Вероятностные тематические модели
Лекция 3.
Латентное размещение Дирихле

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 2016

Задача вероятностного тематического моделирования

Дано: W — словарь терминов

D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$

n_{dw} — частота термина w в документе d

$\hat{p}(w|d) = \frac{n_{dw}}{n_d}$, где n_d — длина документа d

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

$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t

$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Задача стохастического матричного разложения *некорректно поставлена*, т. к. её решение не единственно и не устойчиво:

$$\underset{W\times D}{\left(\frac{n_{dw}}{n_d}\right)} \approx \underset{W\times T}{\Phi} \cdot \underset{T\times D}{\Theta} = (\Phi S)(S^{-1}\Theta) = \underset{W\times T}{\Phi'} \cdot \underset{T\times D}{\Theta'}$$

для невырожденных $\underset{T\times T}{S}$ таких, что Φ', Θ' тоже стохастические

Содержание

Гипотеза об априорных распределениях Дирихле

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} \underbrace{p(w|t)}_{\phi_{wt}} \underbrace{p(t|d)}_{\theta_{td}}$

1. Пусть $\theta_d = (\theta_{td})_{t \in T} \in \mathbb{R}^{|T|}$ — случайные векторы из распределения Дирихле с параметром $\alpha \in \mathbb{R}^{|T|}$:

$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t > 0;$$

2. Пусть $\phi_t = (\phi_{wt})_{w \in W} \in \mathbb{R}^{|W|}$ — случайные векторы из распределения Дирихле с параметром $\beta \in \mathbb{R}^{|W|}$:

$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \ \beta_w > 0;$$

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

Вероятностная модель порождения текста

Тематическая модель LDA (Latent Dirichlet Allocation):

$$p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}, \quad \phi_t \sim \mathrm{Dir}(\phi|\beta), \quad \theta_d \sim \mathrm{Dir}(\theta|\alpha).$$

Процесс порождения документов $d = \{w_1 \dots w_{n_d}\}$ коллекции D:

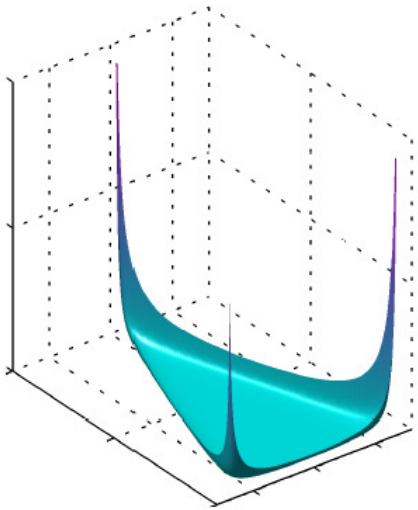
Вход: векторы гиперпараметров β, α;
Выход: коллекция документов;

выбрать вектор ϕ_t из $\mathrm{Dir}(\phi|\beta)$ для каждой темы $t \in T$;
выбрать вектор θ_d из $\mathrm{Dir}(\theta|\alpha)$ для каждого документа $d \in D$;
для всех документов $d \in D$
  для всех позиций слов $i = 1, \dots, n_d$ в документе d
    выбрать тему t_i из $p(t|d) \equiv \theta_{td}$;
    выбрать слово w_i из $p(w|t_i) \equiv \phi_{wt}$;

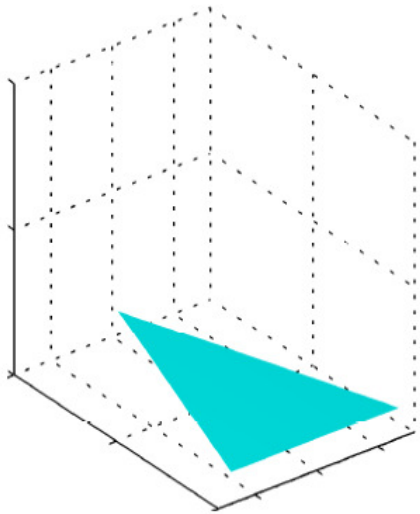
Почему именно распределение Дирихле?

– Может порождать сглаженные или разреженные векторы
– Имеет параметры, управляющие степенью разреженности
– Неплохо описывает кластерные структуры на симплексе
– Является сопряжённым к мультиномиальному распределению

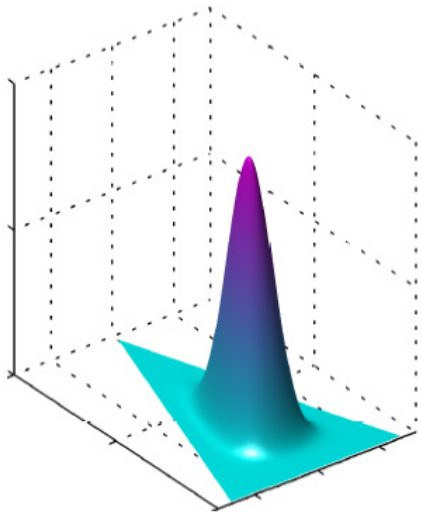
Пример. $\mathrm{Dir}(\theta|\alpha)$ при $|T| = 3$, $\theta, \alpha \in \mathbb{R}^3$:



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$

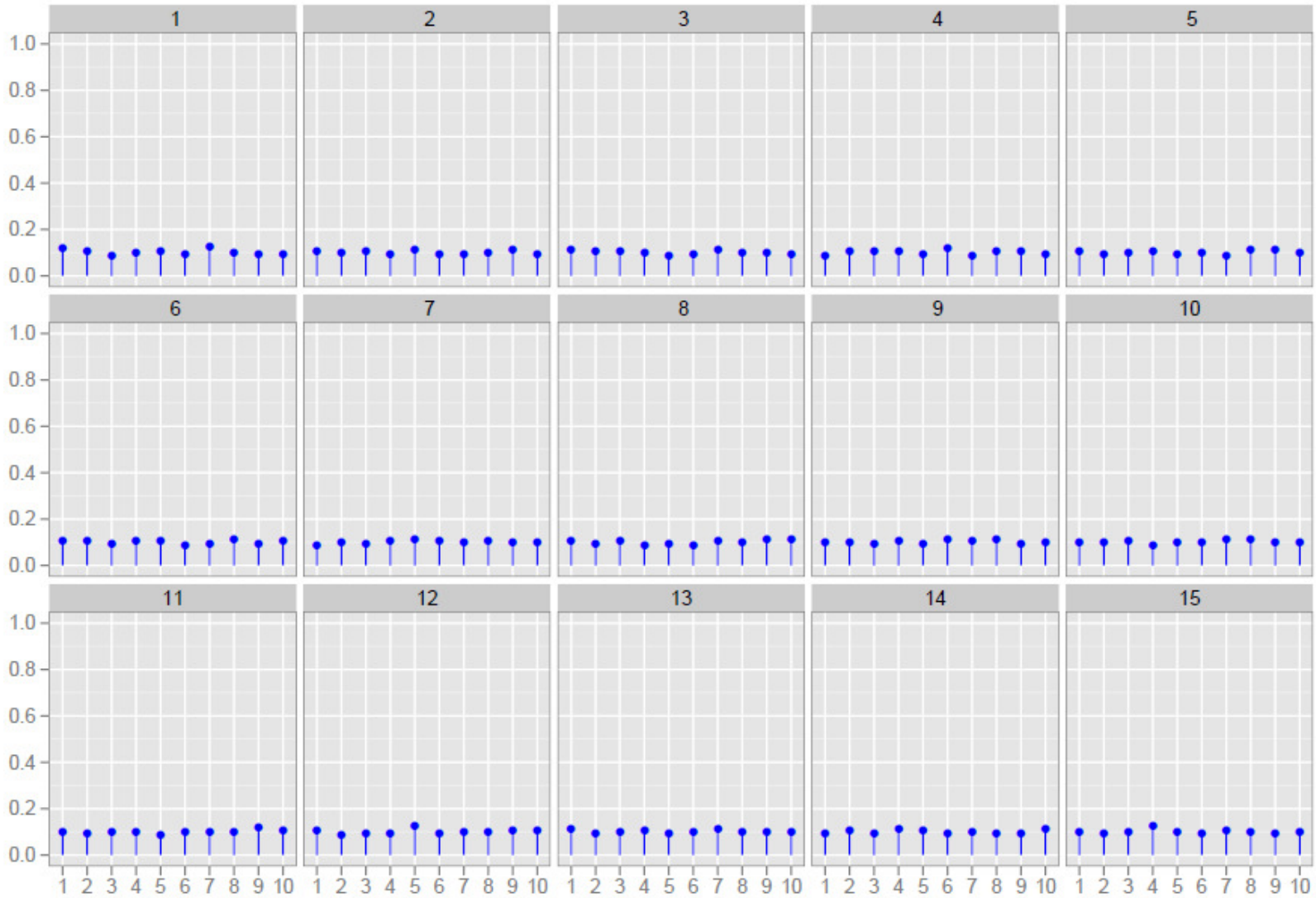


$\alpha_1 = \alpha_2 = \alpha_3 = 1$

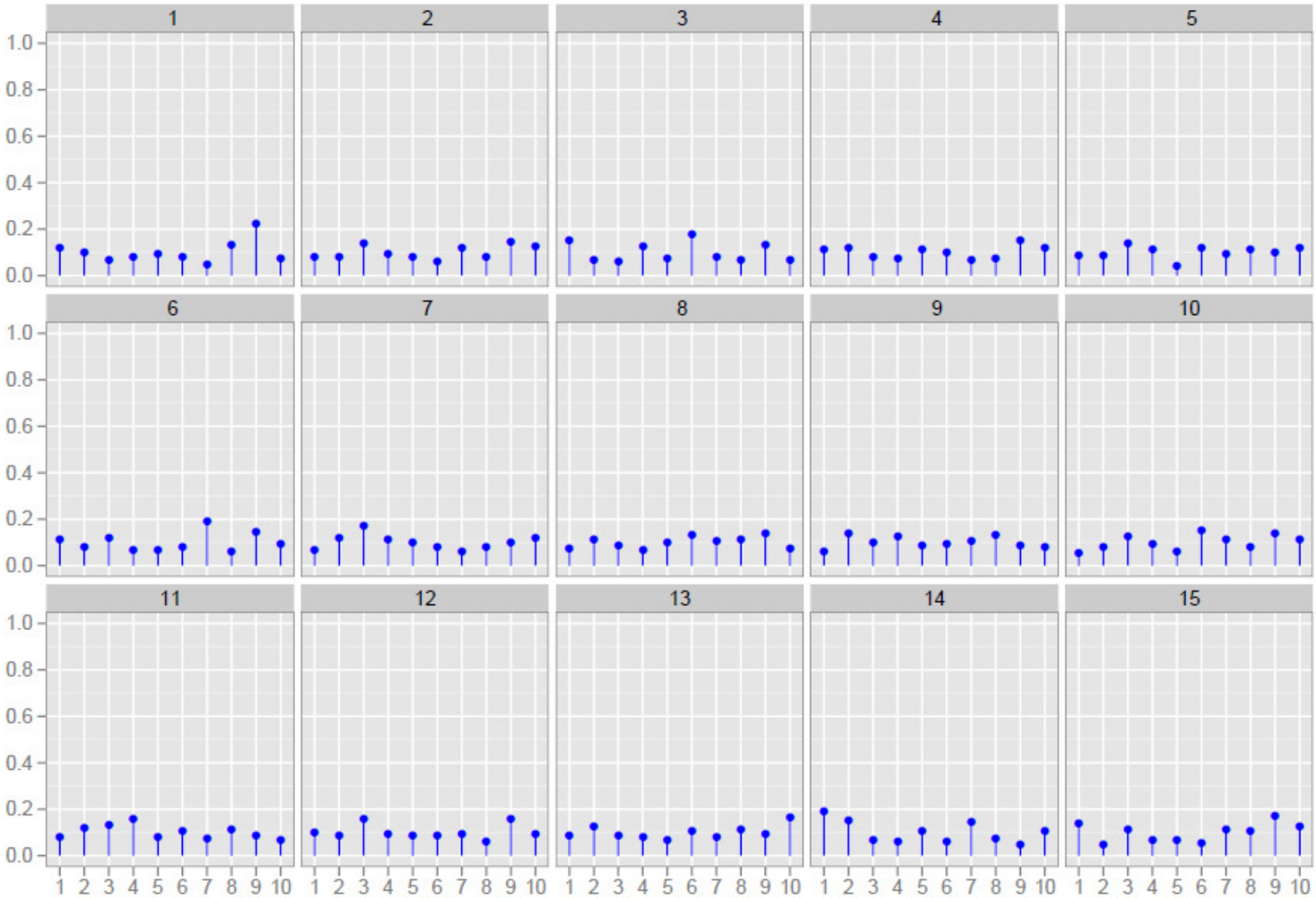


$\alpha_1 = \alpha_2 = \alpha_3 = 10$

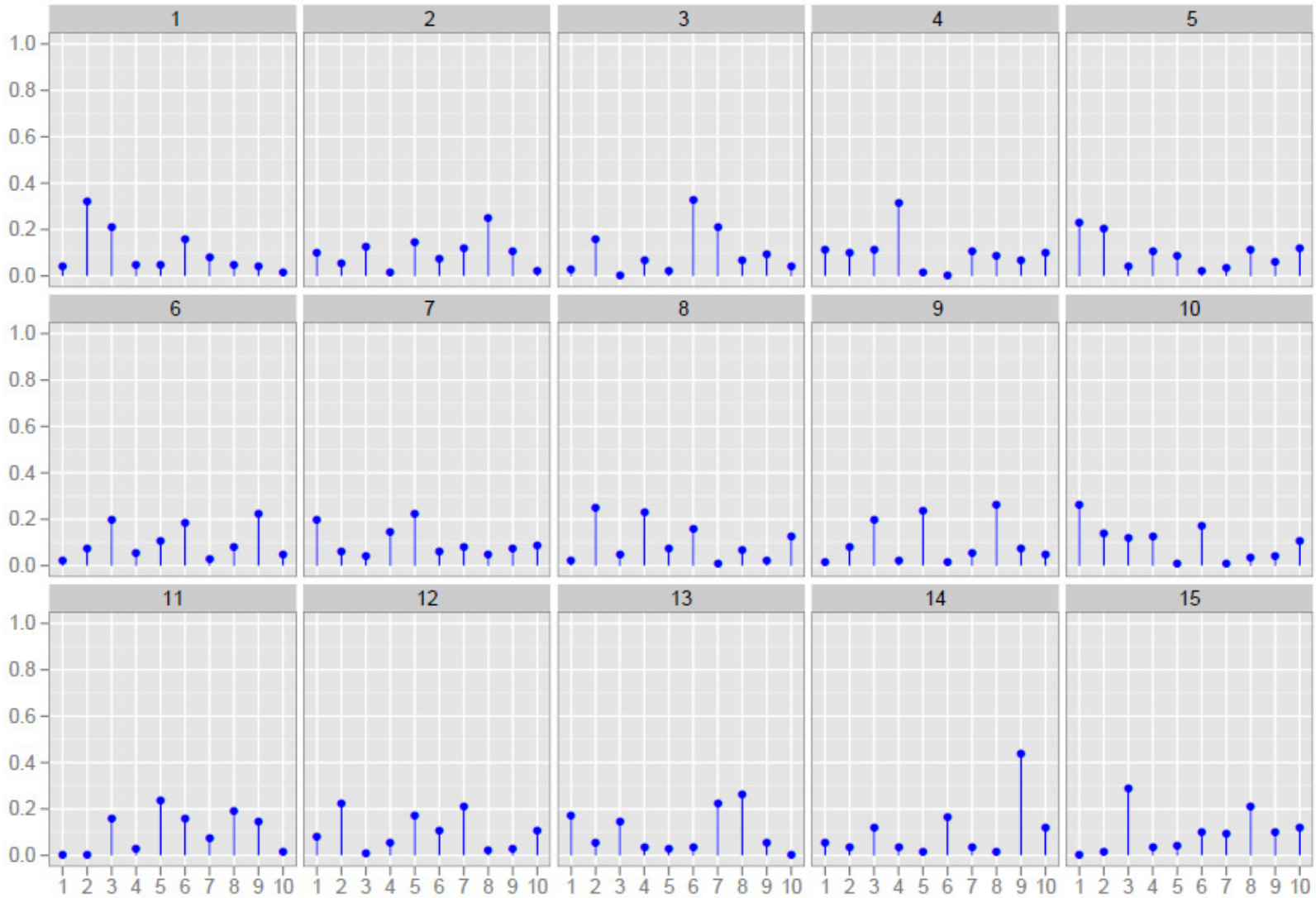
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 100$, 10 тем, 15 документов



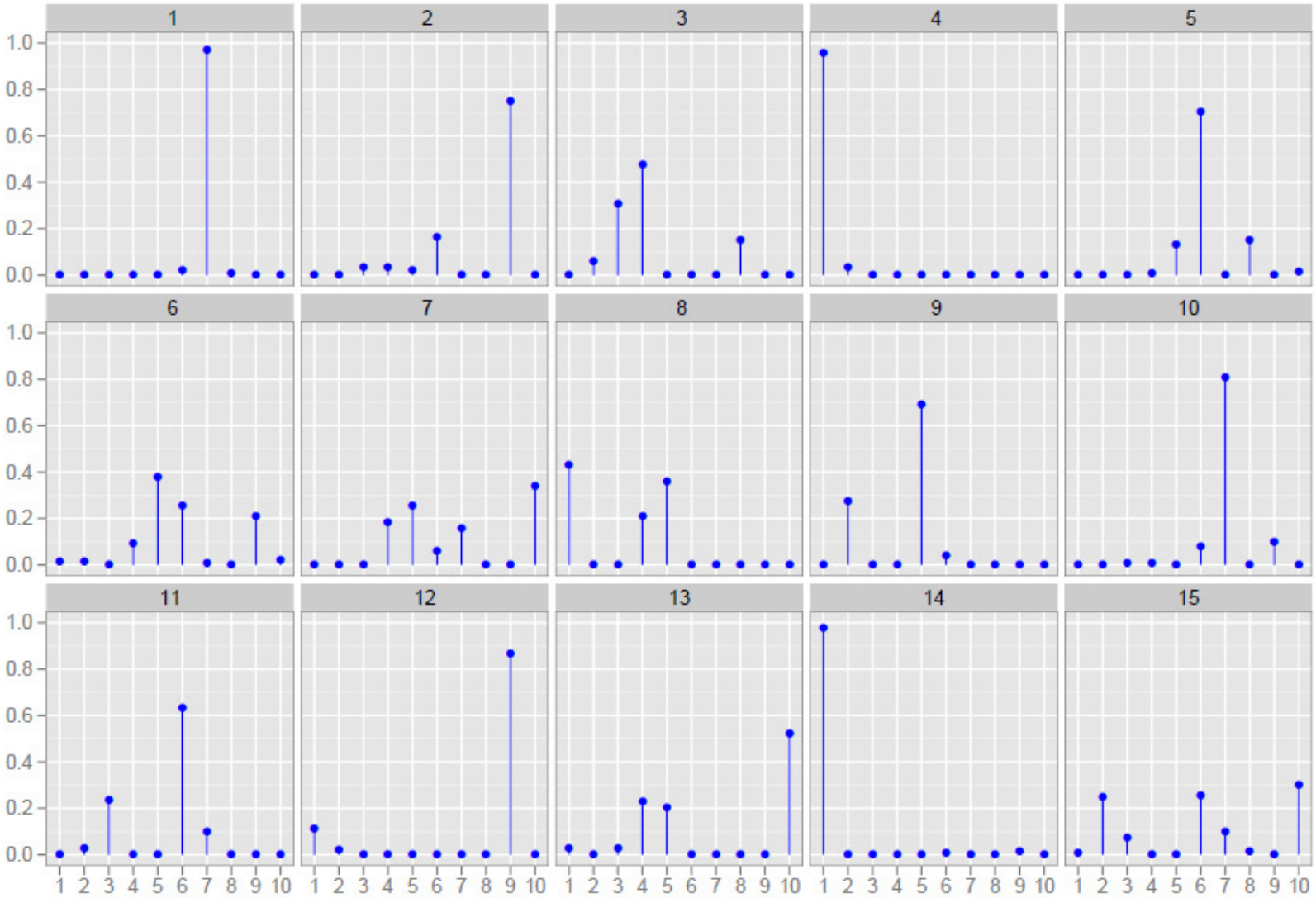
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 10$, 10 тем, 15 документов



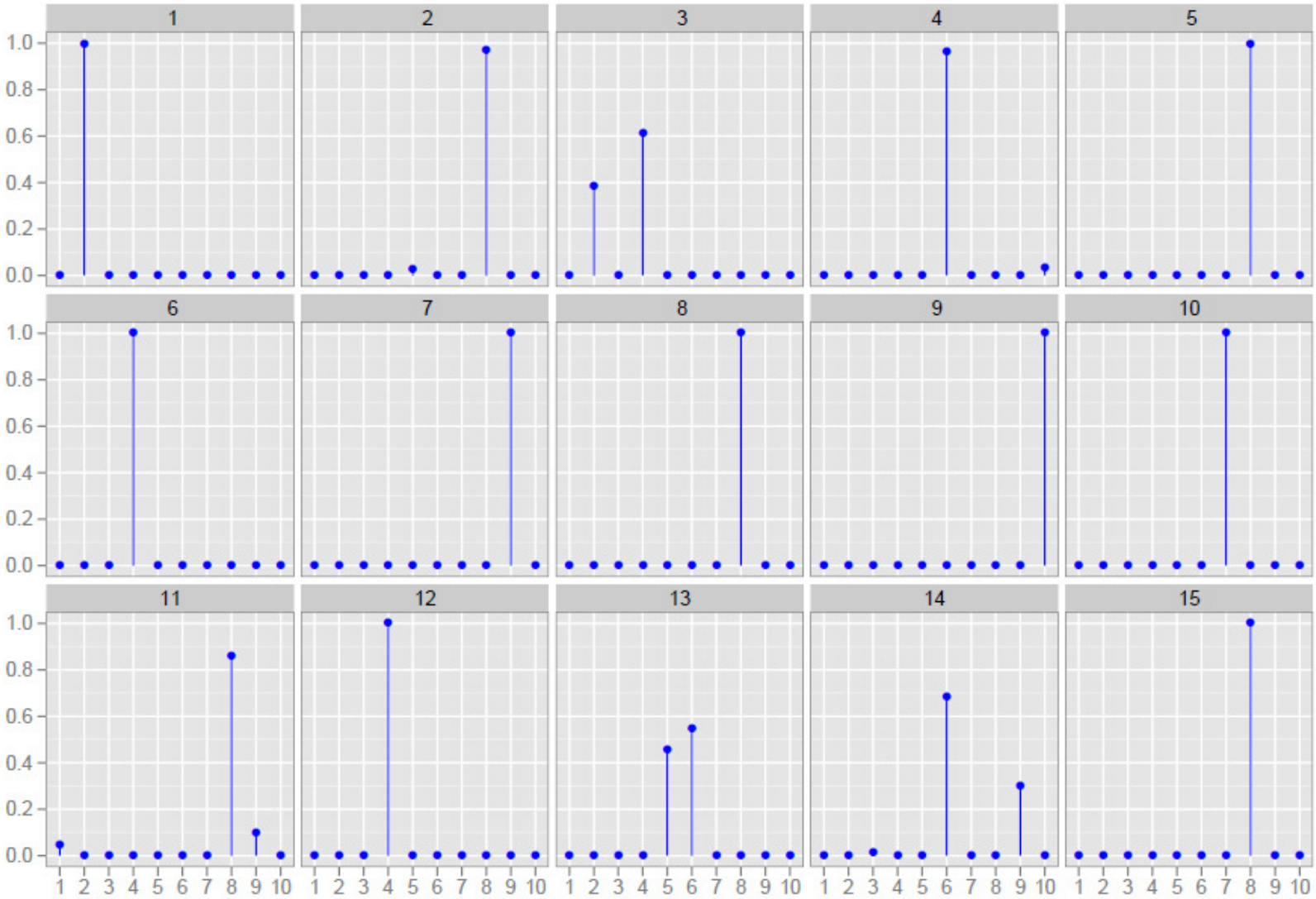
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 1$, 10 тем, 15 документов



Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 0.1$, 10 тем, 15 документов



Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 0.01$, 10 тем, 15 документов



Максимизация апостериорной вероятности для модели LDA

Совместное правдоподобие данных и модели:

$$\ln \prod_{d\in D}\prod_{w\in d} p(d,w|\Phi,\Theta)^{n_{dw}} \prod_{t\in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d|\alpha) \to \max_{\Phi,\Theta}$$

Принцип MAP (maximum a posteriori probability)

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + \\ + \sum_{t\in T}\sum_{w\in W} \ln \phi_{wt}^{\beta_w - 1} + \sum_{d\in D}\sum_{t\in T} \ln \theta_{td}^{\alpha_t - 1} \to \max_{\Phi,\Theta}$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t\in T} \theta_{td} = 1.$$

Регуляризованный EM-алгоритм для модели LDA

Максимум апостериорной вероятности ($\tilde{\beta}_w = \beta_w - 1$, $\tilde{\alpha}_t = \alpha_t - 1$):

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td}}_{\ln \text{ правдоподобия } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{t,w} \tilde{\beta}_w \ln \phi_{wt} + \sum_{d,t} \tilde{\alpha}_t \ln \theta_{td}}_{\text{критерий регуляризации } R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

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

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw} p_{tdw} + \tilde{\beta}_w\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw} p_{tdw} + \tilde{\alpha}_t\right) \end{cases}$$

где $\underset{t\in T}{\text{norm}}\, x_t = \dfrac{\max\{x_t, 0\}}{\sum_{s\in T} \max\{x_s, 0\}}$ — операция нормировки вектора.

EM-алгоритм для PLSA без регуляризации ($\tilde{\beta}_w = 0,\ \tilde{\alpha}_t = 0$)

Максимизация логарифма правдоподобия:

$$\mathscr{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \ \to \ \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$:

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

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

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

Основное различие EM-алгоритма для моделей PLSA и LDA

Оценки условных вероятностей $\phi_{wt} \equiv p(w|t)$, $\theta_{td} \equiv p(t|d)$:

- в PLSA — несмещённые оценки максимума правдоподобия:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad \theta_{td} = \frac{n_{td}}{n_d}$$

- в LDA — сглаженные байесовские оценки:

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \qquad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}$$

Различие проявляется только при малых n_{wt}, n_{td}, исчезая на больших данных и при использовании робастных моделей.

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.

Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных, 2013. — Т. 1, № 6. — С. 657–686.

Байесовские методы обучения параметров в модели LDA

- Сэмплирование Гиббса (GS — Gibbs Sampling)
  можно рассматривать как специальный случай EM
  — с обновлением по каждой словопозиции (n_{dw} раз);
  — с сэмплированием 1 темы для каждой словопозиции;
  — с регуляризацией Дирихле.
- VB, CVB — (Collapsed) Variational Bayesian inference
  можно рассматривать как специальный случай EM
  — без сэмплирования;
  — с регуляризацией Дирихле.

Griffiths T., Steyvers M. Finding scientific topics. 2004.

Teh Y. W., Newman D., Welling M. A collapsed variational Bayesian inference algorithm for Latent Dirichlet Allocation // NIPS-2006.

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.

Алгоритм сэмплирования Гиббса

Вход: коллекция D, число тем $|T|$, параметры α, β;
Выход: распределения Φ и Θ;

$n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
для всех $i := 1, \ldots, i_{\max}$ (итерация = один проход коллекции)
  для всех документов $d \in D$ и терминов $w_1, \ldots, w_{n_d} \in d$
    если $i \geqslant 2$ **то** $t := t_{dw}$; $--n_{wt}$; $--n_{td}$; $--n_t$; $--n_d$;
    $p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\frac{n_{wt}+\beta_w}{n_t+\beta_0} \cdot \frac{n_{td}+\alpha_t}{n_d+\alpha_0}\right)$ для всех $t \in T$;
    сэмплировать одну тему t из распределения $p(t|d,w)$;
    $t_{dw} := t$; $++n_{wt}$; $++n_{td}$; $++n_t$; $++n_d$;

$\phi_{wt} := n_{wt}/n_t$ для всех $w \in W$, $t \in T$;
$\theta_{td} := n_{td}/n_d$ для всех $d \in D$, $t \in T$;

Griffiths T., Steyvers M. Finding scientific topics. 2004.

Модель SWB с фоновой и шумовой компонентами

Гипотеза: каждый термин в документе (d, w)
— либо связан с какой-то темой t,
— либо специфичен для данного документа (шум),
— либо является общеупотребительным (фон).

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

$$p(w|d) = \frac{p_0(w|d) + \gamma\pi_{dw} + \varepsilon\pi_w}{1 + \gamma + \varepsilon}; \qquad p_0(w|d) = \sum_{t\in T} \phi_{wt}\theta_{td},$$

$\pi_{dw} \equiv p_{\text{ш}}(w|d)$ — *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 // Advances in Neural Information Processing Systems, MIT Press, 2006. — Vol. 19. — Pp. 241–248.

Робастная тематическая модель

Если PLSA не может предсказать слово w в документе d

$$p_0(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td} = 0,$$

то естественно считать такие слова w нетематическими:

$$p(w|d) = \nu_d p_0(w|d) + \big[p_0(w|d) = 0\big]\pi_{dw},$$

где $\pi_{dw} = \frac{n_w}{n}$ — униграммная модель языка.

Нормировочный множитель ν_d находится аналитически:

$$\nu_d = \sum_{w \in W} \big[p_0(w|d) > 0\big]\pi_{dw}$$

Внутренний критерий для сравнения моделей

Перплексия тестовой коллекции D' (hold-out perplexity):

$$\mathcal{P}(D') = \exp\left(-\frac{\sum\limits_{d\in D'}\sum\limits_{w\in d''} n_{dw}\ln p(w|d)}{\sum\limits_{d\in D'}\sum\limits_{w\in d''} n_{dw}}\right),$$

$d = d' \sqcup d''$ — 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''.

Интерпретация перплексии:

1) чем меньше, тем лучше;
2) если документ состоит из n_d равновероятных слов, то $\mathcal{P} = n_d$

Эксперименты с робастными PLSA и LDA

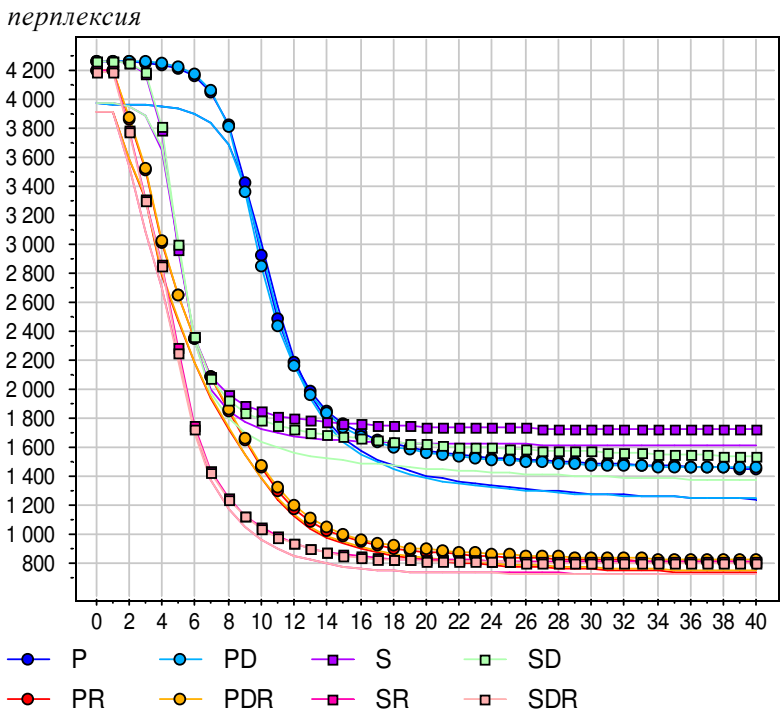
Использовались две коллекции:

- NIPS:
  - $|D| = 1566$ статей конференции NIPS на английском языке;
  - суммарной длины $n = 2.3\,\text{M}$,
  - словарь $|W| = 13\,\text{K}$.
  - Контрольная коллекция: $|D'| = 174$.
- RuDis:
  - $|D| = 2000$ авторефератов диссертаций на русском языке;
  - суммарной длины $n = 8.7\,\text{M}$,
  - словарь $|W| = 30\,\text{K}$.
  - Контрольная коллекция: $|D'| = 200$.

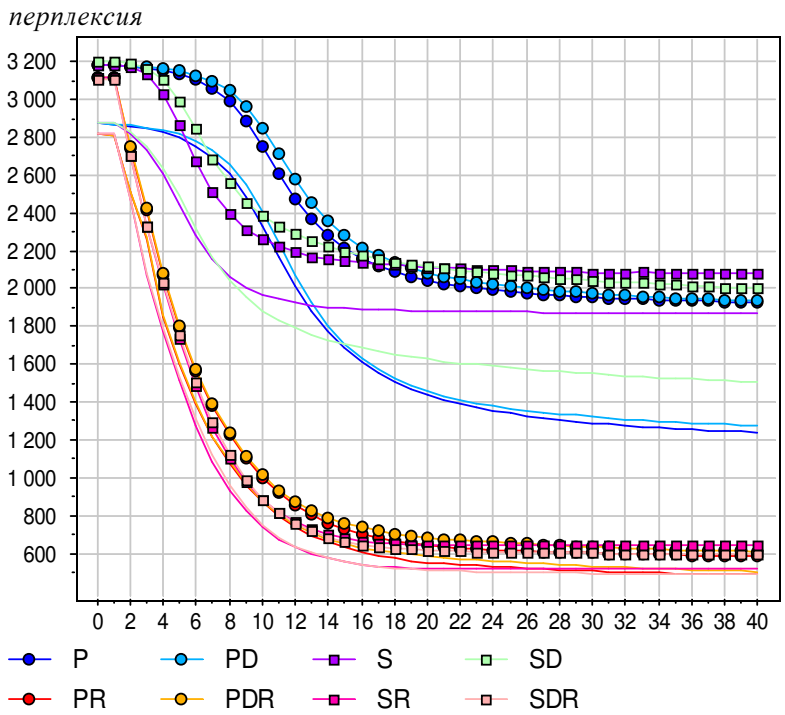
Предобработка: лемматизация, удаление стоп-слов.

Строятся графики зависимости перплексии от числа итераций (проходов коллекции); число итераций 40; число тем $|T| = 100$;

Эксперименты с робастными PLSA и LDA



RuDis



NIPS

Обозначения: P – PLSA
D – LDA ($\alpha_t = 0.5,\ \beta_w = 0.01$)
S – сэмплирование
R – робастность (шум $\gamma = 0.3$, фон $\varepsilon = 0.01$)

Выводы

1. LDA не уменьшает переобучение,
   а лишь точнее моделирует вероятности редких слов
2. Робастный PLSA лучше, чем LDA

Недостатки LDA:

1. Формально противоречит гипотезе разреженности
2. Не имеет убедительных лингвистических обоснований
3. Усложняет построение композитных моделей

Potapenko A. A., Vorontsov K. V., Robust PLSA Performs Better Than LDA // European Conference on Information Retrieval ECIR-2013, Moscow, 24–27 March 2013. — Pp. 784–787.

Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных, 2013. — Т. 1, № 6. — С. 657–686.

Расстояние между распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$

- Среднеквадратичное отклонение не подходит!
- Дивергенция Кульбака-Лейблера (несимметричная):

  $\mathrm{KL}(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i}$
- Симметричные расстояния:

  $\mathrm{KL}(P,Q) = \frac{1}{2}(P\|Q) + \frac{1}{2}(Q\|P)$ — Кульбака-Лейблера

  $\mathrm{JS}(P,Q) = \frac{1}{2}\mathrm{KL}\left(P\|\frac{P+Q}{2}\right) + \frac{1}{2}\mathrm{KL}\left(Q\|\frac{P+Q}{2}\right)$ — Йенсена-Шеннона

  $H^2(P,Q) = \frac{1}{2}\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$ — Хеллингера
- Статистика χ^2:

  $X^2(P,Q) = \sum_i \frac{(p_i - q_i)^2}{q_i}$

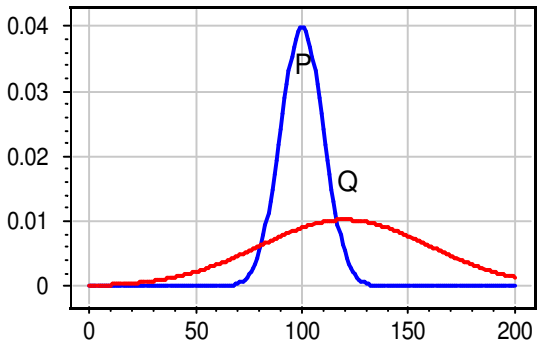
Свойства дивергенции Кульбака–Лейблера

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

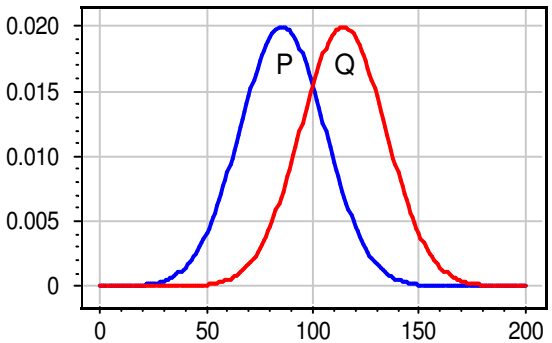
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

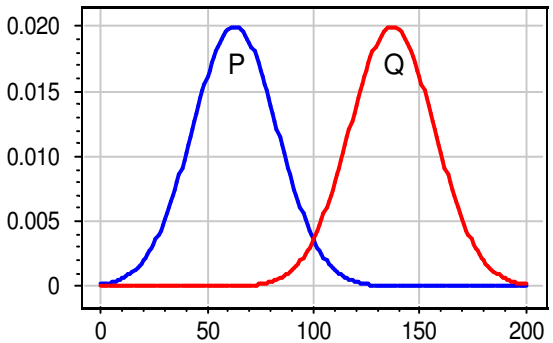
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P вложено в Q:



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



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



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

Цель эксперимента

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:

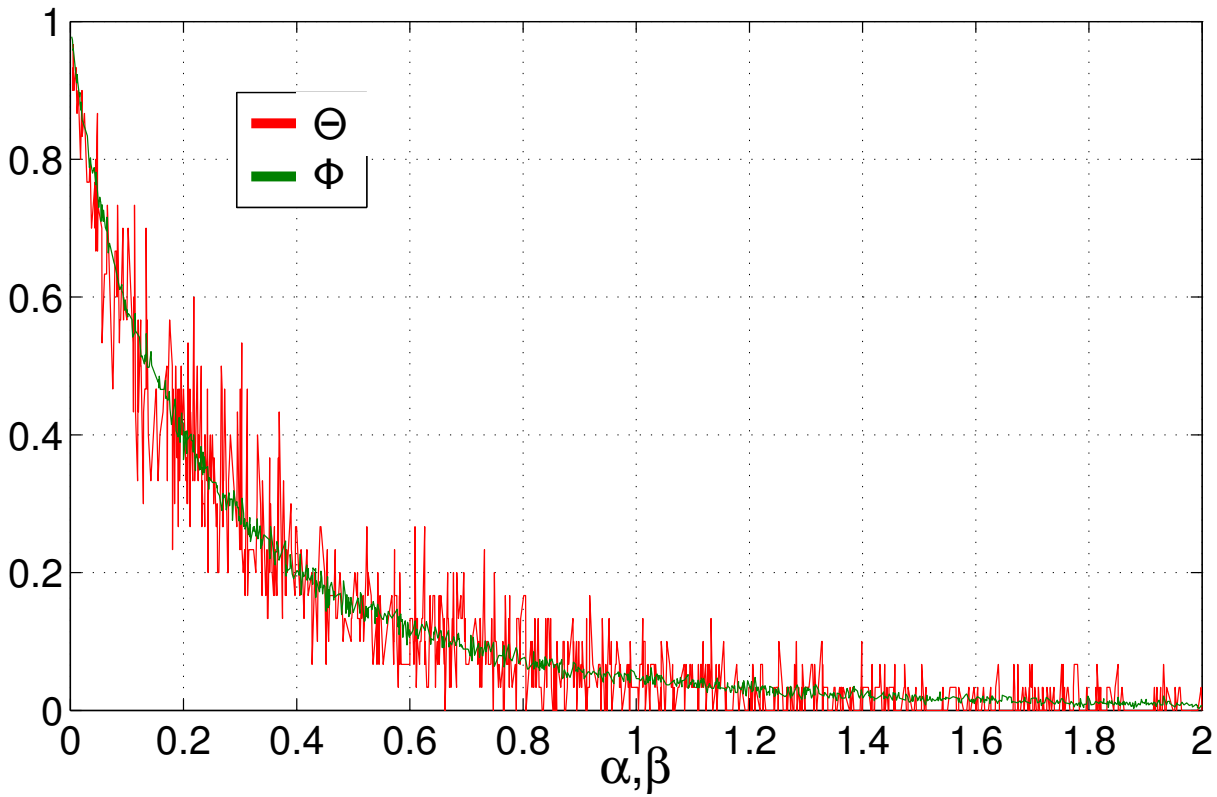
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

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

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

Генерация синтетических данных из распределения Дирихле

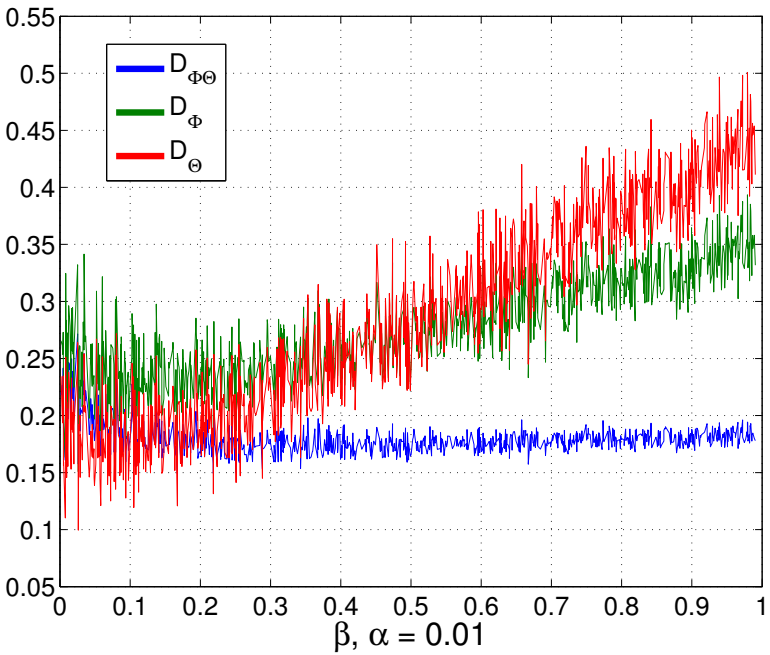
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \mathrm{Dir}(\alpha)$ и $\phi_t^0 \sim \mathrm{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



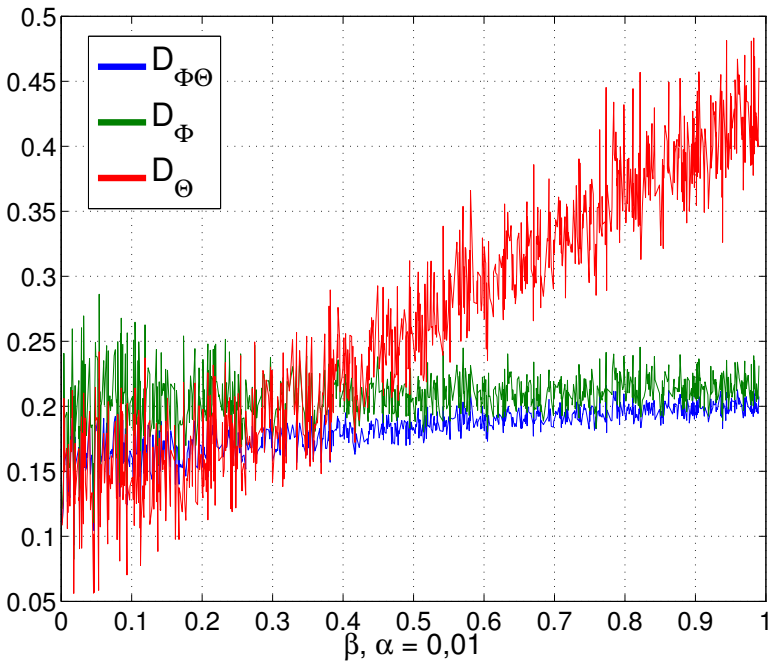
Неустойчивость восстановления Φ, Θ

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

PLSA



LDA

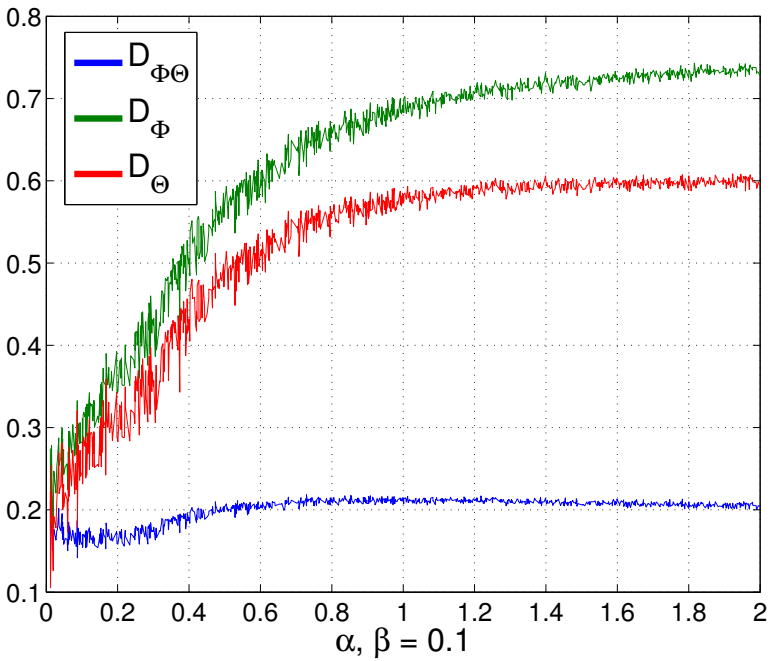


Виталий Глушаченков. Устойчивость матричных разложений в задачах тематического моделирования // Магистерская диссертация. МФТИ, 2013.

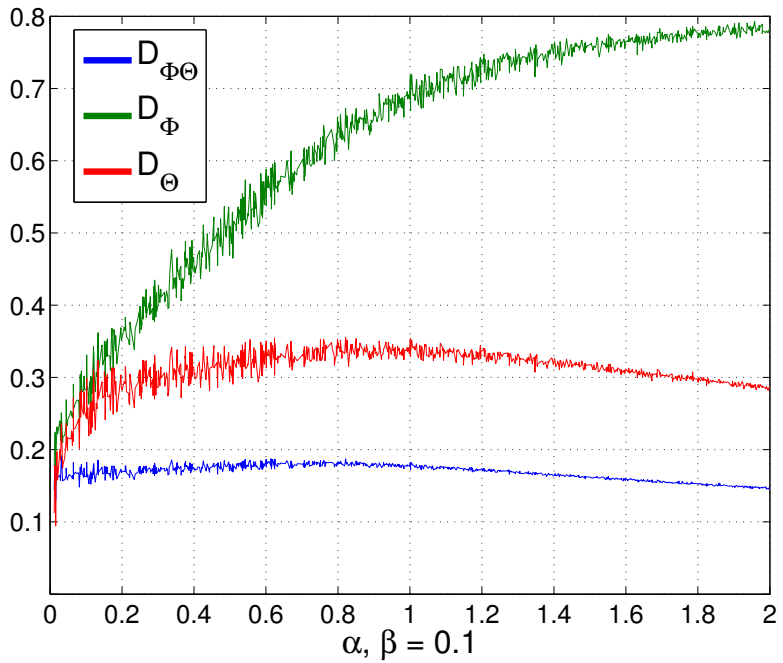
Неустойчивость восстановления Φ, Θ

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 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.

Измеряется нормированная KL-близость пар тем t и s:

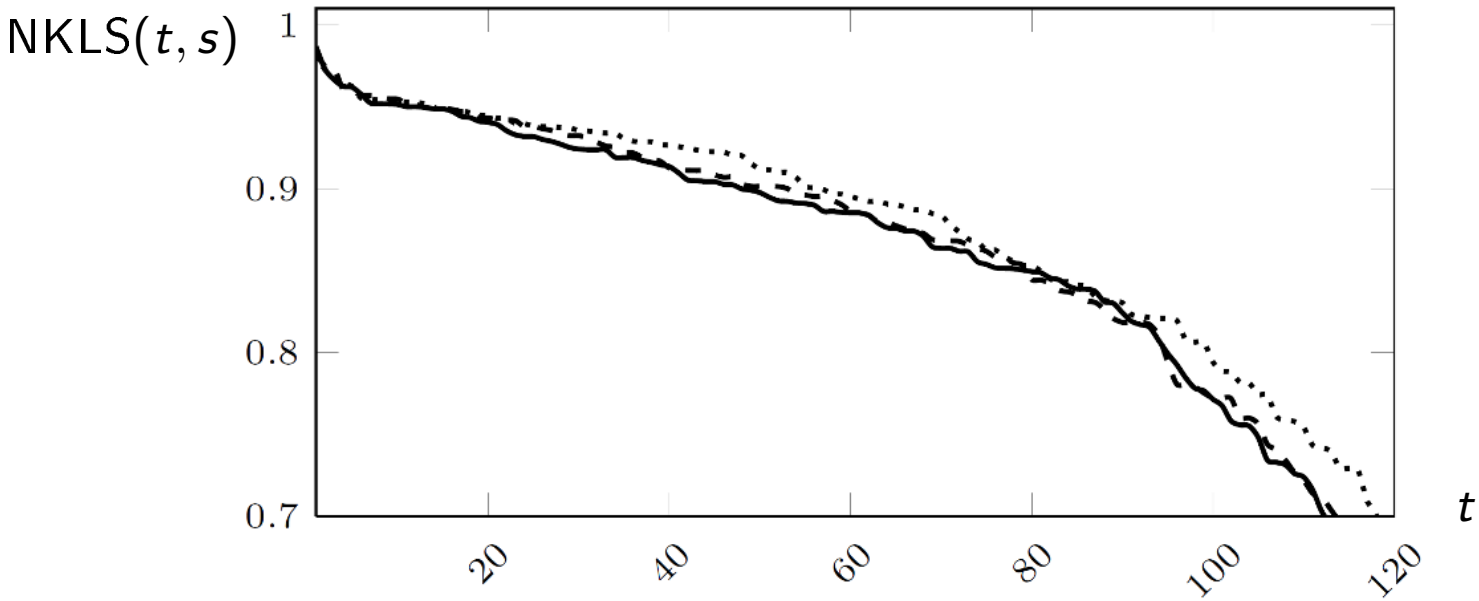
$$\text{NKLS}(t,s) = \left(1 - \frac{\text{KL}(t,s)}{\max\limits_{t',s'} \text{KL}(t',s')}\right)$$

При $\text{NKLS}(t,s) > 0.9$ 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-GS 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. Volume 101, Issue 1-3 Pp. 303–323.

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

Резюме

- Задача тематического моделирования некорректно поставлена, её решение не единственно и не устойчиво.
- Регуляризация — стандартный приём решения таких задач.
- EM-алгоритм LDA отличается от PLSA сглаженными частотными оценками условных вероятностей.
- Модель LDA лучше описывает вероятности редких слов, но для выявления тематики они как раз не важны.
- Модель LDA — недостаточно сильный регуляризатор, она не решает проблему неединственности и неустойчивости.