор декоррелирования тем

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

Минимизируем ковариации между вектор-столбцами  $\phi_t$ :

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

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

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

## Напоминание. Разреживающий регуляризатор для отбора тем

**Цель:** избавиться от незначимых тем (topic selection).

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

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

Подставляем, получаем:

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

**Эффект:** обнуляются строки матрицы  $\Theta$  с малыми  $n_t$ , заодно (неожиданно) удаляются зависимые и расщеплённые темы.

---

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

## Разреживание, сглаживание, декоррелирование, отбор тем

M-шаг при комбинировании  $b$  регуляризаторов:

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

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

**Данные:** статьи NIPS (Neural Information Processing System)

$|D| = 1566$  статей,  $n = 2.3$  М,  $|W| = 13$  К,

контрольная коллекция:  $|D'| = 174$ .

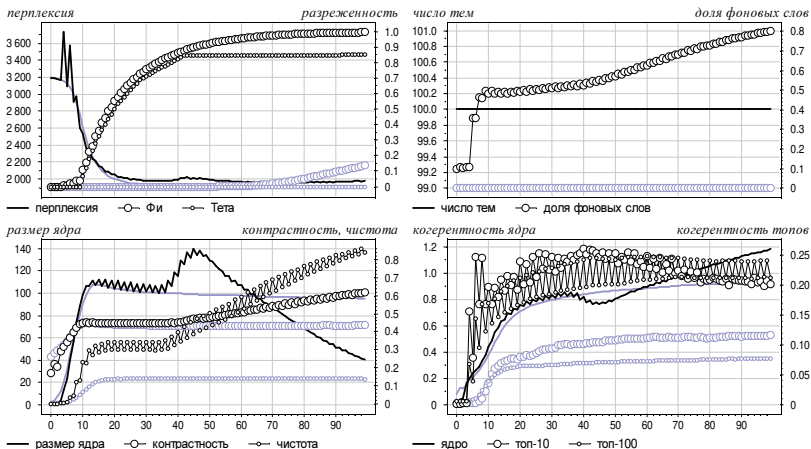
---

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



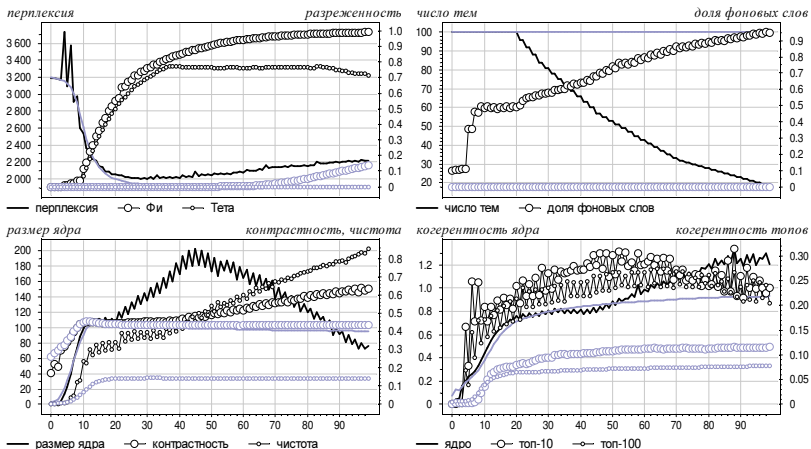
## Разреживание, сглаживание, декоррелирование

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



## Те же регуляризаторы, плюс отбор тем

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



## Выводы по результатам экспериментов

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

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

**Рекомендации по выбору траектории регуляризации:**

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

## Постановка задачи планирования экспериментов (AutoML)

$\Lambda = (\lambda_1, \dots, \lambda_K)$  — вектор гиперпараметров ( $|T|$ ,  $\tau_i$ ,  $\tau_m$  и др.)

$Q = (q_1, \dots, q_M)$  — вектор критериев качества модели

### Гипотезы:

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

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

---

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

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

## Гипотеза условной независимости

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

**Гипотеза семантической однородности темы  $t$**

— в теме  $t$  термины 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  $t$**

— термины темы  $t$  порождаются независимо от документов:

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

**Гипотеза согласованности термина  $w$  с темой  $t$**

— тема  $t$  распределена по документам независимо от терминов:

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

## Мера семантической неоднородности темы $t$ в коллекции

Статистика для проверки гипотезы  $H_0(t)$ :

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

По определению условной вероятности и формуле Е-шага:

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

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

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  $\gamma_i$

## 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:

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

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

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:

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

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

Возможные применения меры несогласованности  $S_{wt}$ :

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



## Средневзвешенные статистики с произвольной функцией потерь

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

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

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

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

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

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

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

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

$\ln \mathcal{P}_{td} = \text{avg}_{w \in d}(n_{tdw}, \ell(d, w))$  — 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  $n_{dw} \rightarrow 1$ ,  $n_{tdw} \rightarrow p_{tdw}$
- бинарная функция потерь  $\ell(d, w) = [p(w|d) < \frac{\alpha}{n_d}]$   
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:

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

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

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

### Модифицированный EM-алгоритм

$$p_{tdw} = \operatorname{norm}_{t \in T} (\phi_{wt} \theta_{td})$$

$$\beta_{dw} = \sum_t \frac{p_{tdw}}{p_t}$$

$$\phi_{wt} = \operatorname{norm}_{w \in W} \left( \sum_d \tilde{n}_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$$

$$\tilde{n}_{dw} = n_{dw} (1 + \tau \beta_{dw})$$

$$\theta_{td} = \operatorname{norm}_{t \in T} \left( \sum_w \tilde{n}_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$$

$$p_t = \frac{1}{n} \sum_{dw} n_{dw} p_{tdw}$$

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

### Открытые проблемы

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