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 • 18 сентября 2019

Содержание

Напоминания. Задача тематического моделирования

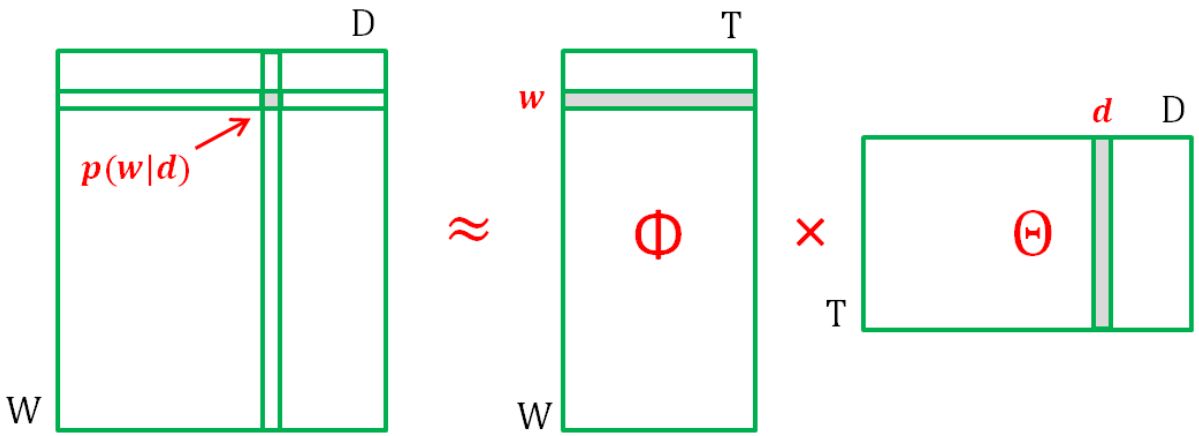
Дано: коллекция текстовых документов, $p(w|d) = \frac{n_{dw}}{n_d}$

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:



Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

Напоминания. ARTM — аддитивная регуляризация

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

где $\underset{t\in T}{\text{norm}}(x_t) = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — операция нормирования вектора.

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014.

Напоминания. Комбинирование регуляризаторов в ARTM

Максимизация log правдоподобия с k регуляризаторами R_i:

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

где τ_i — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Vorontsov K., Potapenko A. Additive regularization of topic models. Machine Learning, 2015.

Напоминания. Регуляризаторы сглаживания и разреживания

Общий вид регуляризаторов сглаживания и разреживания:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T}\sum_{w\in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d\in D}\sum_{t\in T} \alpha_{td} \ln \theta_{td} \to \max,$$

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:

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

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

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

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation. 2010.

Критерии качества тематических моделей

Внешние критерии:

- Полнота и точность тематического поиска
- Качество ранжирования при тематическом поиске
- Качество тематических рекомендаций
- Качество категоризации 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):

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

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 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):

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

$d = d' \sqcup d''$ — случайное разбиение тестового документа на две половины равной длины;

параметры ϕ_{wt} оцениваются по обучающей коллекции D;
параметры θ_{td} оцениваются по первой половине d';
перплексия вычисляется по второй половине d''.

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
	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 внутренняя мера интерпретируемости

Когерентность (согласованность) темы t по k топовым словам:

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

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

$\mathrm{PMI}(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — 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, различности и невырожденности тем

- Разреженность — доля нулевых элементов в Φ и Θ
- Характеристики интерпретируемости тем:
  - размер ядра темы: $|W_t|$, ядро $W_t = \{w : p(t|w) > 0.25\}$
  - чистота темы: $\sum_{w \in W_t} p(w|t)$
  - контрастность темы: $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$
- Вырожденность тематической модели:
  - доля фона в коллекции: $\frac{1}{n} \sum_{d,w} \sum_{t \in B} p(t|d,w)$
  - доля нетематичных документов: $\frac{1}{|D|} \sum_{d \in D} \Big[\sum_{t \in B} p(t|d) > 0.95 \Big]$
  - доля нетематичных термов: $\frac{1}{|W|} \sum_{w \in W} \Big[\sum_{t \in B} p(t|w) > 0.95 \Big]$

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

Робастная тематическая модель

Гипотеза: каждое слово в документе (d, w) является
— либо тематическим, связанным с какой-то темой t,
— либо специфичным для данного документа (шум),
— либо общеупотребительным (фон).

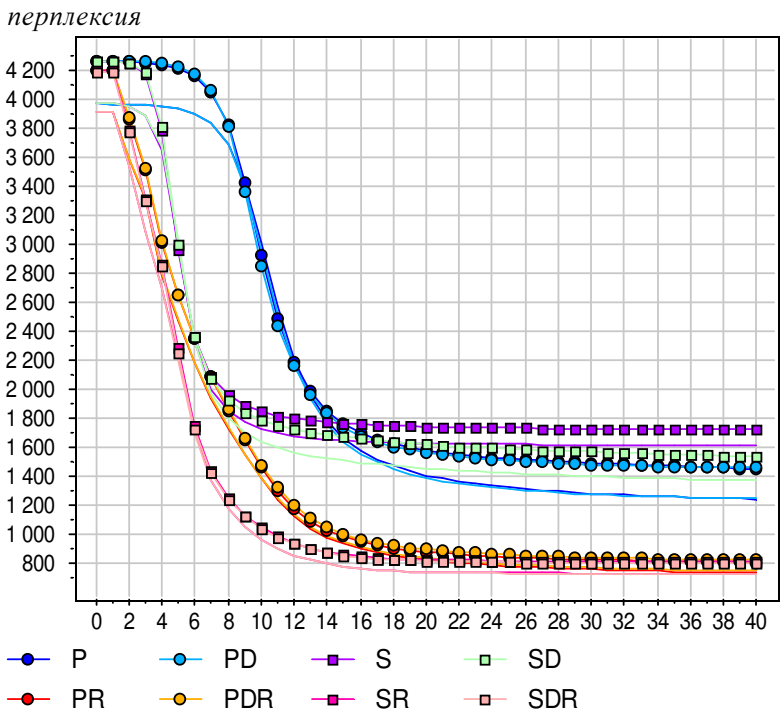
Модель смеси тематической, шумовой и фоновой компонент SWB (Special Words with Background):

$$p(w|d) = \gamma\pi_{dw} + \varepsilon\pi_w + (1 - \gamma - \varepsilon)\sum_{t\in T}\phi_{wt}\theta_{td}$$

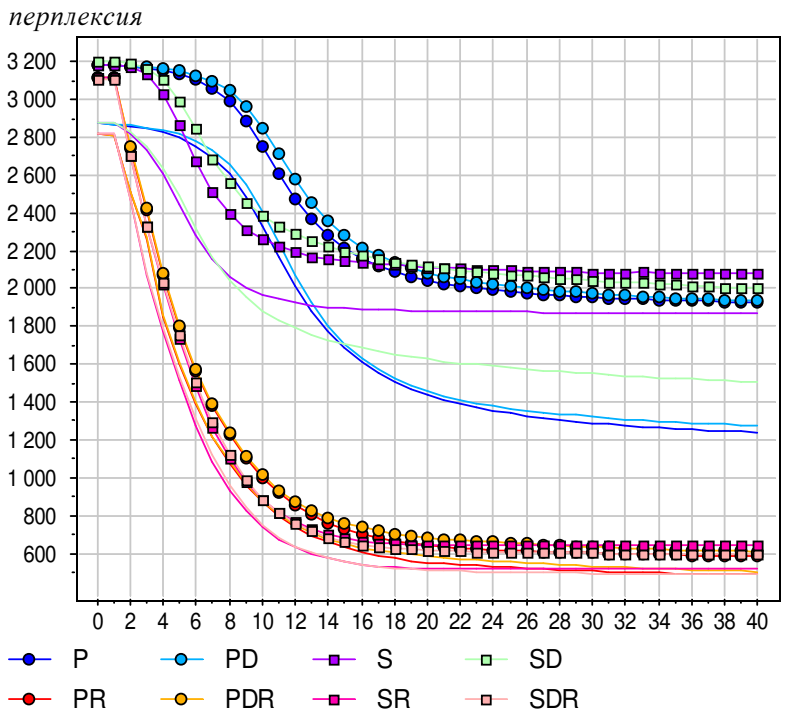
$\pi_{dw} \equiv p_{\text{ш}}(w|d)$ — *шумовая* компонента, γ — параметр;
$\pi_w \equiv p_{\text{ф}}(w)$ — *фоновая* компонента, ε — параметр.

Chemudugunta C., Smyth P., Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model. NIPS, 2006.

Эксперименты с робастными PLSA и LDA



Коллекция RuDis



Коллекция NIPS

Обозначения: P – PLSA, D – LDA ($\alpha_t = 0.5$, $\beta_w = 0.01$)
S – сэмплирование темы из $p(t|d, w)$ для каждого d, w
R – робастность (шум $\gamma = 0.3$, фон $\varepsilon = 0.01$)

A.Potapenko, K.Vorontsov. Robust PLSA performs better than LDA. ECIR-2013.

Выводы

1. Переобучение проявляется только на редких словах
2. LDA точнее моделирует вероятности редких слов
3. Но они как раз не интересны для тематической модели!
4. Робастные PLSA и LDA почти одинаковы по перплексии
5. Робастные PLSA и LDA почти не переобучаются
6. Робастный PLSA лучше, чем обычный LDA
7. Перплексия — не вполне адекватная мера качества

Potapenko A. A., Vorontsov K. V. Robust PLSA performs better than LDA // European Conference on Information Retrieval ECIR-2013.

Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // JMLDA, 2013.

Способны ли PLSA и LDA восстановить истинные темы?

Матрицы Φ_0 и Θ_0 порождаются распределением Дирихле.
Синтетическая коллекция порождается матрицами Φ_0 и Θ_0.
Размеры: $|D| = 500$, $|W| = 1000$, $|T| = 30$, $n_d \in [100, 600]$.

Цель — сравнить восстановленные распределения $p(i|j)$
с исходными синтетическими распределениями $p_0(i|j)$
по среднему расстоянию Хеллингера:

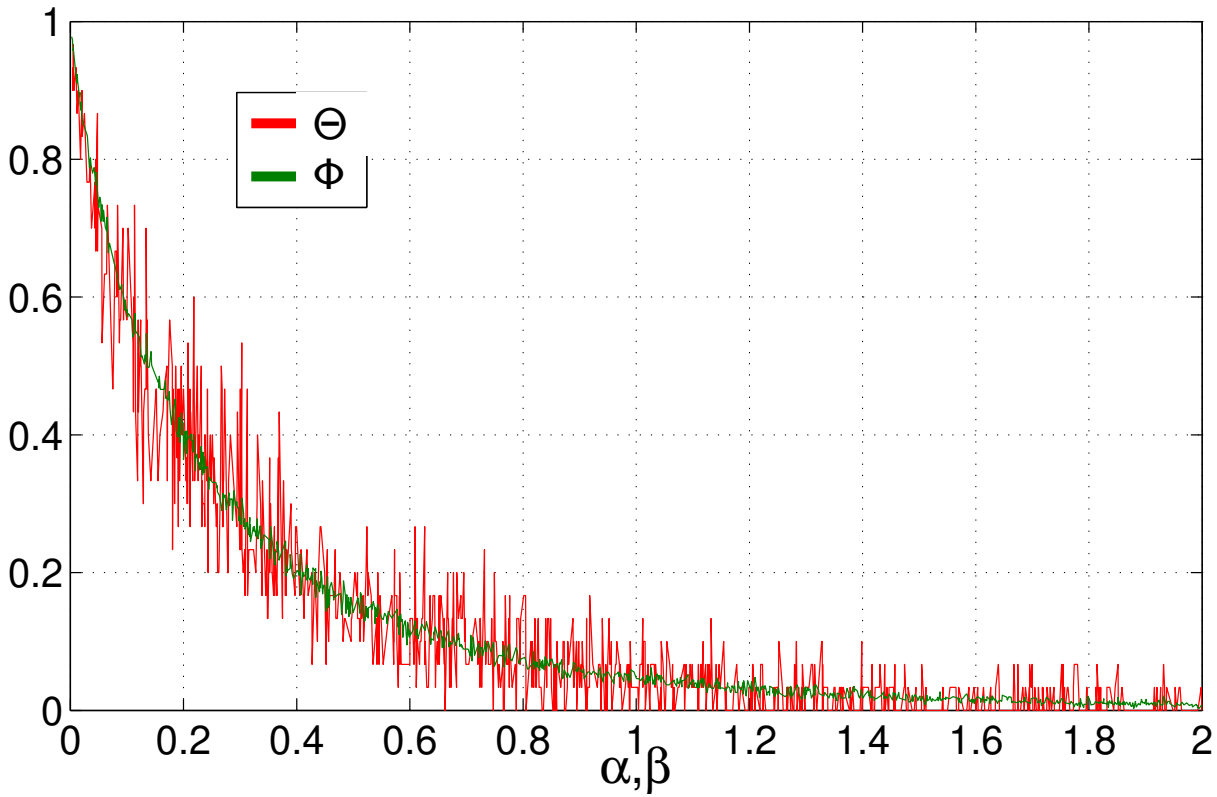
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

как для самих матриц Φ и Θ, так и для их произведения:

$$D_\Phi = H(\Phi, \Phi_0);$$
$$D_\Theta = H(\Theta, \Theta_0);$$
$$D_{\Phi\Theta} = H(\Phi\Theta, \Phi_0\Theta_0).$$

Разреженность векторов, порождаемых распределением Dir

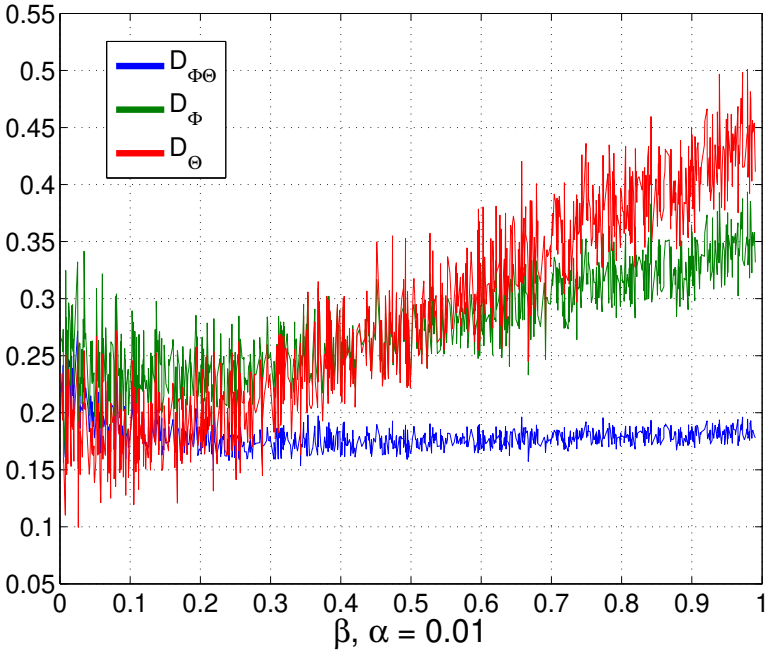
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \text{Dir}(\alpha)$ и $\phi_t^0 \sim \text{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



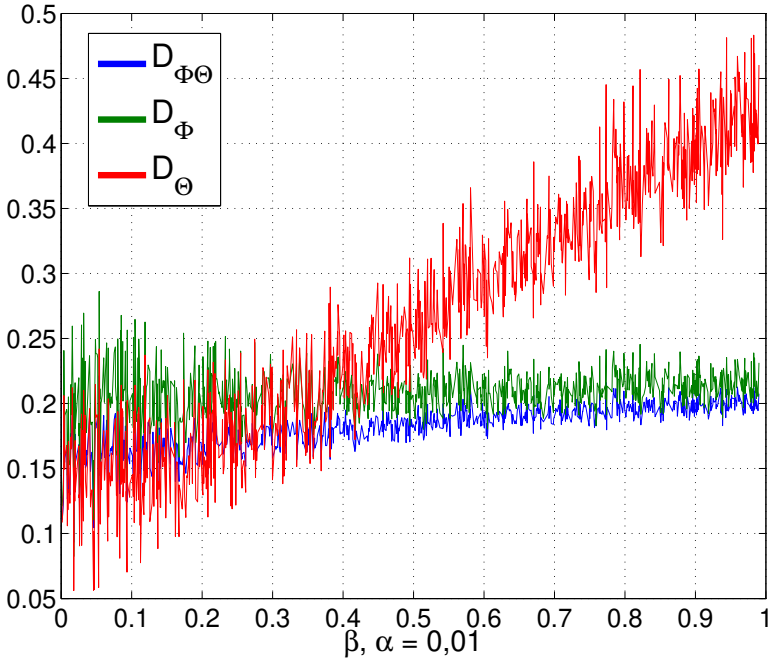
Неустойчивость восстановления матриц Φ и Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Φ_0 при фиксированном $\alpha = 0.01$

PLSA



LDA

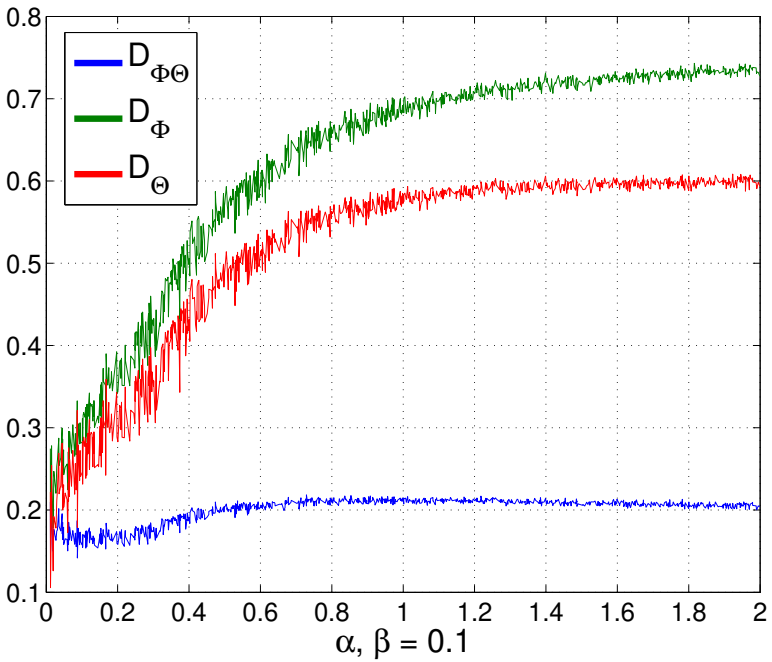


Виталий Глушаченков. Устойчивость матричных разложений в задачах тематического моделирования. Магистерская диссертация, МФТИ, 2013.

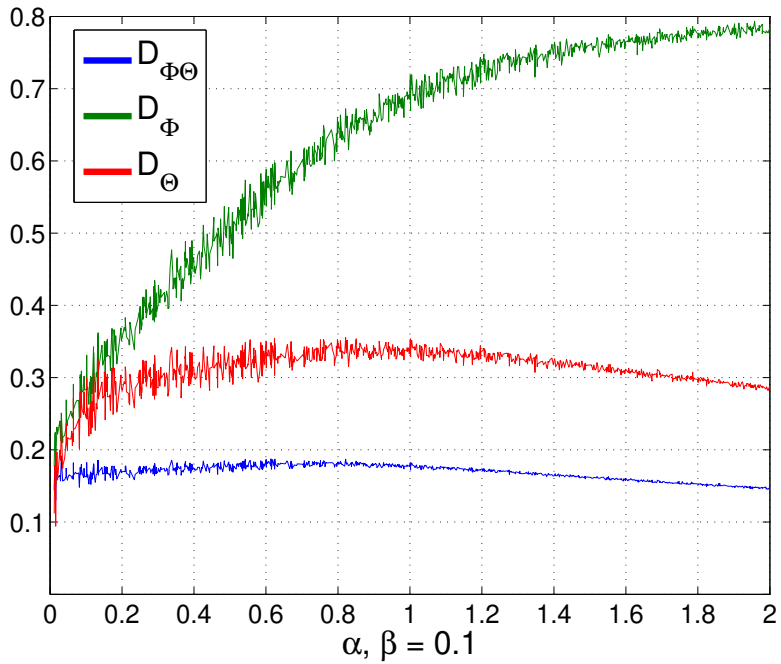
Неустойчивость восстановления матриц Φ и Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Θ_0 при фиксированном $\beta = 0.1$

PLSA



LDA



Виталий Глушаченков. Устойчивость матричных разложений в задачах тематического моделирования. Магистерская диссертация, МФТИ, 2013.

Второй эксперимент — на реальных данных

Посты ЖЖ: $|D| = 300\,\text{K}$, $|W| = 154\,\text{K}$, $n = 35\,\text{M}$, $|T| = 120$.
LDA: симметричное распределение Дирихле, $\beta = 0.1$, $\alpha = 0.5$.

Цель эксперимента — оценить различность тем, получаемых в нескольких запусках алгоритма LDA Gibbs Sampling.

Проблема «проклятия размерности»:
длинные хвосты мешают сравнивать распределения.

Доля существенных слов в темах (word ratio):

$$WR = \frac{1}{|W|}\frac{1}{|T|}\sum_{w\in W}\sum_{t\in T}\left[\phi_{wt} > \frac{1}{|W|}\right] \quad \text{(в эксперименте } \sim 3.5\%\text{)}$$

Доля существенных тем в документах (document ratio):

$$DR = \frac{1}{|D|}\frac{1}{|T|}\sum_{d\in D}\sum_{t\in T}\left[\theta_{td} > \frac{1}{|T|}\right] \quad \text{(в эксперименте } \sim 11.5\%\text{)}$$

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

Методика эксперимента

Оставлены слова w, имеющие $\phi_{wt} > \frac{1}{|W|}$ хотя бы в одной теме
Сокращение словаря (vocabulary reduction): 154 K $\to$ 8 K.

Дивергенция Кульбака–Лейблера между темами t и s:

$$\mathrm{KL}(t,s) = \sum_{w \in W} p(w|t) \ln \frac{p(w|t)}{p(w|s)}$$

Нормированная KL-близость пар тем t и s:

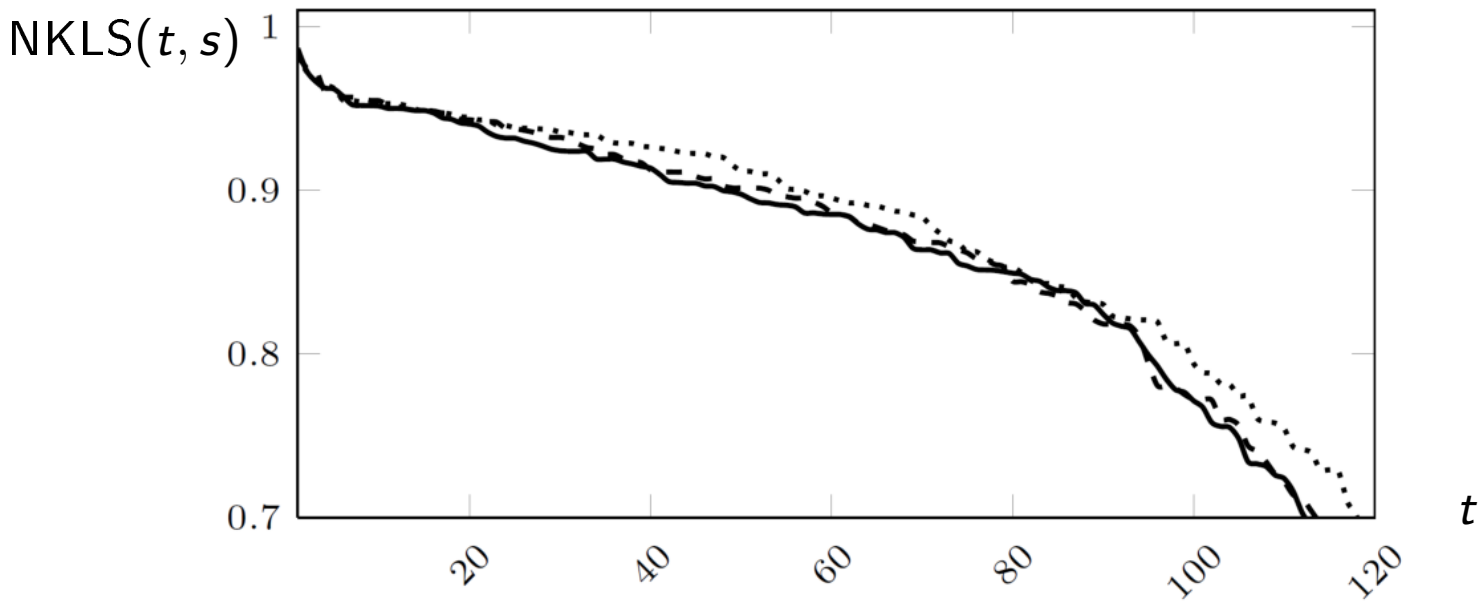
$$\mathrm{NKLS}(t,s) = \left(1 - \frac{\mathrm{KL}(t,s)}{\max\limits_{t',s'} \mathrm{KL}(t',s')}\right)$$

При $\mathrm{NKLS}(t,s) > 0.9$ в темах совпадают 30–50 топовых слов, и эксперты-социологи признают такие темы одинаковыми.

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

Неустойчивость LDA в разных запусках

Результат эксперимента: нормированная KL-близость NKLS между темой t и ближайшей к ней s в другом запуске.



1. Менее 50% тем воспроизводятся от запуска к запуску.
2. Плохо воспроизводятся как мусорные темы, так и хорошие.

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

Выводы из экспериментов

- Матрицы Φ, Θ устойчиво восстанавливаются только при сильной разреженности Φ_0, Θ_0 (более 90% нулей)
- Произведение $\Phi\Theta$ восстанавливается устойчиво, независимо от разреженности исходных Φ_0, Θ_0
- В разных запусках со случайной инициализацией или сэмплированием строятся существенно различные темы
- PLSA не переобучается, а лишь хуже моделирует малые вероятности редких слов, которые не интересны.
- Распределение Дирихле — слишком слабый регуляризатор

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models. Machine Learning. Springer, 2015.

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

Вместо резюме. Мифы про LDA

- LDA существенно меньше переобучается, чем PLSA
- LDA строит разреженные тематические модели
- LDA имеет меньше параметров по сравнению с PLSA
- LDA == тематическое моделирование

На самом деле,

- LDA и PLSA почти не отличаются на больших данных
- LDA не максимизирует разреженность моделей
- LDA имеет больше параметров по сравнению с PLSA
- LDA — лишь самая простая базовая модель
- LDA не имеет убедительных лингвистических обоснований

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.

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:

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

Подставляем, получаем:

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

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

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

Эксперименты с 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^0_{dw}) из полученных Φ и Θ:

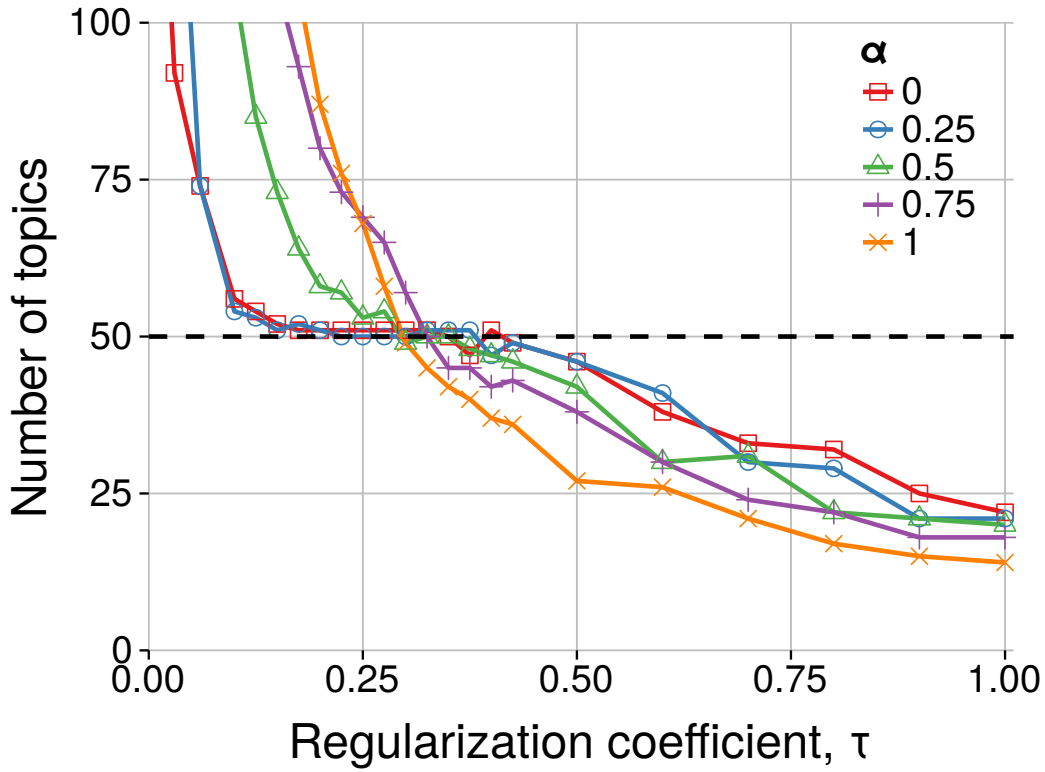
$$n^0_{dw} = n_d \sum_{t \in T_0} \phi_{wt}\theta_{td}$$

Параметрическое семейство полусинтетических данных:

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

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

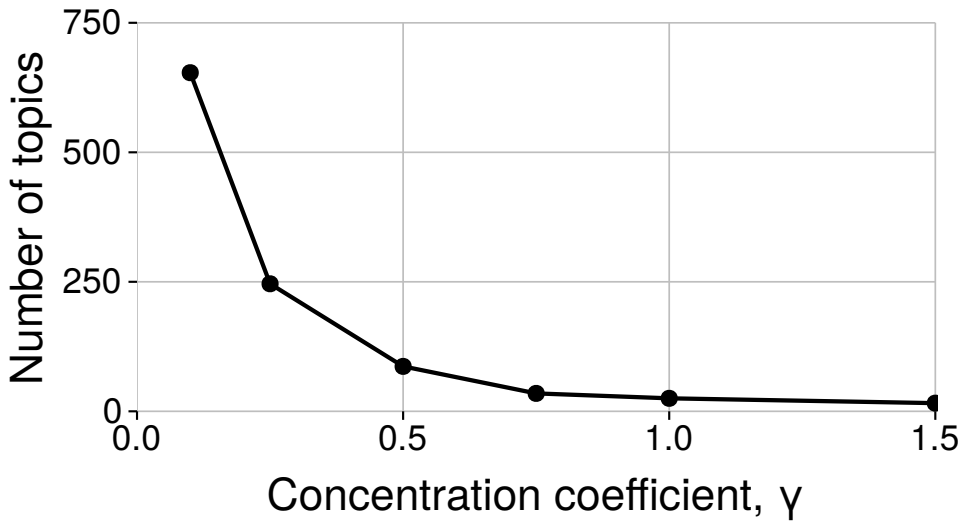
Попытка определения числа тем



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

Сравнение с байесовской тематической моделью HDP

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

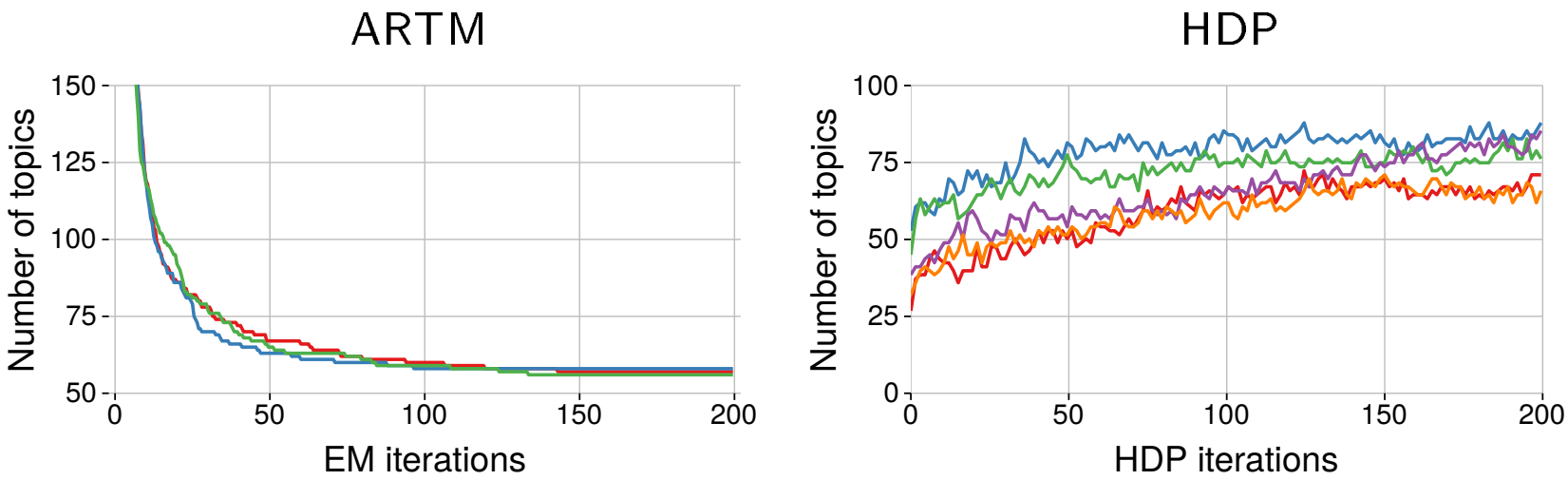


- *Коэффициент концентрации* γ в HDP влияет на $|T|$ так же сильно, как выбор коэффициента τ в ARTM.

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

Сравнение ARTM и HDP по устойчивости

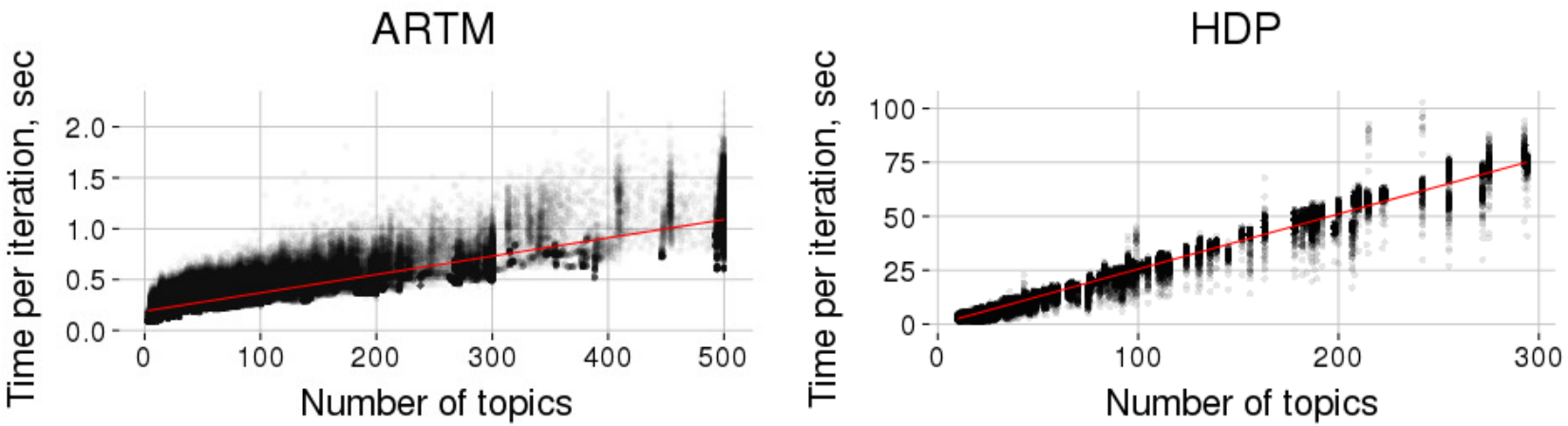
Запуск ARTM и HDP много раз из случайных инициализаций:



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

Сравнение ARTM и HDP по времени вычислений

Сравнение времени одного прохода коллекции (sec)

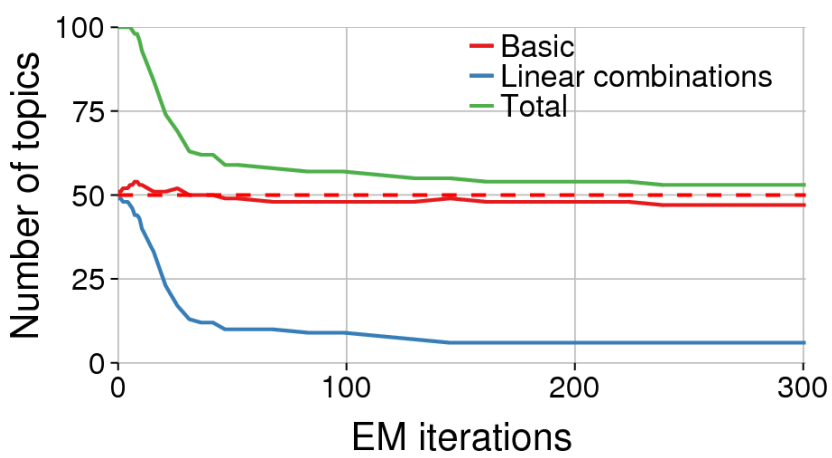


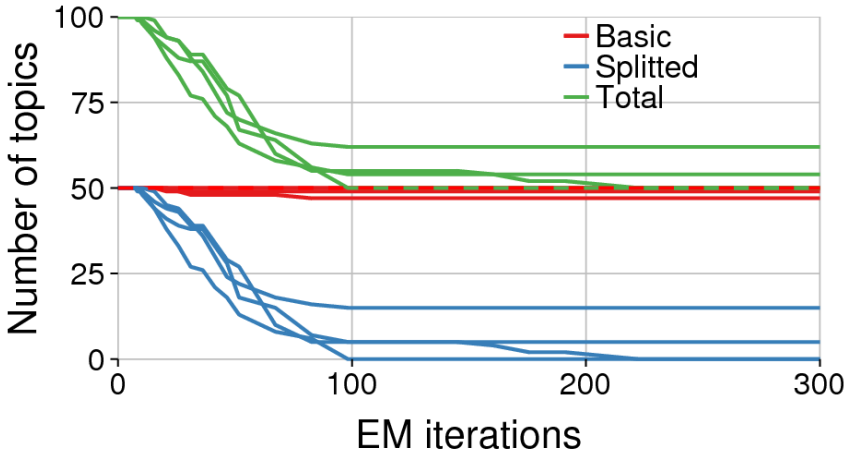
- ARTM в 100 раз быстрее!

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

Удаление линейно зависимых и расщеплённых тем

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





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

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

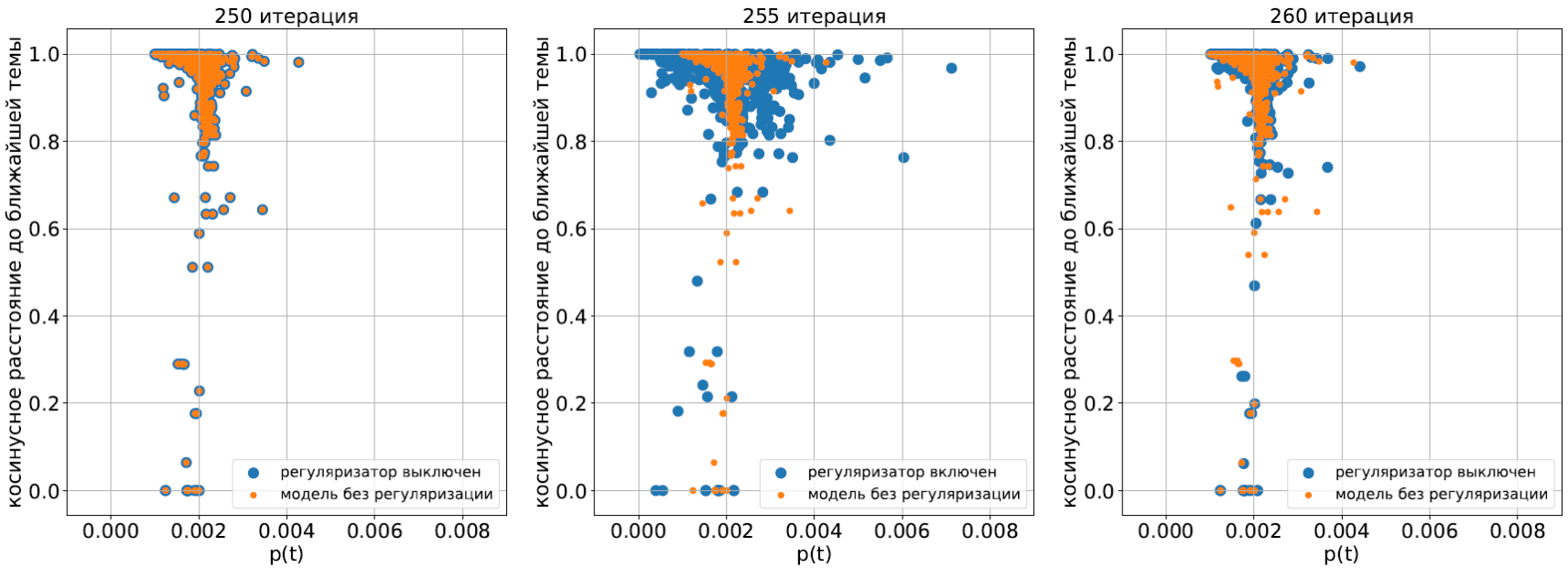
Выводы по результатам экспериментов

- Регуляризатор отбора тем удаляет незначимые темы и определяет оптимальное число тем, если оно существует
- Увы, в реальных данных его не существует!
  Оно задаётся исходя из целей моделирования.
- Значит, надо иерархически дробить темы на подтемы, пусть пользователь выбирает нужную ему детализацию
- Есть простой метод для удаления лишних тем,
  но как добавлять темы в ARTM — **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!

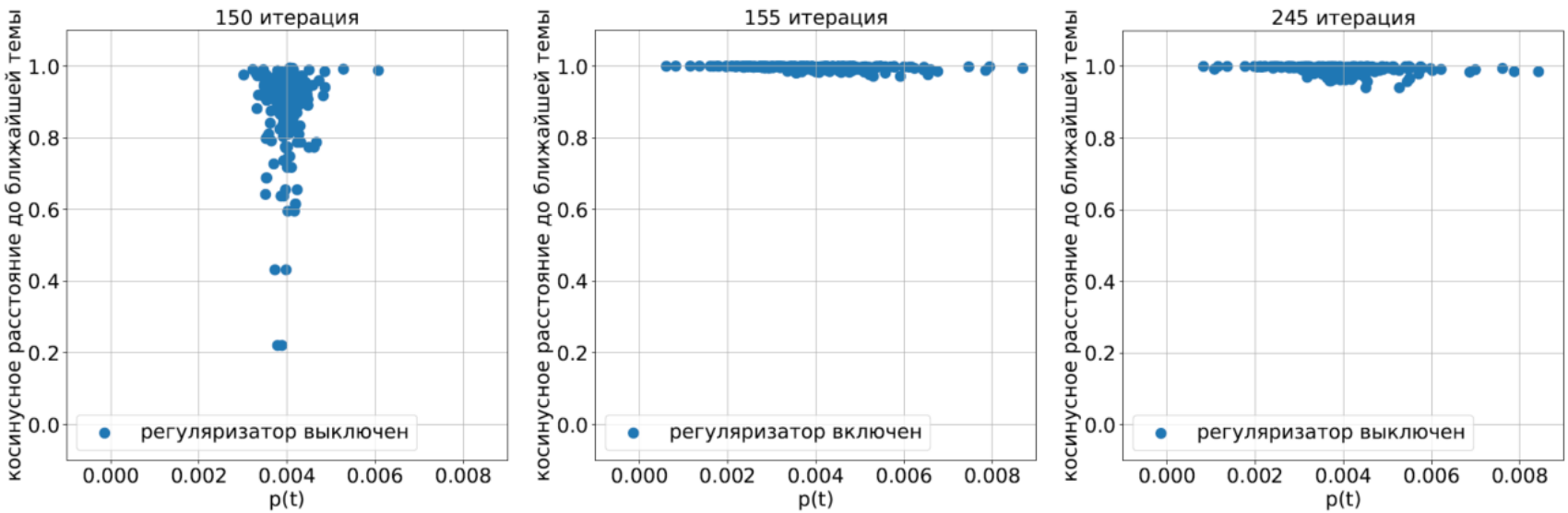


Г.Фоминская. Выявление тем-дубликатов в тематических моделях. Курсовая работа, ВМК МГУ, 2018.

Проблема малых тем и тем-дубликатов

Эксперимент на коллекции postnauka.ru

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

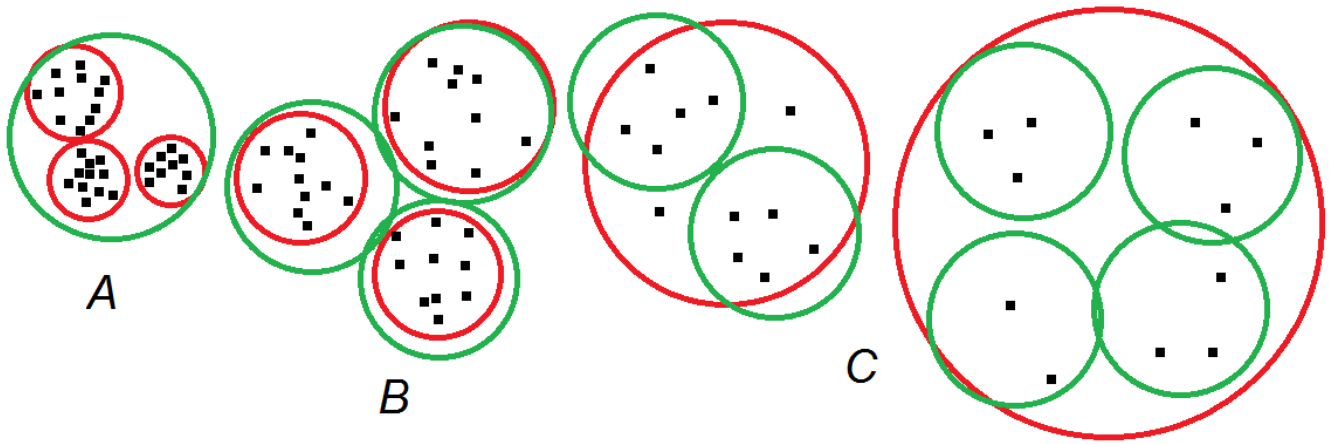


Г.Фоминская. Выявление тем-дубликатов в тематических моделях. Курсовая работа, ВМК МГУ, 2018.

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



Разреживание, сглаживание, декоррелирование, отбор тем

M-шаг при комбинировании 6 регуляризаторов:

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

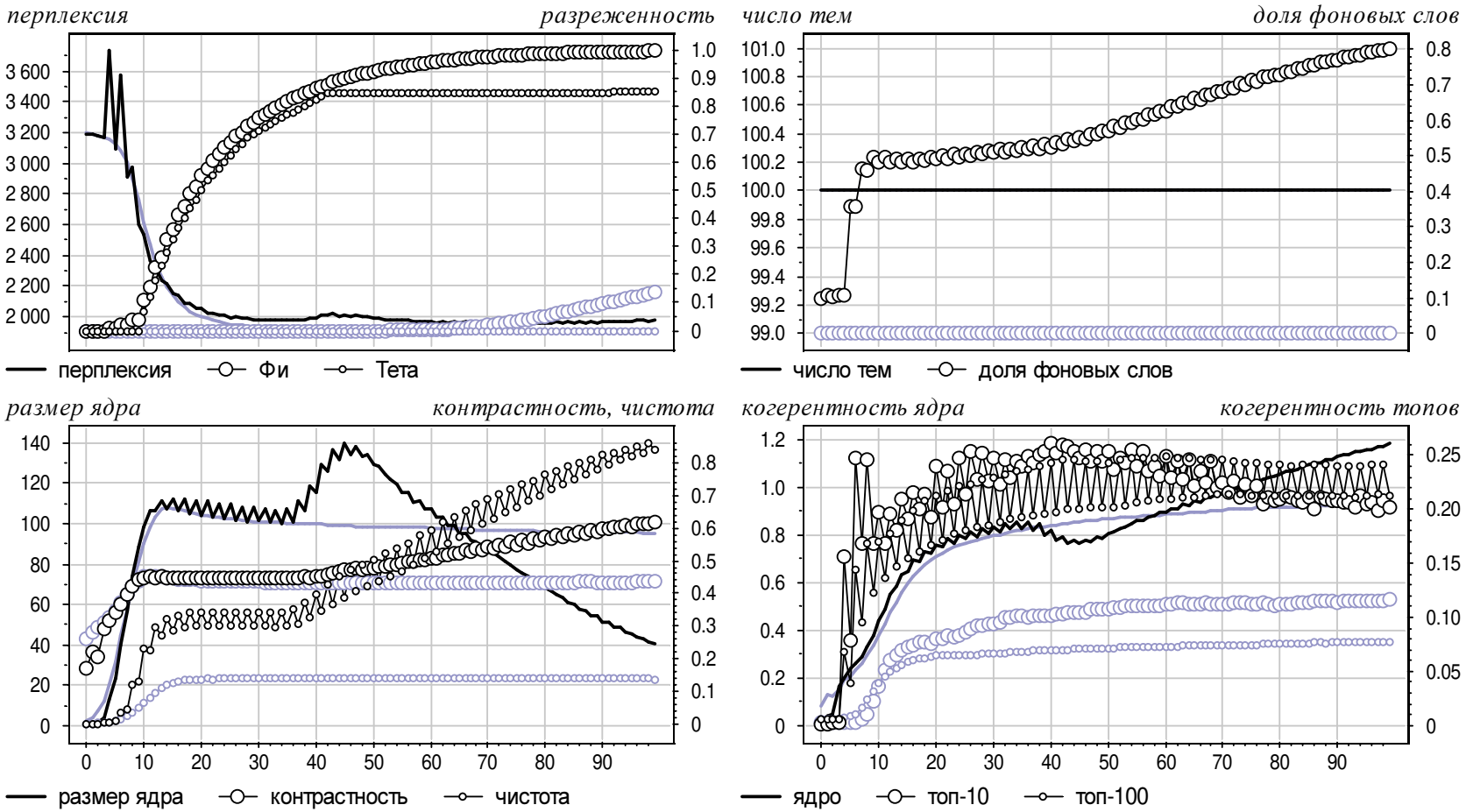
$$\theta_{td} = \underset{t}{\text{norm}}\Big(n_{td} + \tau_4 \underbrace{\alpha_t[t \in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \tau_5 \underbrace{\alpha_t[t \in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} - \tau_6 \underbrace{\frac{n_d}{n_t}\theta_{td}}_{\substack{\text{удаление}\\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

Данные: статьи NIPS (Neural Information Processing System)
$|D| = 1566$ статей, $n = 2.3\,\text{M}$, $|W| = 13\,\text{K}$,
контрольная коллекция: $|D'| = 174$.

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

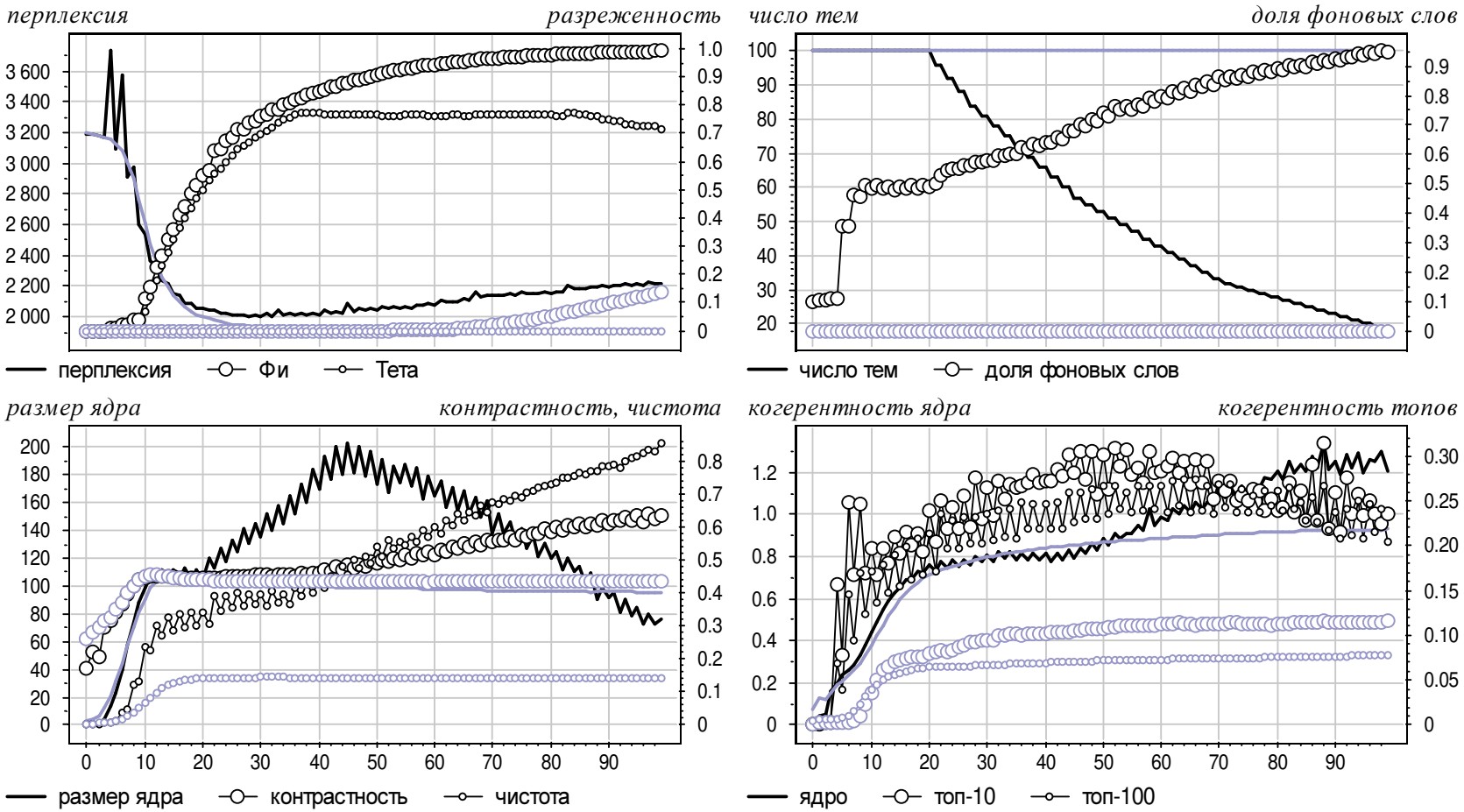
Разреживание, сглаживание, декоррелирование

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:

- *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
- почти без потери *перплексии* (правдоподобия) модели

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

Резюме

- 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