Вероятностные тематические модели
Лекция 4. Регуляризаторы для АРТМ

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 • весна 2017

Содержание

Напоминание. Задача тематического моделирования

Дано: W — словарь терминов (слов или словосочетаний),
D — коллекция текстовых документов $d \subset W$,
n_{dw} — сколько раз термин w встретился в документе d.

Найти: модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$ с параметрами $\underset{W \times T}{\Phi}$ и $\underset{T \times D}{\Theta}$:
$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t,
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d.

Критерий максимума логарифма правдоподобия:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \;\to\; \max_{\phi,\theta};$$

$$\phi_{wt} \geqslant 0; \quad \textstyle\sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \textstyle\sum_t \theta_{td} = 1.$$

Проблема: задача стохастического матричного разложения *некорректно поставлена*: $\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$.

Напоминание. ARTM и регуляризованный EM-алгоритм

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

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

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

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

PLSA: $R(\Phi,\Theta) = 0$

LDA: $R(\Phi,\Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}$

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:

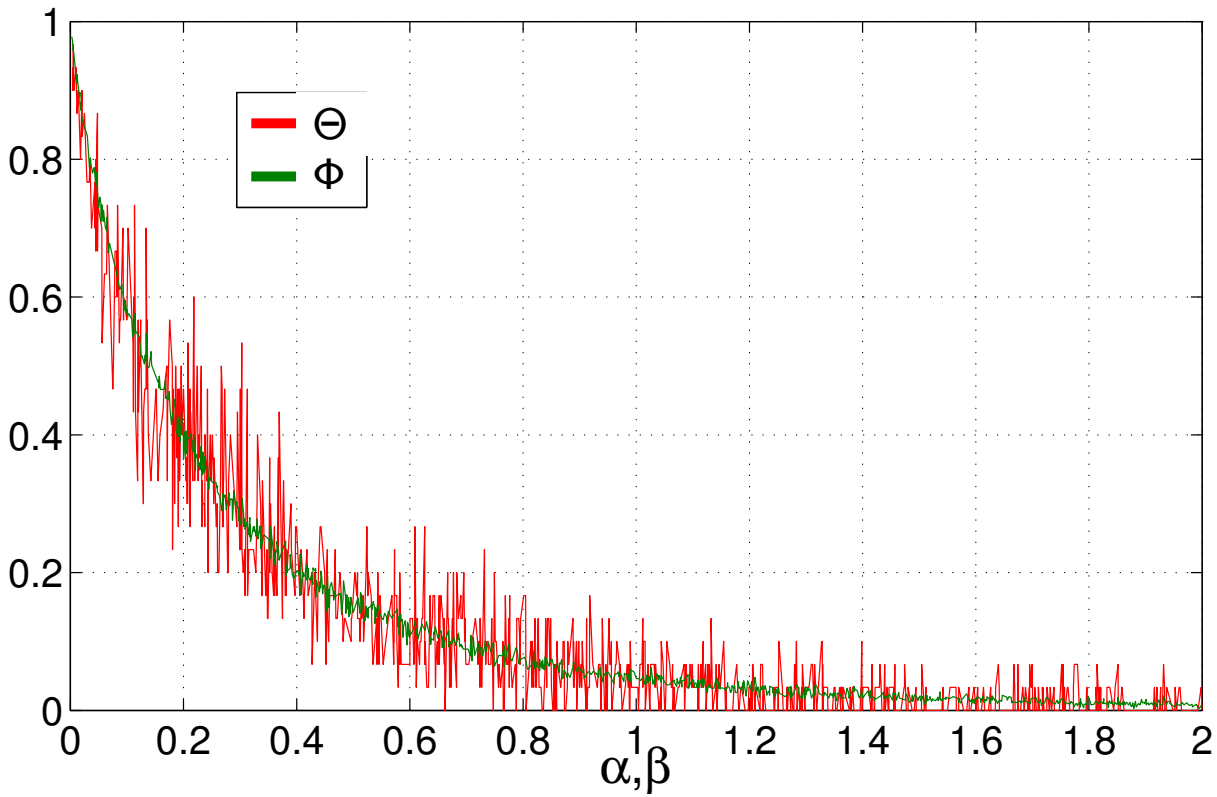
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

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

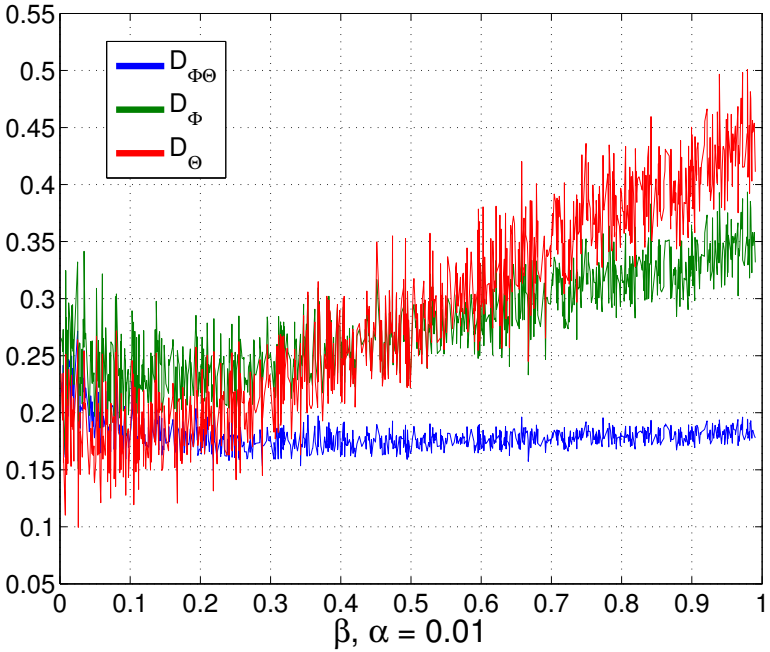
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \mathrm{Dir}(\alpha)$ и $\phi_t^0 \sim \mathrm{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



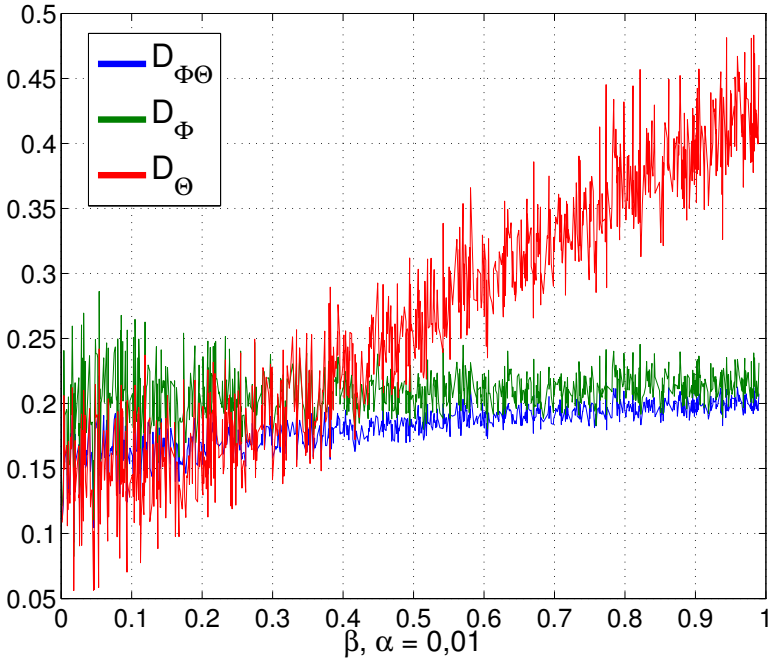
Неустойчивость восстановления матриц Φ и Θ

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

PLSA



LDA

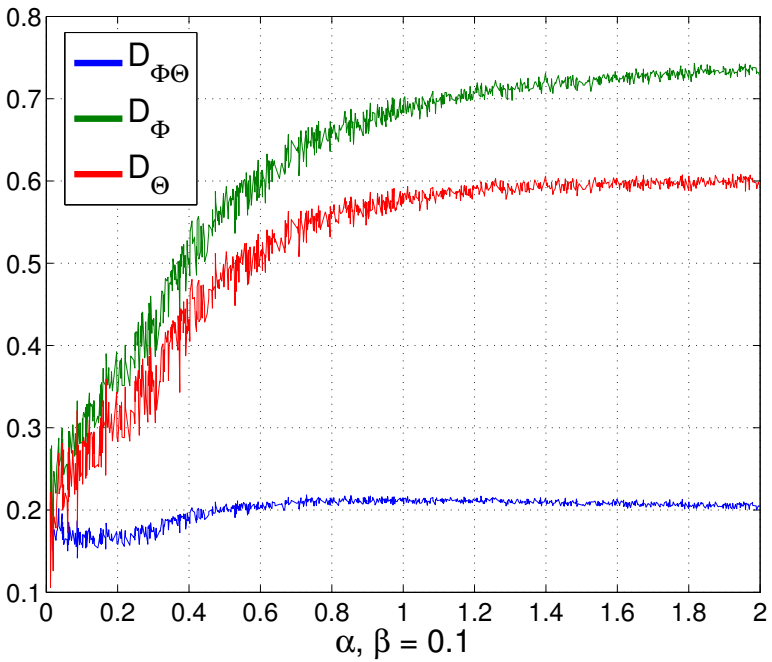


Виталий Глушаченков. Устойчивость матричных разложений в задачах тематического моделирования // Магистерская диссертация. МФТИ, 2013.

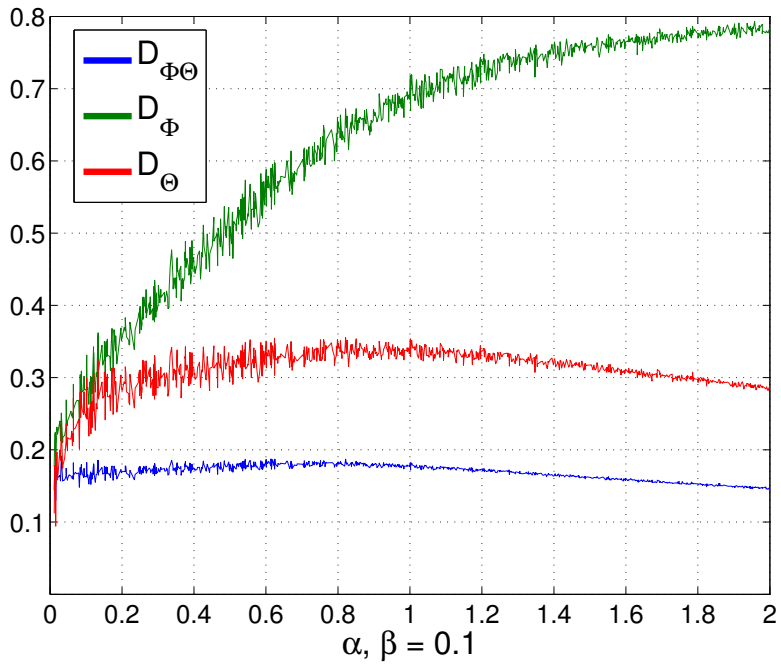
Неустойчивость восстановления матриц Φ и Θ

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.

Цель эксперимента

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 терминов 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:

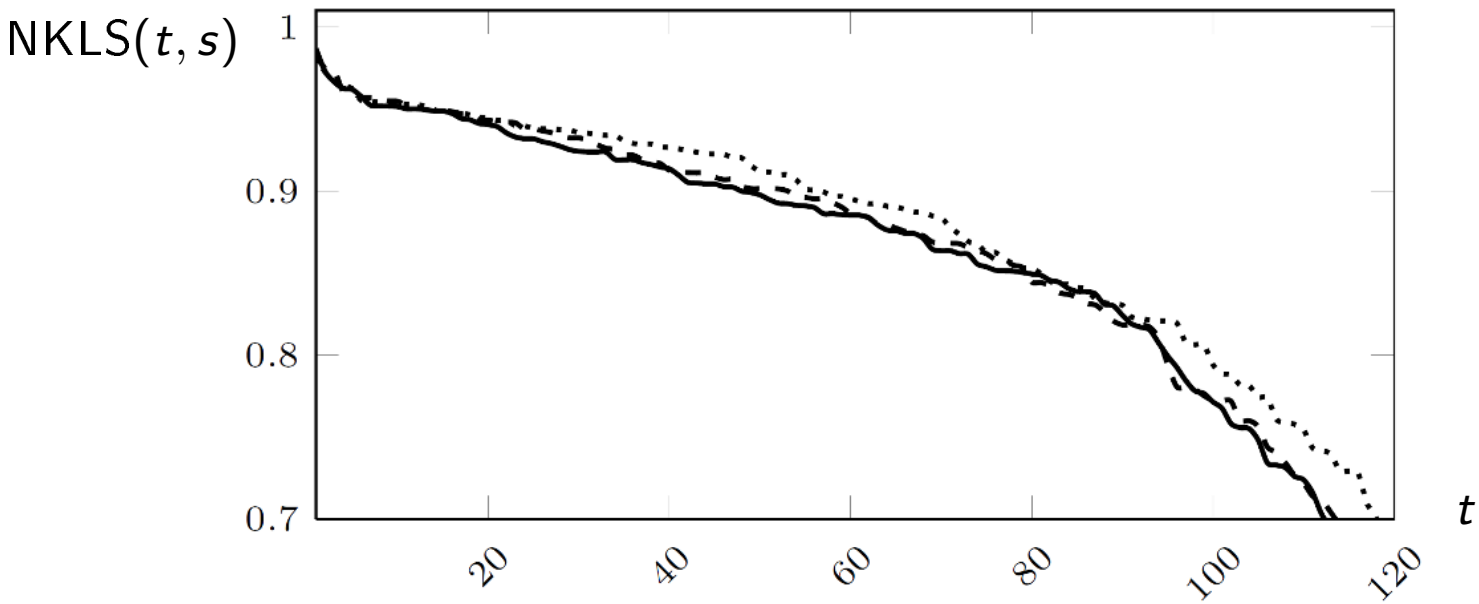
$$\mathrm{NKLS}(t,s) = \left(1 - \frac{\mathrm{KL}(t,s)}{\max\limits_{t',s'} \mathrm{KL}(t',s')}\right)$$

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

1. Матрицы Φ, Θ устойчиво восстанавливаются только при сильной разреженности Φ_0, Θ_0 (более 90% нулей)
2. Произведение $\Phi\Theta$ восстанавливается устойчиво, независимо от разреженности исходных Φ_0, Θ_0
3. В разных запусках с использованием случайных начальных приближений или сэмплирования EM-алгоритм находит существенно различающиеся наборы тем
4. 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.

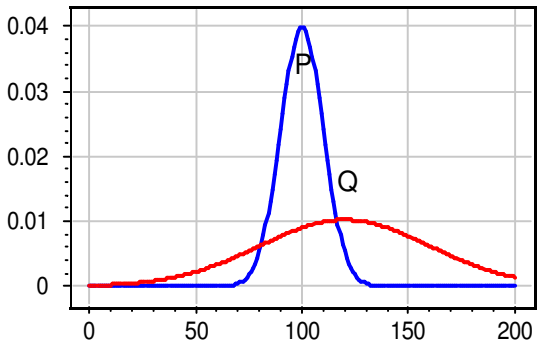
Напоминание. Дивергенция Кульбака–Лейблера

1. $\mathrm{KL}(P\|Q) \geqslant 0; \qquad \mathrm{KL}(P\|Q) = 0 \;\Leftrightarrow\; P = Q;$

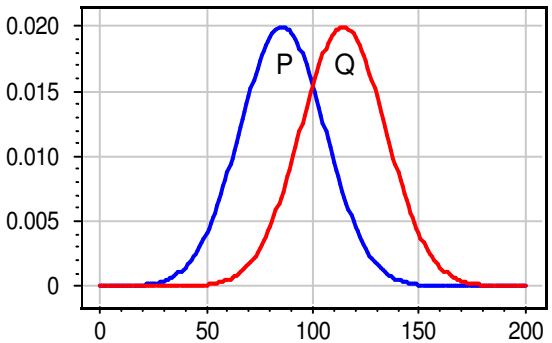
2. Минимизация KL эквивалентна максимизации правдоподобия:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_{\alpha} \quad \Leftrightarrow \quad \sum_{i=1}^{n} p_i \ln q_i(\alpha) \to \max_{\alpha}$$

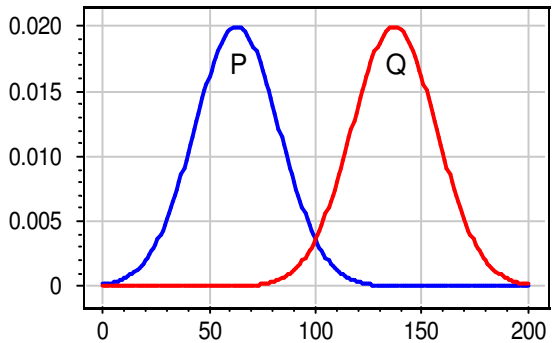
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P вложено в Q:



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 2.97$



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 0.44$



$\mathrm{KL}(P\|Q) = 2.97$
$\mathrm{KL}(Q\|P) = 2.97$

Регуляризатор сглаживания (переосмысление LDA)

Гипотеза сглаженности:
распределения ϕ_{wt} близки к заданному распределению β_w;
распределения θ_{td} близки к заданному распределению α_t.

$$\sum_{t\in T} \mathrm{KL}(\beta_w \| \phi_{wt}) \to \min_{\Phi}; \qquad \sum_{d\in D} \mathrm{KL}(\alpha_t \| \theta_{td}) \to \min_{\Theta}.$$

Максимизируем сумму регуляризаторов:

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

Подставляем, получаем формулы M-шага LDA:

$$\phi_{wt} = \underset{w\in W}{\mathrm{norm}}(n_{wt} + \beta_0\beta_w), \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}(n_{td} + \alpha_0\alpha_t).$$

Этого вы не найдёте в *D.Blei, A.Ng, M.Jordan.* Latent Dirichlet allocation // Journal of Machine Learning Research, 2003. — Vol. 3. — Pp. 993–1022.

Регуляризатор разреживания (обобщение LDA)

Гипотеза разреженности: среди ϕ_{wt}, θ_{td} много нулей;
распределения ϕ_{wt} далеки от заданного распределения β_w;
распределения θ_{td} далеки от заданного распределения α_t.

$$\sum_{t\in T} \mathrm{KL}(\beta_w \| \phi_{wt}) \to \max_{\Phi}; \qquad \sum_{d\in D} \mathrm{KL}(\alpha_t \| \theta_{td}) \to \max_{\Theta}.$$

Максимизируем сумму регуляризаторов:

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

Подставляем, получаем «анти-LDA»:

$$\phi_{wt} = \underset{w\in W}{\mathrm{norm}}\big(n_{wt} - \beta_0\beta_w\big), \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}\big(n_{td} - \alpha_0\alpha_t\big).$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010.

Объединение 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

Частичное обучение (semi-supervised learning) темы t:

- $\beta_{wt} = [w \in W_t]$ — *белый список* W_t терминов темы t
- $\alpha_{td} = [d \in D_t]$ — *белый список* D_t документов темы t
- $\beta_{wt} = -[w \in W_t]$ — *чёрный список* W_t терминов темы t
- $\alpha_{td} = -[d \in D_t]$ — *чёрный список* D_t документов темы t

Обобщённая KL-дивергенция

KL-дивергенция — это мера сходства векторов (β_w) и $(\ln \phi_w)$:

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

Почему бы не заменить $\ln x$ другой монотонной функцией $\mu(x)$?

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

M-шаг для регуляризатора обобщённой KL-дивергенции:

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\big(n_{wt} + \beta_0\beta_{wt} f(\phi_{wt})\big), \quad \theta_{td} = \underset{t\in T}{\operatorname{norm}}\big(n_{td} + \alpha_0\alpha_{td} f(\theta_{td})\big),$$

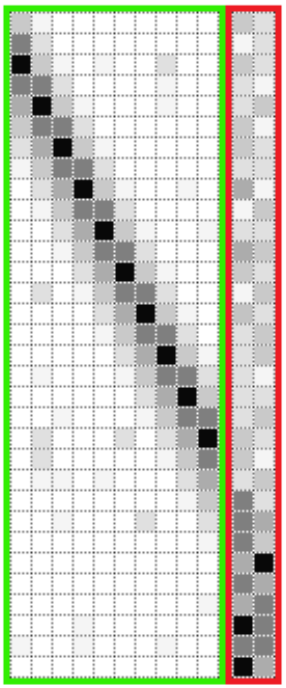
где $f(x) = x\mu'(x)$; в случае KL-дивергенции $\mu \equiv \ln$, $f(x) = 1$.

Разделение тем на предметные и фоновые

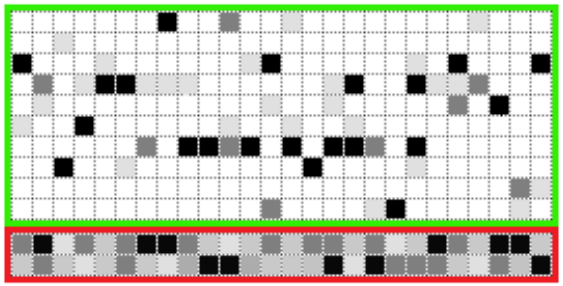
Предметные темы S содержат термины предметной области,
$p(w|t)$, $p(t|d)$, $t \in S$ — разреженные, существенно различные

Фоновые темы B содержат слова общей лексики,
$p(w|t)$, $p(t|d)$, $t \in B$ — существенно отличные от нуля

$\Phi_{W \times T}$



$\Theta_{T \times D}$



Регуляризатор декоррелирования тем

Цель — выделить *лексическое ядро* каждой темы, набор терминов, 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:

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

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Регуляризатор для сокращения числа тем

Цель: избавиться от «мелких» незначимых тем.

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

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

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 получается удалить 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.

Некоторые критерии качества тематической модели

Построение BTM — многокритериальная оптимизация.
Поэтому критериев для контроля качества модели тоже много.

- Перплексия контрольной коллекции: $\mathcal{P} = \exp\left(-\frac{1}{n}\mathscr{L}\right)$
- Разреженность — доля нулевых элементов в Φ и Θ
- Характеристики интерпретируемости тем:
  - когерентность темы: [Newman, 2010]
  - размер ядра темы: $|W_t|$, ядро $W_t = \{w\colon p(t|w) > 0.25\}$
  - чистота темы: $\sum\limits_{w \in W_t} p(w|t)$
  - контрастность темы: $\frac{1}{|W_t|} \sum\limits_{w \in W_t} p(t|w)$
- Вырожденность тематической модели:
  - число тем: $|T|$
  - доля фона в коллекции: $\frac{1}{n} \sum\limits_{d,w} \sum\limits_{t \in B} p(t|d,w)$

Оценки интерпретируемости: когерентность

Когерентность темы t

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

где w_i — i-й термин в порядке убывания ϕ_{wt}.

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

P_{uv} — доля документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов),

P_u — доля 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.

Разреживание + Сглаживание + Декорреляция + Отбор тем

M-шаг при комбинировании 6 регуляризаторов:

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

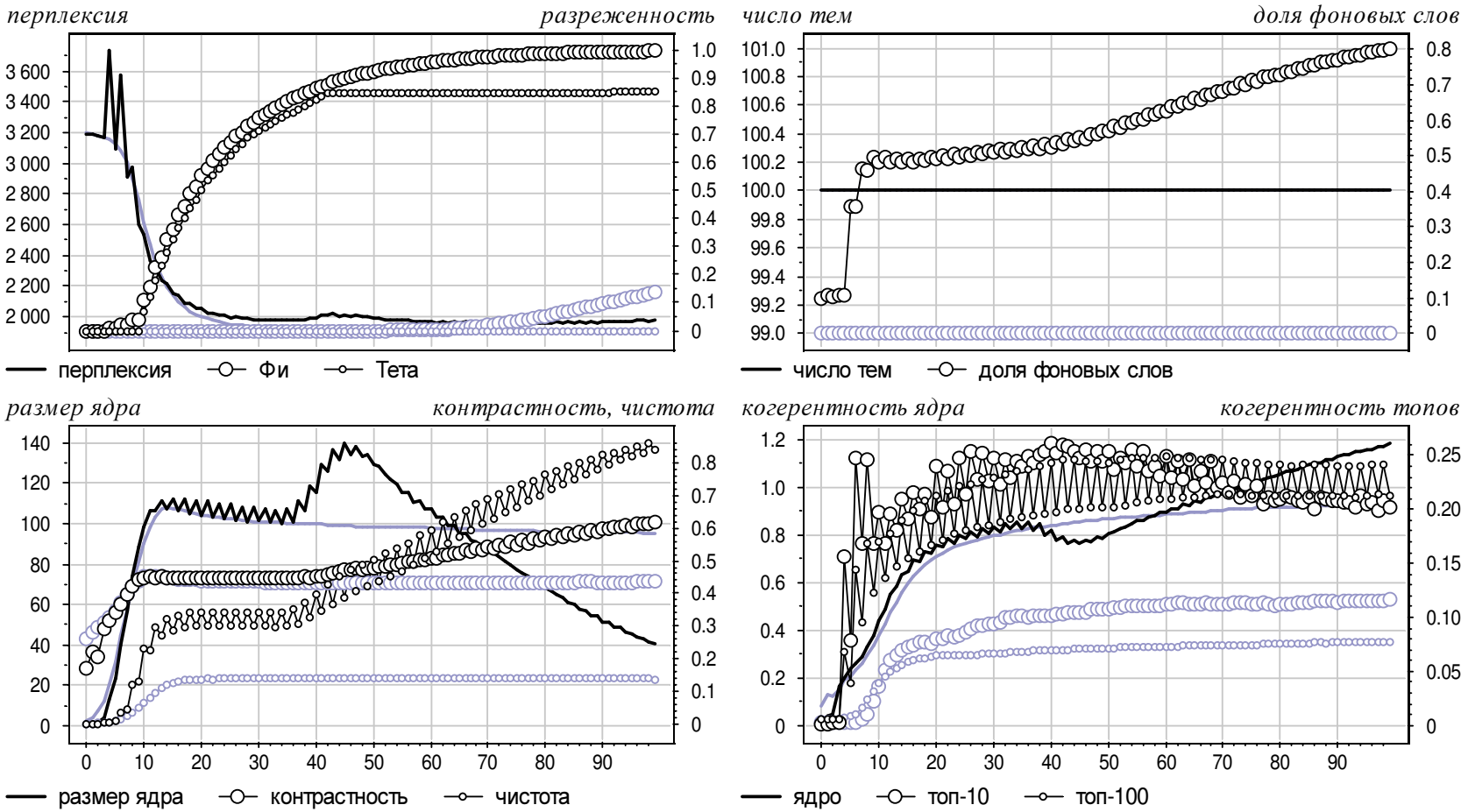
$$\theta_{td} = \underset{t}{\text{norm}}\Big(n_{td} + \tau_4 \underbrace{\alpha_t[t \in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \tau_5 \underbrace{\alpha_t[t \in S]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} - \tau_6 \underbrace{\frac{n_d}{n_t}\theta_{td}}_{\substack{\text{удаление}\\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

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.

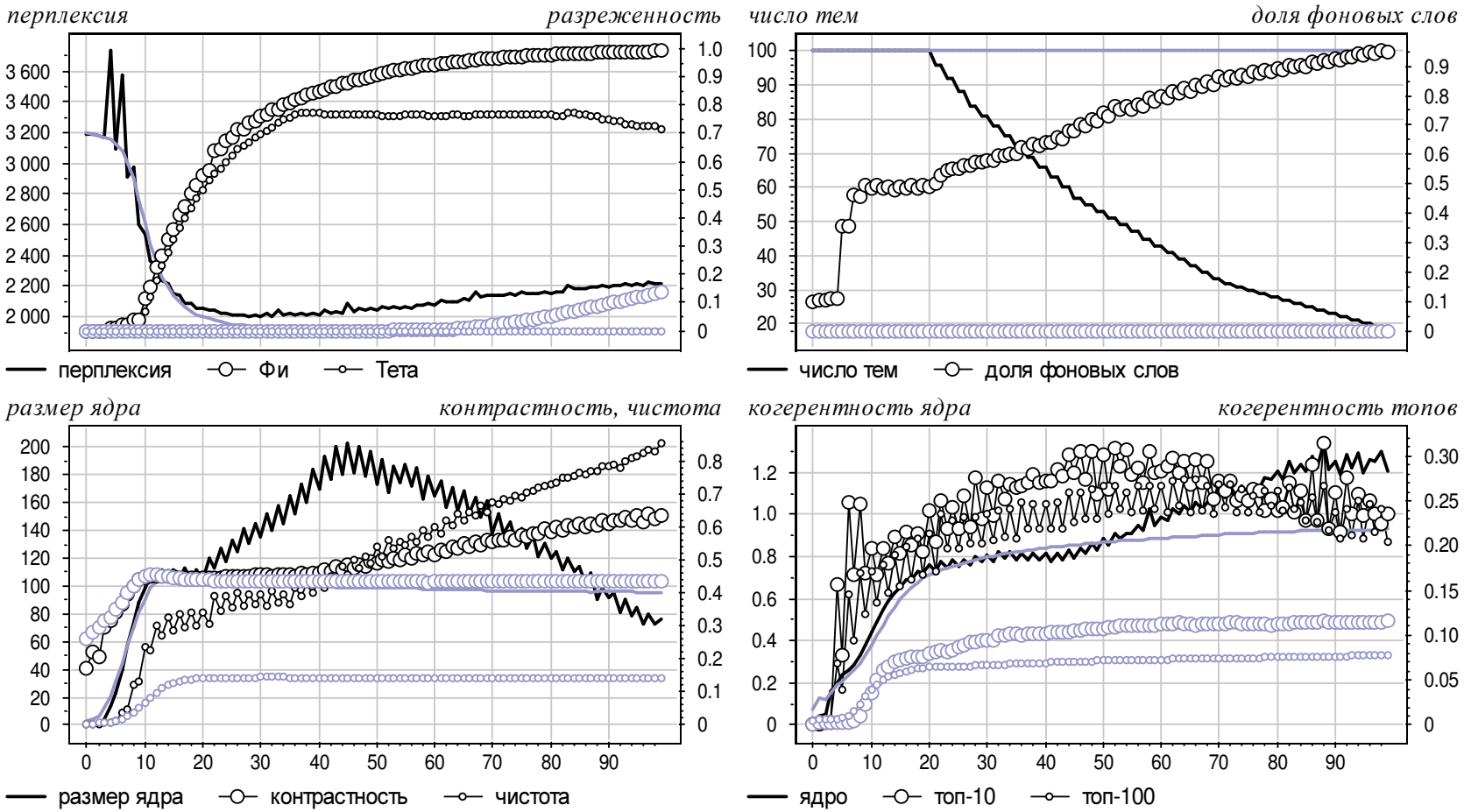
Разреживание, сглаживание, декорреляция

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

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

Подобраны траектории регуляризации:

- разреживание включать постепенно после 10-20 итераций
- сглаживание включать сразу
- декорреляцию включать сразу и как можно сильнее
- сокращение числа тем включать постепенно,
- никогда не совмещая с декорреляцией 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 (n^0_{dw}) из полученных Φ и Θ:

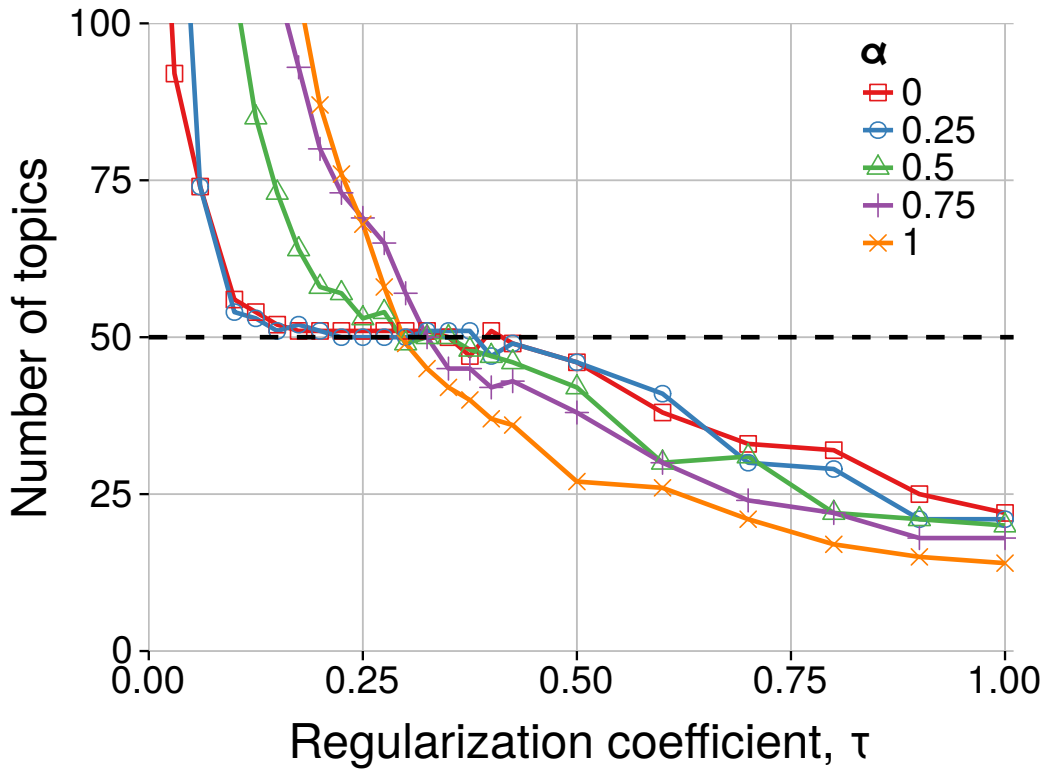
$$n^0_{dw} = n_d \sum_{t \in T} \phi_{wt} \theta_{td}$$

Параметрическое семейство полусинтетических данных:

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

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

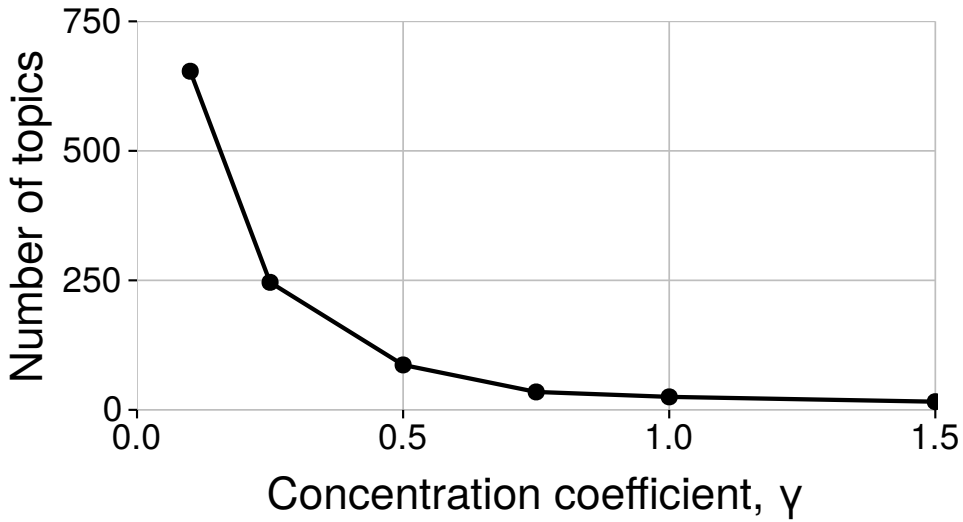
Попытка определения числа тем



- На синтетических данных надёжно находим $|T| = 50$,
- в широком интервале значений коэффициента τ;
- однако на реальных данных нет столь чёткого интервала.

Сравнение с байесовской тематической моделью HDP

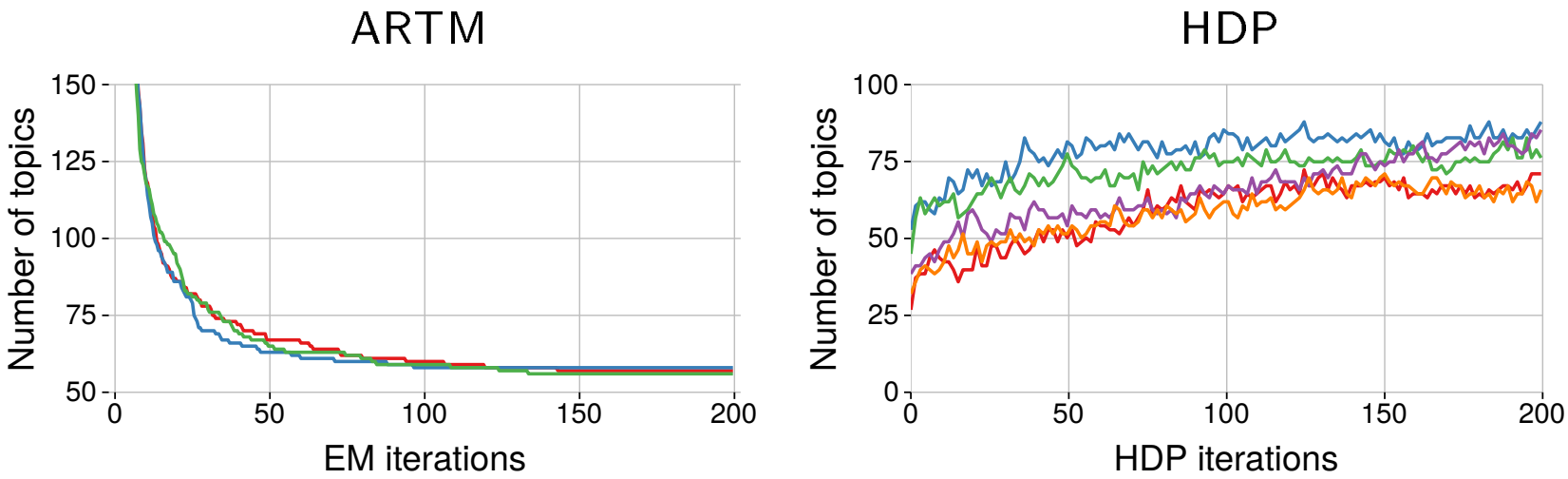
HDP, Hierarchical Dirichlet Process [Tech et.al, 2006] — «state-of-the-art» байесовский подход к определению числа тем



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

Сравнение ARTM и HDP по устойчивости

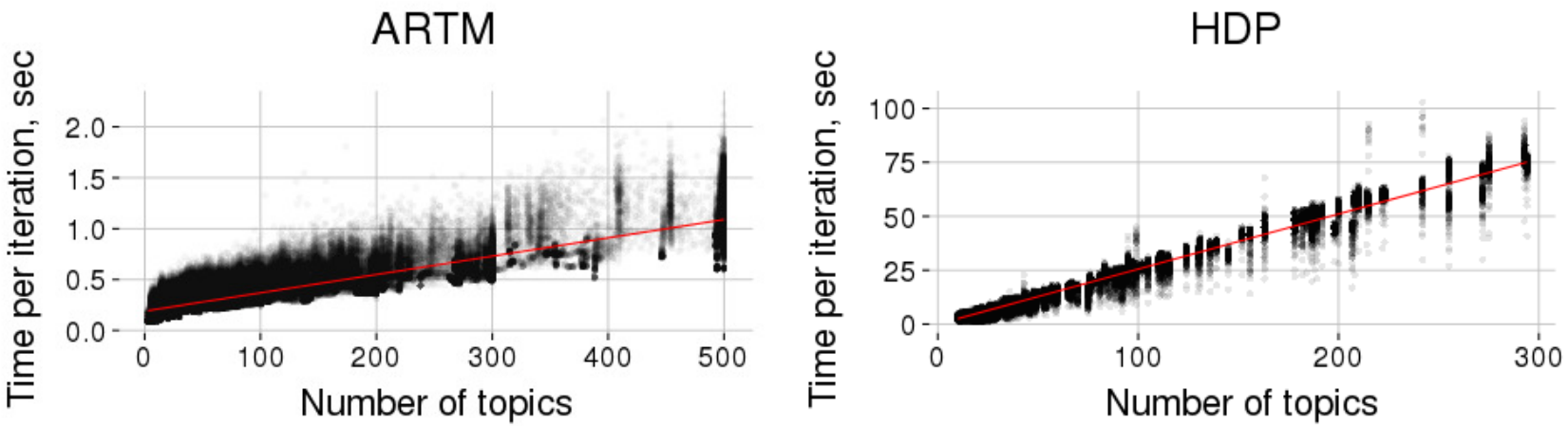
Запуск ARTM и HDP много раз из случайных инициализаций:



- 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)

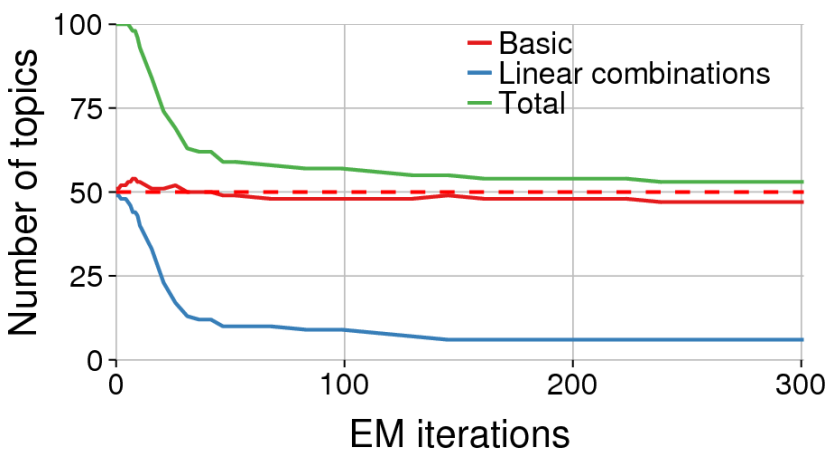


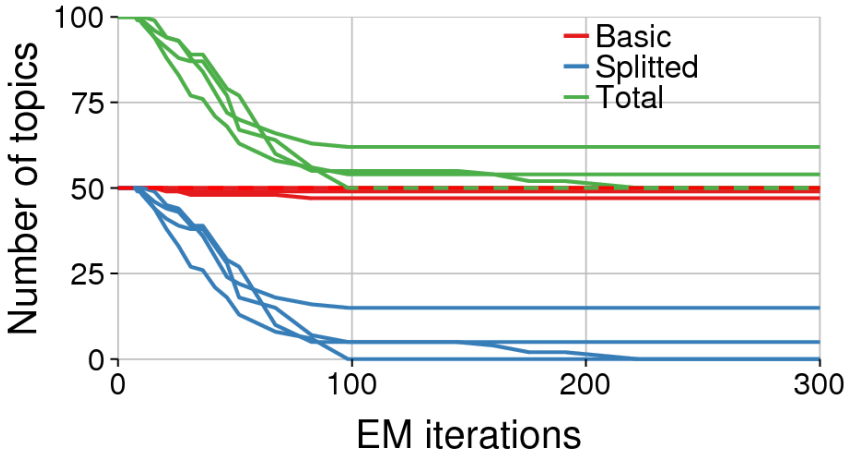
- ARTM в 100 раз быстрее!

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization // SLDS 2015, Royal Holloway, University of London, UK. pp. 193–202.

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 более 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, Royal Holloway, University of London, UK. pp. 193–202.

Резюме

- Решение задач анализа текстов *в стиле ARTM* — это построение моделей с заданными свойствами путём включения нужного набора регуляризаторов.
- Разреживание, сглаживание и декоррелирование — «джентльменский набор» регуляризаторов для повышения интерпретируемости и различности тем.
- Регуляризатор отбора тем — для удаления незначимых, зависимых, расщеплённых тем.
- Оптимального числа тем вообще не существует!
- Коэффициенты регуляризации пока подбираем вручную, их автоматическая настройка — в стадии разработки.