

Вероятностные тематические модели

Лекция 3. Оценивание качества тематических моделей

К. В. Воронцов
vokov@forecsys.ru

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

МФТИ – ФИЦ ИУ РАН • 23 сентября 2020

- 1 Измерение качества тематических моделей**
 - Правдоподобие и перплексия
 - Интерпретируемость и когерентность
 - Разреженность и различность
- 2 Эксперименты с аддитивной регуляризацией**
 - Комбинирование регуляризаторов
 - Проблема определения числа тем
 - Проблема несбалансированности тем
- 3 Проверка гипотезы условной независимости**
 - Статистики на основе KL-дивергенции и их обобщения
 - Регуляризатор семантической однородности
 - Применения оценок семантической однородности

Напоминания. Задача тематического моделирования

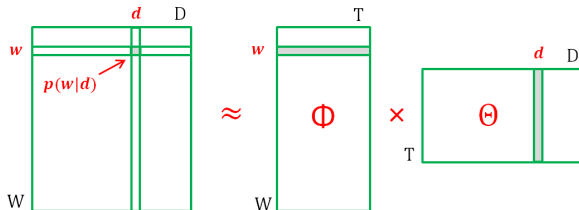
Дано: коллекция текстовых документов, $p(w|d) = \frac{n_{dw}}{n_d}$

Вероятностная тематическая модель:

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

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

Это задача стохастического матричного разложения:



Максимизация \log правдоподобия с регуляризатором R :

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

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

где $\mathop{\text{norm}}_{t \in T}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$ — операция нормирования вектора.

Максимизация \log правдоподобия с k регуляризаторами R_i :

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + \sum_{i=1}^k \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

где τ_i — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \sum_{i=1}^k \tau_i \frac{\partial R_i}{\partial \phi_{wt}} \right) \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} + \theta_{td} \sum_{i=1}^k \tau_i \frac{\partial R_i}{\partial \theta_{td}} \right) \end{cases} \end{cases}$$

Критерии качества тематических моделей

Внешние критерии:

- Полнота и точность тематического поиска
- Качество ранжирования при тематическом поиске
- Качество классификации / категоризации документов
- Качество суммаризации / сегментации документов
- Экспертные оценки качества тем

Внутренние критерии:

- Правдоподобие и перплексия
- Средняя когерентность (согласованность) тем
- Разреженность матриц Φ и Θ
- Различность 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 различности или неопределённости слов в тексте
- коэффициент ветвления (branching factor) текста

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 ϕ_{wt} оцениваются по обучающей коллекции D ;

параметры θ_{td} оцениваются по первой половине d' ;

перплексия вычисляется по второй половине d'' .

Интерпретируемости и когерентность

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 его при определении

Нужна автоматически вычисляемая мера интерпретируемости, коррелирующая с экспертными оценками.

Ею оказалась *когерентность* (согласованность, coherence).

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

Эксперимент. Связь когерентности и интерпретируемости

Измерялась ранговая корреляция Спирмена между 15 метрикам и экспертными оценками интерпретируемости.

PMI — лучшая метрика.

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

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	-0.20	0.19
	LCH	-0.31	-0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	-0.08	0.27
	WuP	0.41	0.26
Wikipedia	RACO	0.62	0.69
	MiW	0.68	0.70
	DOC SIM	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 внутренняя мера интерпретируемости

Когерентность (согласованность) темы t по k топовым словам:

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

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

$\text{PMI}(u, v) = \ln \frac{|D|N_{uv}}{N_u N_v}$ — поточечная взаимная информация (pointwise mutual information),

N_{uv} — число документов, в которых слова u, v хотя бы один раз встречаются рядом (в окне 10 слов),

N_u — число документов, в которых u встретился хотя бы 1 раз.

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

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 K \rightarrow 8 K

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

Лексическое ядро, чистота и контрастность темы

Лексическое ядро W_t темы t , варианты определения:

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

Характеристики лексического ядра темы:

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

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

Критерии различности тем

Среднее расстояние от темы t до ближайшей к ней темы

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

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

Напоминания. 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,

β_{wt} , α_{td} — параметры, задаваемые пользователем:

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

Возможные применения сглаживания и разреживания:

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

Напоминания. Регуляризатор декоррелирования тем

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

Минимизируем ковариации между вектор-столбцами ϕ_t :

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

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

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

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 Θ с малыми n_t , заодно (неожиданно) удаляются зависимые и расщеплённые темы.

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

Разреживание, сглаживание, декоррелирование, отбор тем

M-шаг при комбинировании 6 регуляризаторов:

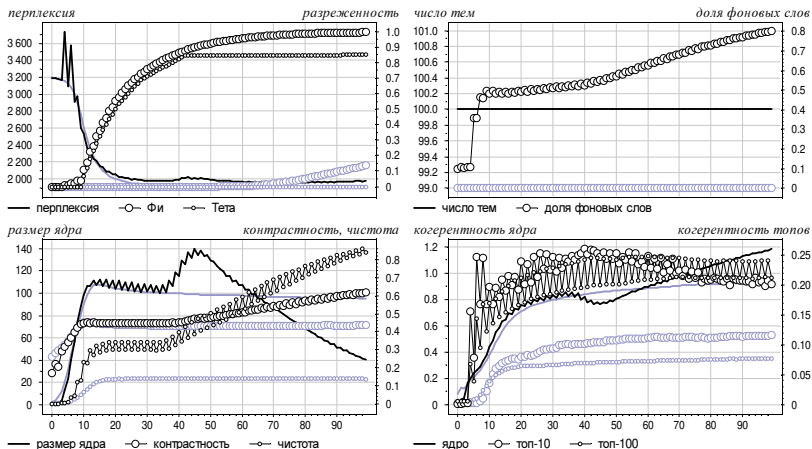
$$\phi_{wt} = \text{norm}_w \left(n_{wt} + \tau_1 \underbrace{\beta_w[t \in B]}_{\substack{\text{сглаживание} \\ \text{фоновых} \\ \text{тем}}} - \tau_2 \underbrace{\beta_w[t \in S]}_{\substack{\text{разреживание} \\ \text{предметных} \\ \text{тем}}} - \tau_3 \underbrace{\phi_{wt} \sum_{s \in S \setminus t} \phi_{ws}}_{\text{декоррелирование}} \right)$$
$$\theta_{td} = \text{norm}_t \left(n_{td} + \tau_4 \underbrace{\alpha_t[t \in B]}_{\substack{\text{сглаживание} \\ \text{фоновых} \\ \text{тем}}} - \tau_5 \underbrace{\alpha_t[t \in S]}_{\substack{\text{разреживание} \\ \text{предметных} \\ \text{тем}}} - \tau_6 \underbrace{\frac{n_d}{n_t} \theta_{td}}_{\text{удаление} \\ \text{малых тем}} \right)$$

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.

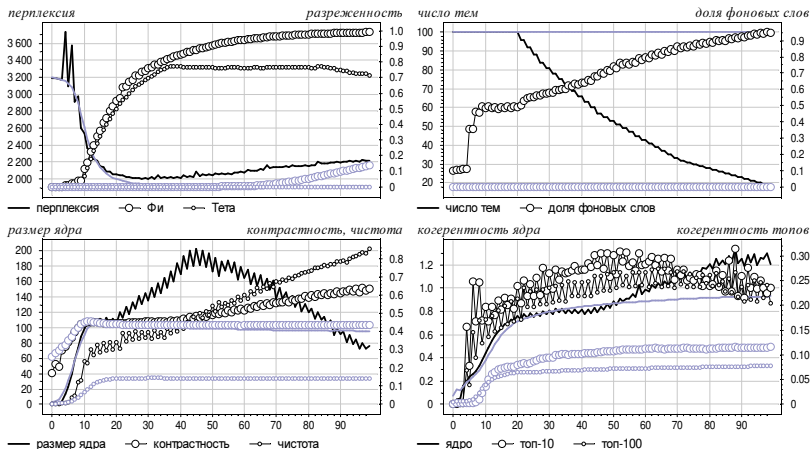
Разреживание, сглаживание, декоррелирование

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 почти без потери перплексии (правдоподобия) модели:

- *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

Эксперименты с регуляризатором отбора тем

Коллекция статей NIPS (Neural Information Processing System)

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

Синтетическая коллекция:

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

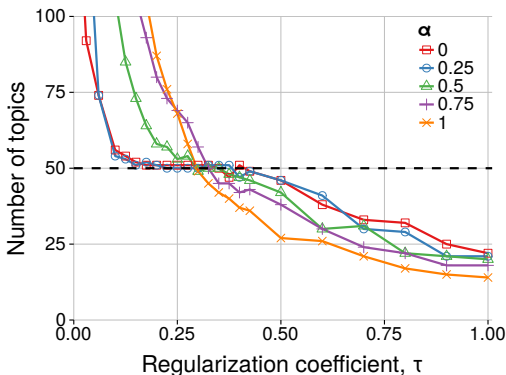
$$n_{dw}^0 = n_d \sum_{t \in T_0} \phi_{wt} \theta_{td}$$

Параметрическое семейство полусинтетических данных:

- n_{dw}^α — смесь синтетических данных n_{dw}^0 и реальных n_{dw} :

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

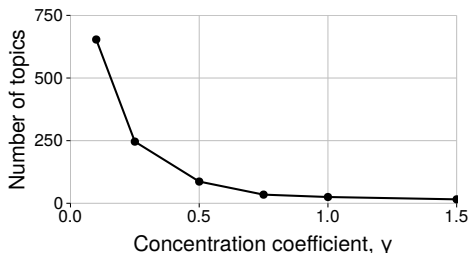
Попытка определения числа тем



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

Сравнение с байесовской тематической моделью HDP

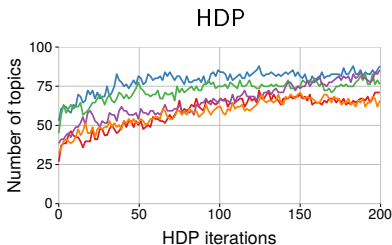
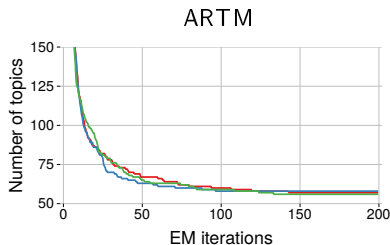
HDP, Hierarchical Dirichlet Process [Teh et.al, 2006] —
«state-of-the-art» байесовский подход к определению числа тем



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

Сравнение ARTM и HDP по устойчивости

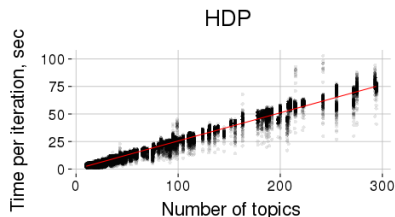
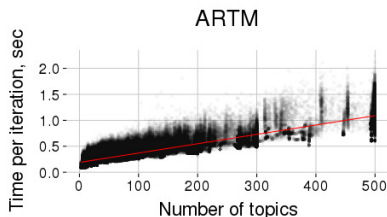
Запуск ARTM и HDP много раз из случайных инициализаций:



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

Сравнение ARTM и HDP по времени вычислений

Сравнение времени одного прохода коллекции (sec)

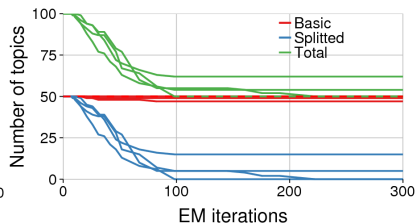
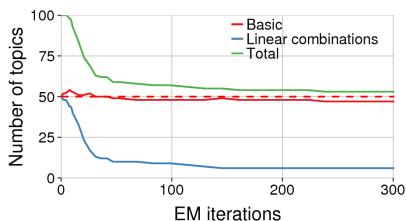


- ARTM в 100 раз быстрее!

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

Удаление линейно зависимых и расщеплённых тем

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



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

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

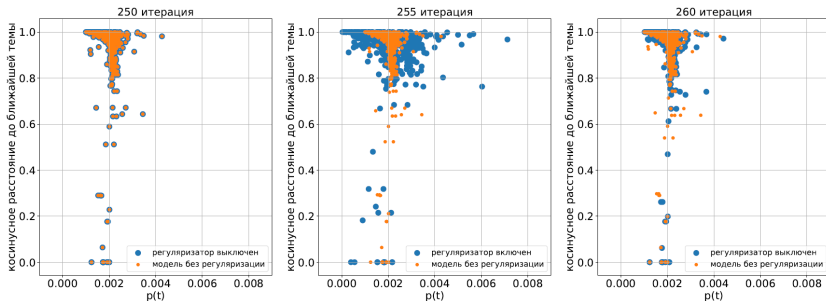
Выводы по результатам экспериментов

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

Проблема малых тем и тем-дубликатов

Эксперимент на коллекции postnauka.ru

- Самой модели не выгодно производить малые темы!
- Регуляризатор отбора тем плохо устраняет дубликаты!

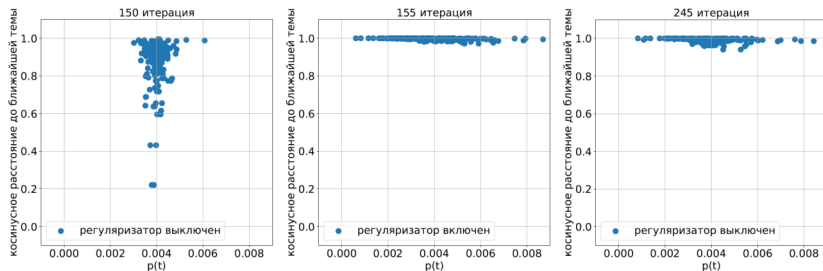


Г.Фоминская. Выявление тем-дубликатов в тематических моделях.
Курсовая работа, ВМК МГУ, 2018.

Проблема малых тем и тем-дубликатов

Эксперимент на коллекции postnauka.ru

- Регуляризатор декоррелирования удаляет дубликаты лучше!
- Заодно он усиливает разброс тем по их мощности $p(t)$

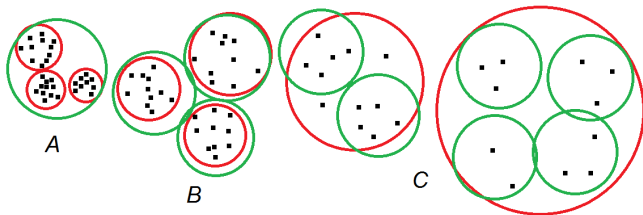


Г.Фоминская. Выявление тем-дубликатов в тематических моделях.
Курсовая работа, ВМК МГУ, 2018.

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 мощностям (красные кластеры).
- Это приводит к появлению тем-дубликатов (A) и семантически разнородных тем (C).
- Выравнивание тем по *радиусу семантической однородности* (зелёные кластеры) должно решать обе проблемы.



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 термины и документы порождаются независимо:

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

Гипотеза согласованность 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 согласованность 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 γ_i

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

Мера несогласованности документа d с темой t

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

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

По определению условной вероятности и формуле E-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 $\ell(d, w) = \frac{\hat{p}(w|d)}{p(w|d)}$ — 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 $\ell(d, w) = \frac{1}{p(w|d)}$ — *перплексия* (чем меньше, тем лучше):

$\ln \mathcal{P} = \text{avg}_{d,w,t}(n_{tdw}, \ln \frac{1}{p(w|d)}) = \text{avg}_{d,w}(n_{dw}, \ln \frac{1}{p(w|d)})$ — коллекции

$\ln \mathcal{P}_d = \text{avg}_{w,t}(n_{tdw}, \ln \frac{1}{p(w|d)}) = \text{avg}_{w \in d}(n_{dw}, \ln \frac{1}{p(w|d)})$ — документа

$\ln \mathcal{P}_t = \text{avg}_{d,w}(n_{tdw}, \ln \frac{1}{p(w|d)})$ — темы t

$\ln \mathcal{P}_{td} = \text{avg}_{w \in d}(n_{tdw}, \ln \frac{1}{p(w|d)})$ — 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.

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, $\beta_{dw} = \sum_{t \in T} \frac{p_{tdw}}{p_t}$:

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

Модифицированный EM-алгоритм

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

$$\beta_{dw} = \sum_{t \in T} \frac{p_{tdw}}{p_t}$$

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

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

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

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

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

- Регуляризация — стандартный приём для решения некорректно поставленных задач
- ARTM позволяет комбинировать регуляризаторы и строить тематические модели с требуемыми свойствами
- Реализация — в проекте с открытым кодом BigARTM
- Сглаживание + разреживание + декоррелирование — наиболее часто используемая комбинация регуляризаторов
- Другие регуляризаторы — в следующих лекциях

Открытые проблемы

- Несбалансированность тем
- Определение числа тем
- Обнаружение новых тем и их добавление в модель
- Оптимальный выбор траектории регуляризации