Вероятностные тематические модели коллекций текстовых документов

По мотивам лекции Воронцова К. В.

ноябрь 2017

Содержание

Понятие «латентной темы»

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.
- *Тема* — вероятностное распределение на терминах: $p(w|t)$ — вероятность встретить термин w в теме t.

Документ имеет ненаблюдаемый *тематический профиль*: $p(t|d)$ — неизвестная частота темы t в документе d.

Когда автор писал термин w в документ d, он думал о теме t.

Документ d состоит из наблюдаемых терминов $w_1, \ldots, w_{n_d}$, $p(w|d)$ — известная частота термина w в документе d.

Тематическая модель пытается выявить латентные темы.

Цели и приложения тематического моделирования

- Выявить скрытую тематическую структуру коллекции текстов
- Выявить тематический профиль каждого документа

Приложения:

- Семантический поиск по текстовому запросу любой длины
- Категоризация, классификация, аннотирование, суммаризация, сегментация текстовых документов
- Поиск научной информации, трендов, фронта исследований
- Поиск специалистов (expert search), рецензентов, проектов
- Анализ и агрегирование новостных потоков
- Рубрикация документов, изображений, видео, музыки
- Рекомендующие системы, коллаборативная фильтрация
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов

Основные предположения

1. Порядок документов в коллекции не важен
2. Порядок слов в документе не важен (bag of words)
3. Слова, встречающиеся «почти во всех» документах, не важны
4. Слово в разных формах — это одно и то же слово
5. Документ обычно относится к небольшому числу тем
6. Тема обычно определяется небольшим числом терминов

Предварительная обработка текстов:

- Приведение всех слов к нормальной форме (лемматизация или стемминг)
- Выделение терминов (term extraction)
- Удаление стоп-слов и слишком редких слов

Вероятностная формализация постановки задачи

Формализация основных предположений:

- каждое слово в документе связано с некоторой темой $t \in T$
- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция — это выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d,t) = p(w|t)$

Вероятностная модель порождения документа d:

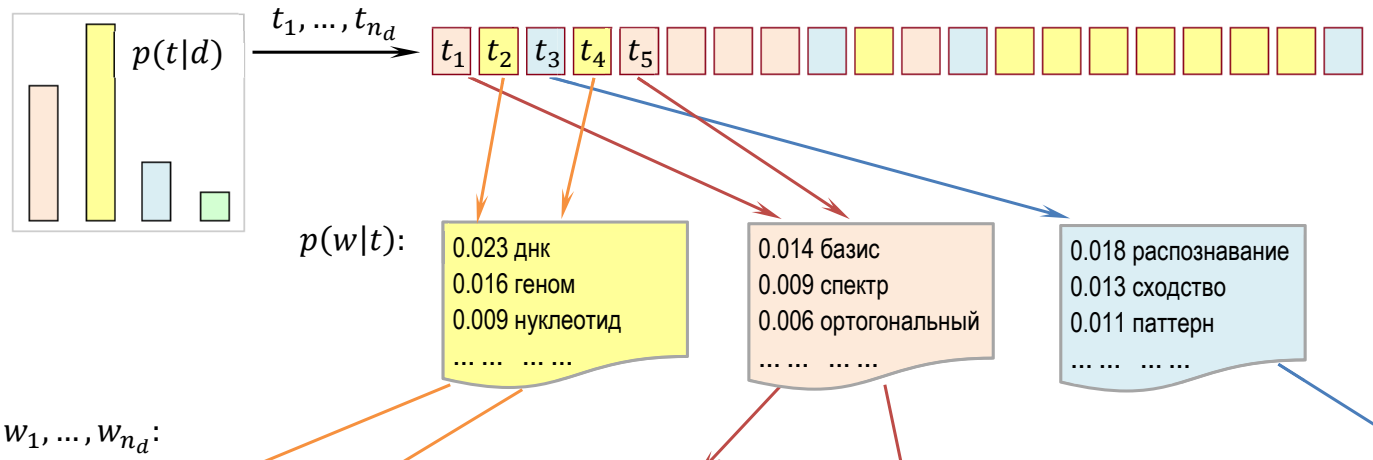
$$p(w|d) = \sum_{t \in T} p(w|d,t)\, p(t|d) = \boxed{\sum_{t \in T} p(w|t)\, p(t|d)}$$

Дано $\hat{p}(w|d) = n_{dw}/n_d$, **найти:**

- $\phi_{wt} \equiv p(w|t)$ — распределение терминов в темах $t \in T$;
- $\theta_{td} \equiv p(t|d)$ — распределение тем в документах $d \in D$.

Вероятностная модель порождения документа

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Принцип максимума правдоподобия

Правдоподобие — это плотность распределения выборки D:

$$p(D) = \prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}},$$

где n_{dw} — число вхождений термина w в документ d.

Пусть $p(w|d, \alpha)$ — параметрическая вероятностная модель документа d, зависящая от вектора параметров $\alpha = (\Phi, \Theta)$.

Логарифм правдоподобия выборки D:

$$\log p(D, \alpha) = \sum_{d \in D} \sum_{w \in d} n_{dw} \log p(w|d, \alpha) p(d) \to \max_{\alpha}.$$

Избавимся от $p(d)$, не влияющего на точку максимума:

$$L(D, \alpha) = \sum_{d \in D} \sum_{w \in d} n_{dw} \log p(w|d, \alpha) \to \max_{\alpha}.$$

Модель PLSA (Probabilistic Latent Semantic Analysis)

Задача: найти максимум правдоподобия

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

при ограничениях неотрицательности и нормировки

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

Интерпретация: стохастическое матричное разложение

$$F \approx \Phi\Theta,$$

$F = \big(\hat{p}(w|d)\big)_{W\times D}$ — известная матрица исходных данных,
$\Phi = \big(\phi_{wt}\big)_{W\times T}$ — искомая матрица терминов тем $\phi_{wt} = p(w|t)$,
$\Theta = \big(\theta_{td}\big)_{T\times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999. Pp. 50–57

Необходимые условия точки максимума правдоподобия

Теорема

Точка максимума правдоподобия Φ, Θ удовлетворяет системе уравнений со вспомогательными переменными n_{dwt}:

$$\text{E-шаг:} \quad n_{dwt} = n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}};$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \dfrac{n_{wt}}{n_t}; \quad n_{wt} = \sum\limits_{d \in D} n_{dwt}; \quad n_t = \sum\limits_w n_{wt} \\ \theta_{td} = \dfrac{n_{td}}{n_d}; \quad n_{td} = \sum\limits_{w \in d} n_{dwt}; \quad n_d = \sum\limits_t n_{td} \end{cases}$$

EM-алгоритм — это чередование шагов E и M до сходимости, т. е. решение системы уравнений методом простых итераций.

EM-алгоритм. Элементарная интерпретация

EM-алгоритм — это чередование E и M шагов до сходимости.

E-шаг: условные вероятности тем $p(t|d,w)$ для всех t,d,w вычисляются через ϕ_{wt}, θ_{td} по формуле Байеса:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}.$$

M-шаг: частотные оценки условных вероятностей вычисляются путём суммирования счётчика $n_{dwt} = n_{dw}p(t|d,w)$:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d\in D} n_{dwt}, \qquad n_t = \sum_{w\in W} n_{wt};$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \qquad n_{td} = \sum_{w\in d} n_{dwt}, \qquad n_d = \sum_{t\in T} n_{td}.$$

Частотные оценки условных вероятностей

Если рассматривать коллекцию как выборку троек (d, w, t), то

$$\hat{p}(w|d) = \frac{n_{dw}}{n_d}, \qquad \hat{p}(w|t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(t|d) = \frac{n_{td}}{n_d};$$

n_{dwt} — число троек (d, w, t) во всей коллекции

$n_{dw} = \sum_t n_{dwt}$ — частота термина w в документе d

$n_{wt} = \sum_d n_{dwt}$ — число употреблений термина w в теме t

$n_{td} = \sum_w n_{dwt}$ — число терминов темы t в документе d

$n_w = \sum_{d,t} n_{dwt}$ — число употреблений термина w в коллекции

$n_t = \sum_{d,w} n_{dwt}$ — число терминов темы t в коллекции

$n_d = \sum_{w,t} n_{dwt}$ — длина документа d

$n = \sum_{d,w,t} n_{dwt}$ — длина коллекции

Рациональный EM-алгоритм

Проблема: необходимость хранить 3D-матрицу n_{dwt}
Идея: E-шаг встраивается внутрь M-шага

Вход: коллекция D, число тем $|T|$, число итераций $i_{\max}$;
Выход: матрицы терминов тем Θ и тем документов Φ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D,\ w \in W,\ t \in T$;
для всех итераций $i = 1, \ldots, i_{\max}$
  $n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D,\ w \in W,\ t \in T$;
  для всех документов $d \in D$ и всех слов $w \in d$
    $n_{dwt} := n_{dw} \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ для всех $t \in T$;
    $n_{wt}, n_{td}, n_t, n_d \mathrel{+}= n_{dwt}$ для всех $t \in T$;
  $\phi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := n_{td}/n_d$ для всех $d \in D,\ t \in T$;

Недостатки PLSA-EM и способы их устранения

1. неединственность и неустойчивость решения

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

для любых $S_{T\times T}$ таких, что Φ', Θ' — стохастические.
— регуляризация: доопределение постановки задачи

2. на малых коллекциях возможно переобучение
— регуляризация: снижение эффективной размерности: сглаживание (LDA), разреживание и др.

3. нет управления *разреженностью* Φ и Θ, т.к.
(в начале $\phi_{wt} = 0$) $\Leftrightarrow$ (в финале $\phi_{wt} = 0$),
(в начале $\theta_{td} = 0$) $\Leftrightarrow$ (в финале $\theta_{td} = 0$)
— регуляризация: постепенное разреживание

4. нет выделения нетематических слов
— регуляризация: сглаживание фоновых тем

Модель LDA (Latent Dirichlet Allocation)

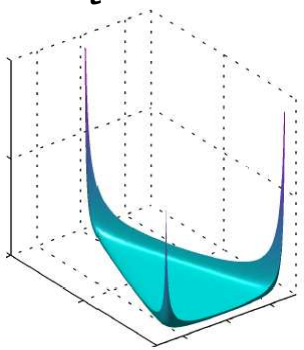
Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

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

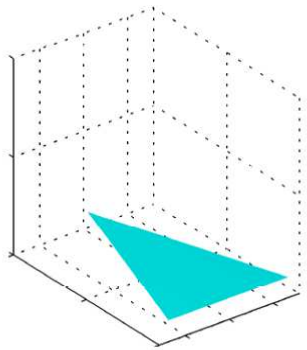
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t \geqslant 0;$$

Пример:

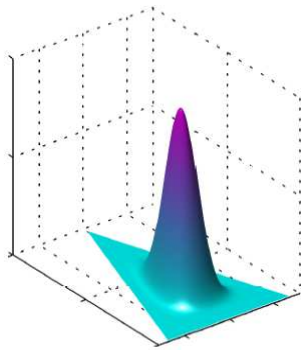
$\mathrm{Dir}(\theta|\alpha)$
$|T| = 3$
$\theta, \alpha \in \mathbb{R}^3$



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



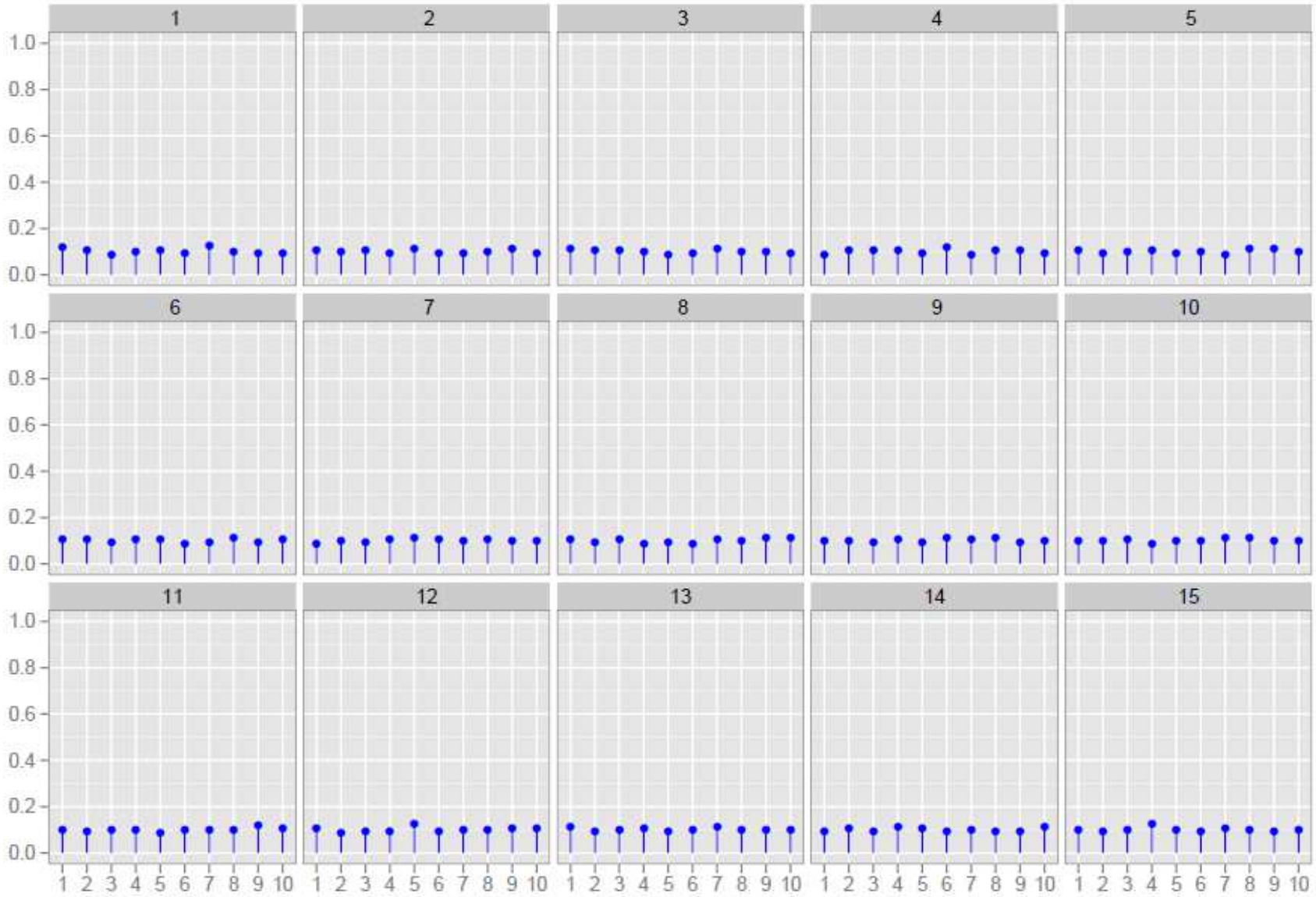
$\alpha_1 = \alpha_2 = \alpha_3 = 1$



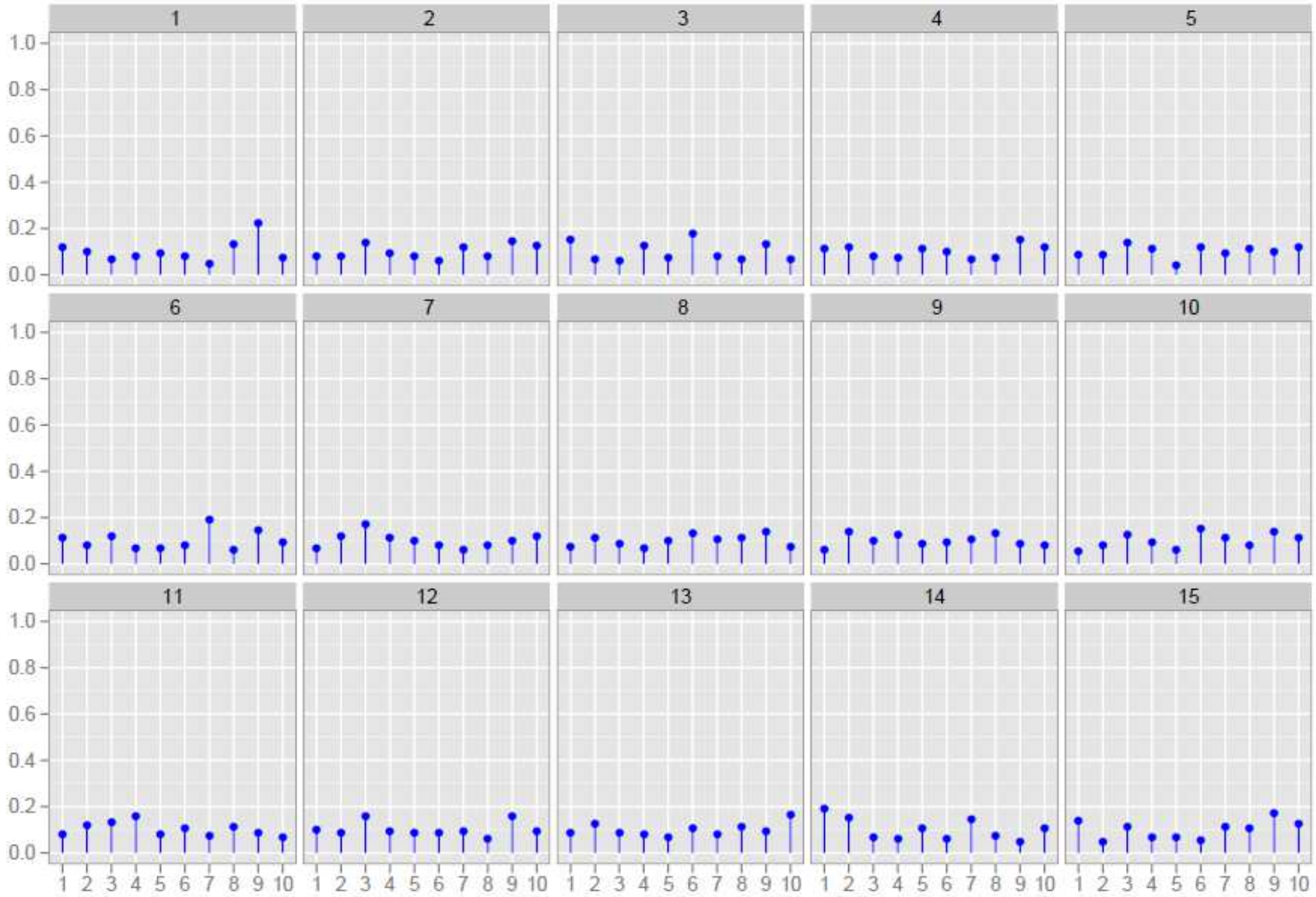
$\alpha_1 = \alpha_2 = \alpha_3 = 10$

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

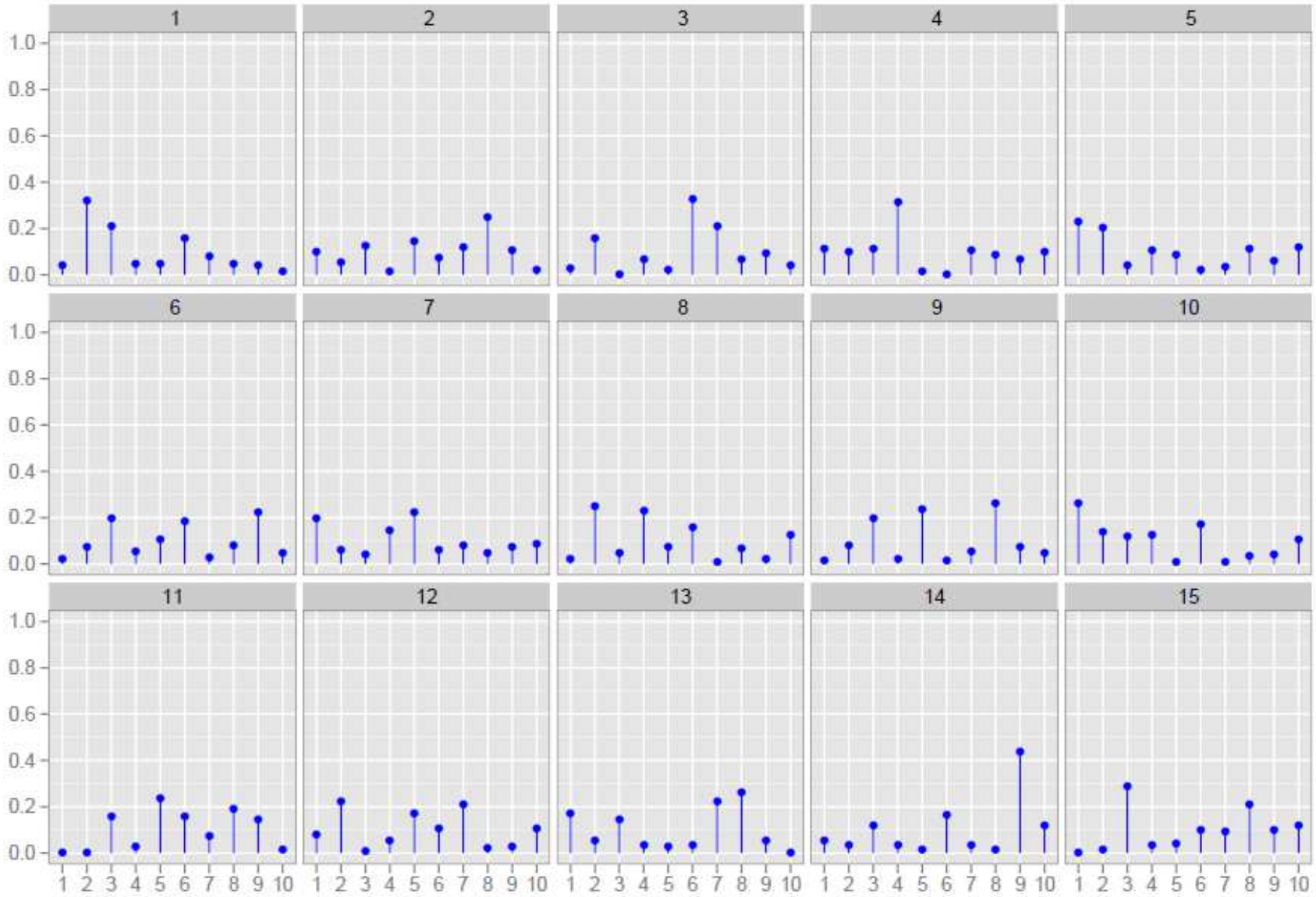
Распределение Дирихле при $\alpha_t \equiv 100$, 10 тем, 15 документов



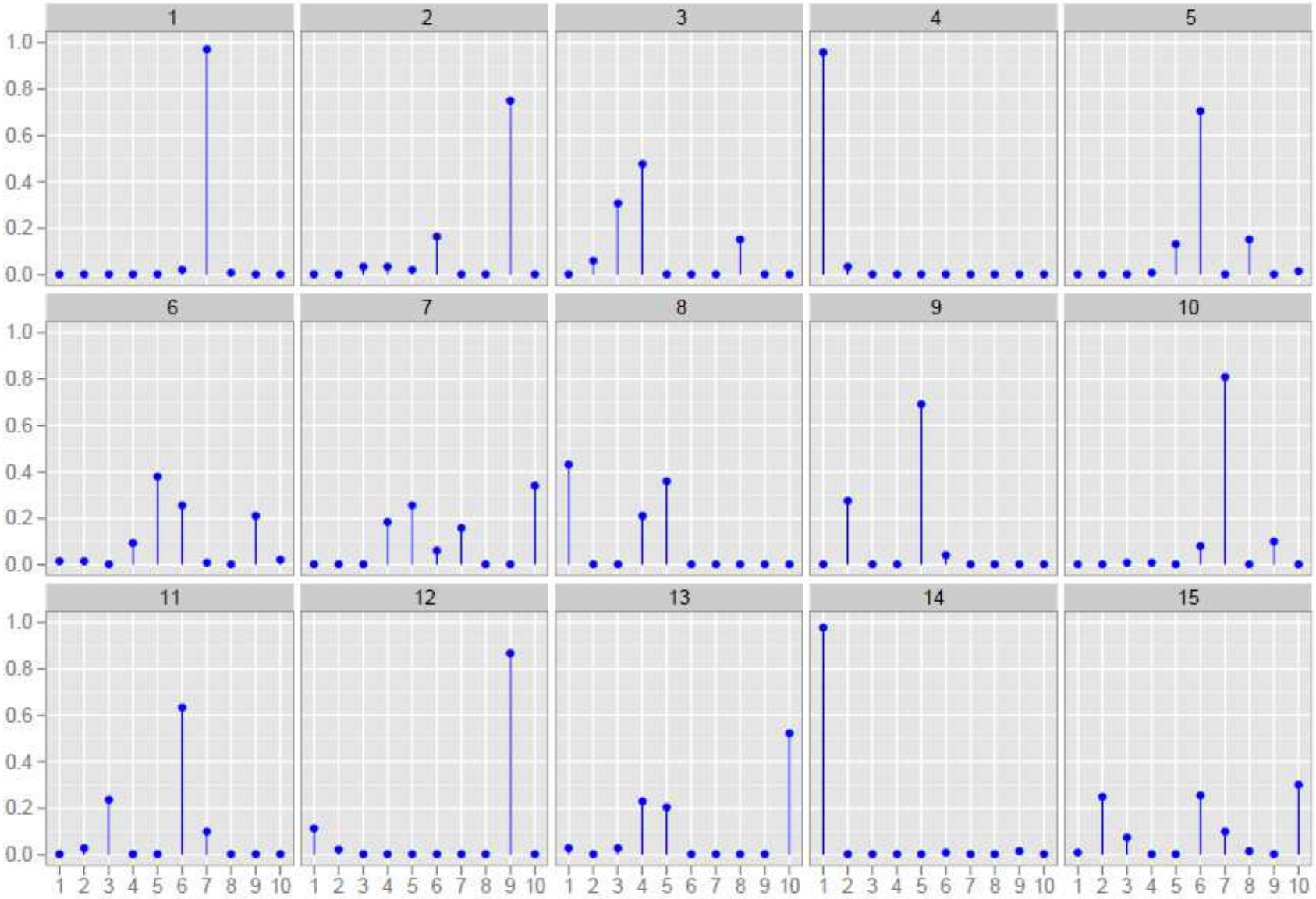
Распределение Дирихле при $\alpha_t \equiv 10$, 10 тем, 15 документов



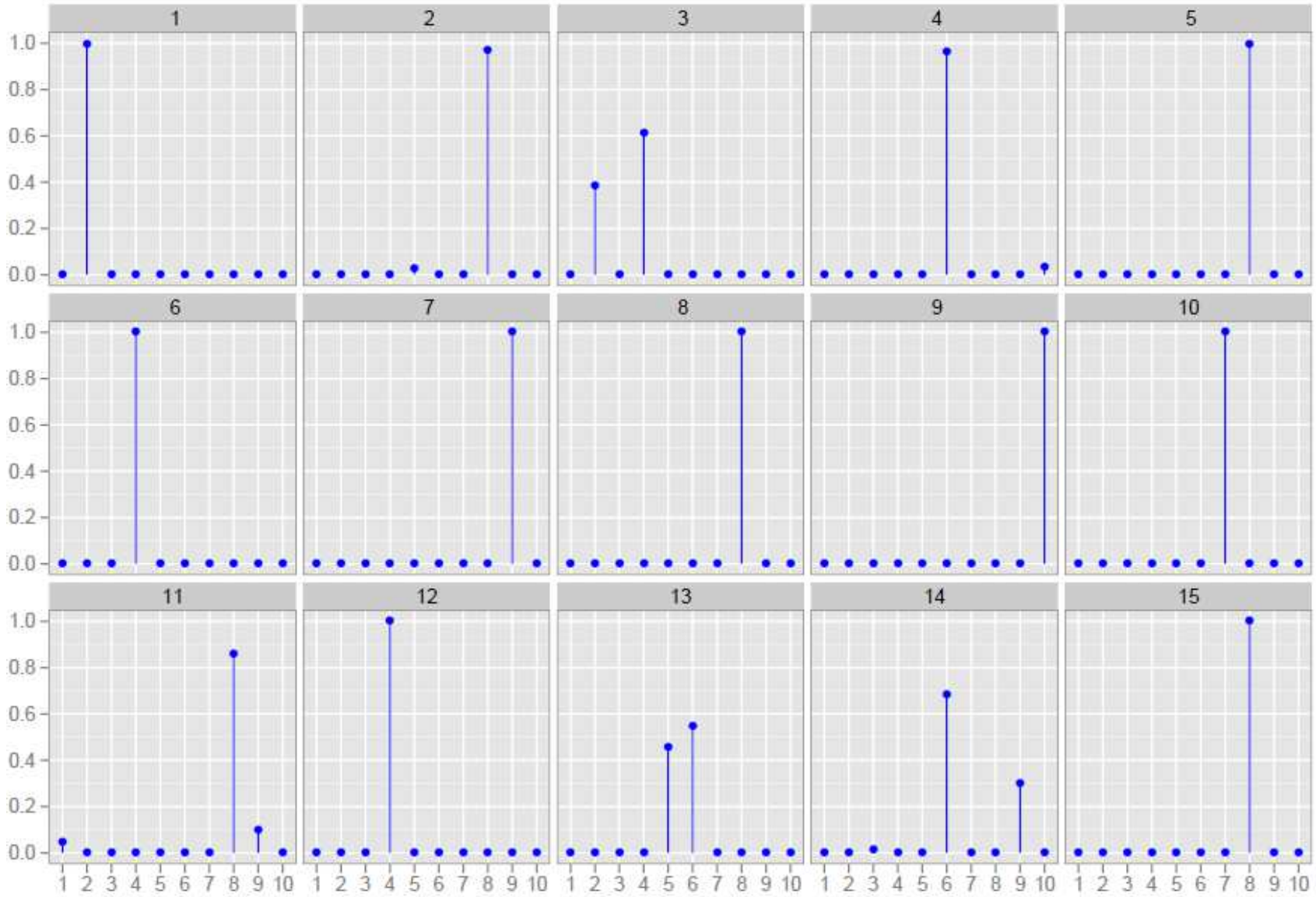
Распределение Дирихле при $\alpha_t \equiv 1$, 10 тем, 15 документов



Распределение Дирихле при $\alpha_t \equiv 0.1$, 10 тем, 15 документов



Распределение Дирихле при $\alpha_t \equiv 0.01$, 10 тем, 15 документов



Главное отличие LDA от PLSA

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}.
Робастные LDA и PLSA почти одинаковы по качеству.

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.

Potapenko A. A., Vorontsov K. V. Robust PLSA Performs Better Than LDA. ECIR-2013, Moscow, Russia, 24-27 March 2013. LNCS, Springer. Pp. 784–787.

Неустойчивость! Эксперимент на модельных данных

Модельные коллекции порождаются заданными матрицами Φ_0 и Θ_0 при $|D| = 500$, $|W| = 1000$, $|T| = 30$, $n_d \in [100, 600]$.

Отклонение восстановленных распределений $p(i|j)$ от исходных модельных распределений $p_0(i|j)$ измеряются средним расстоянием Хеллингера:

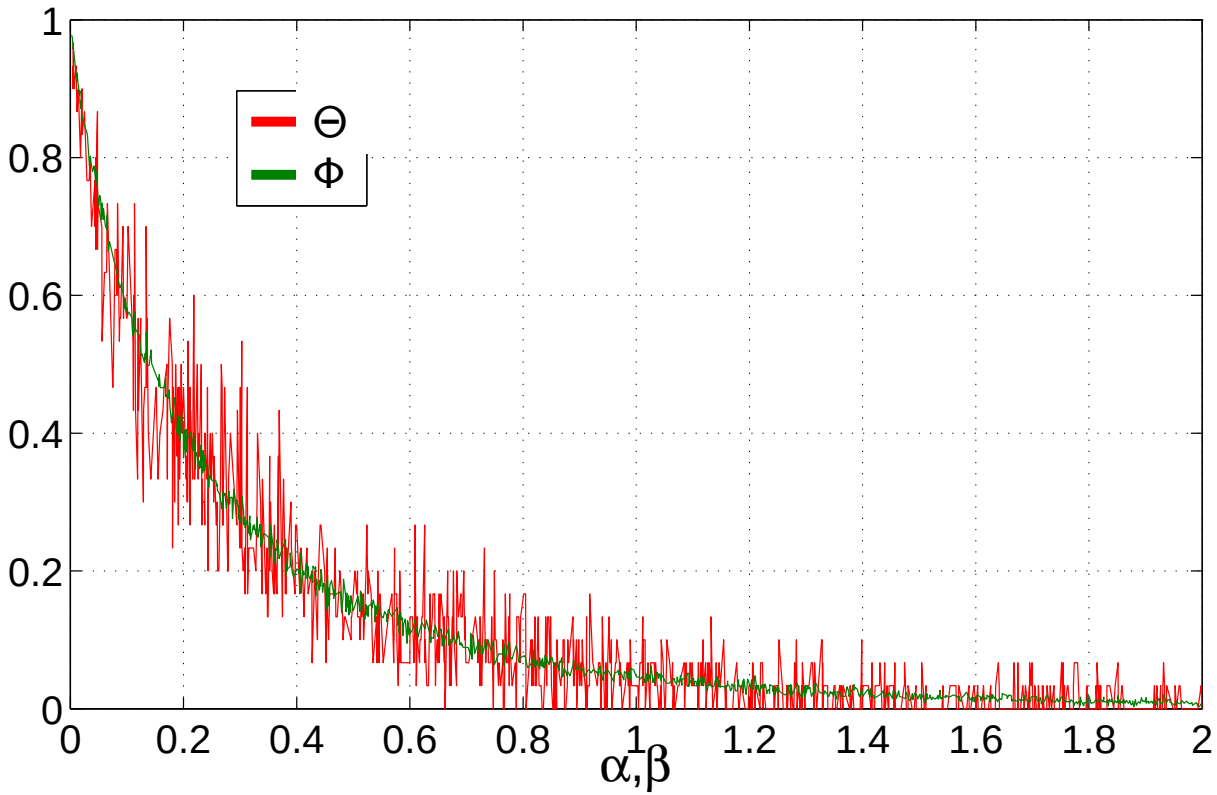
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

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 модельных данных различной разреженности

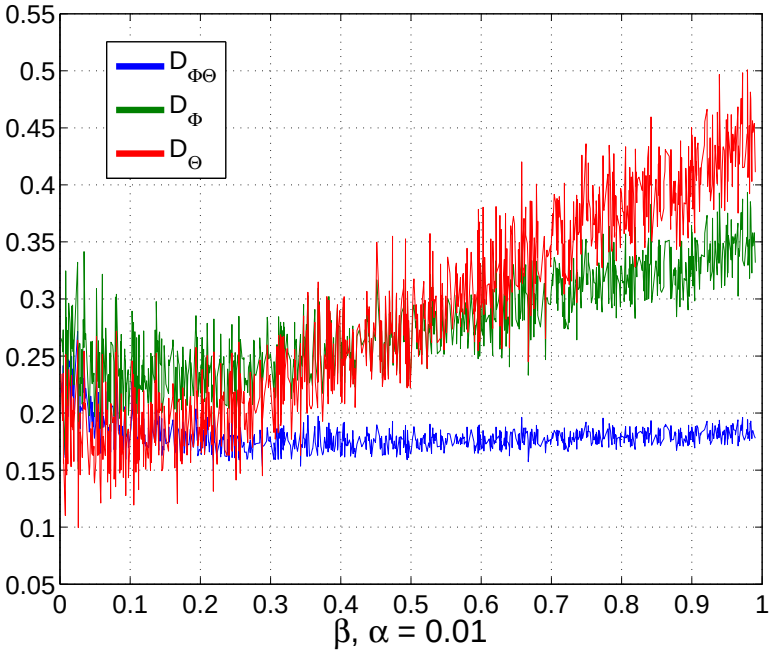
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \text{Dir}(\alpha)$ и $\phi_t^0 \sim \text{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



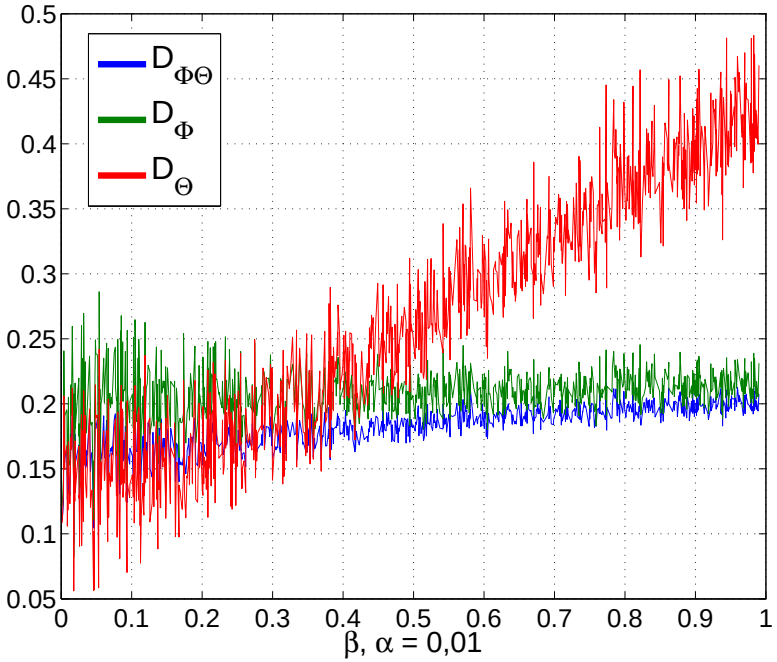
Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от параметра β, т.е. от НЕразреженности матрицы Φ_0

PLSA



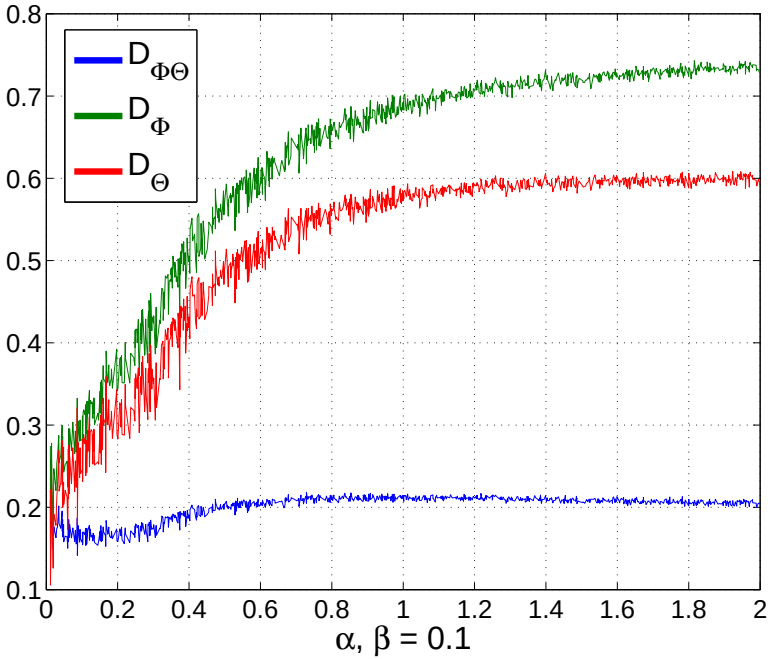
LDA



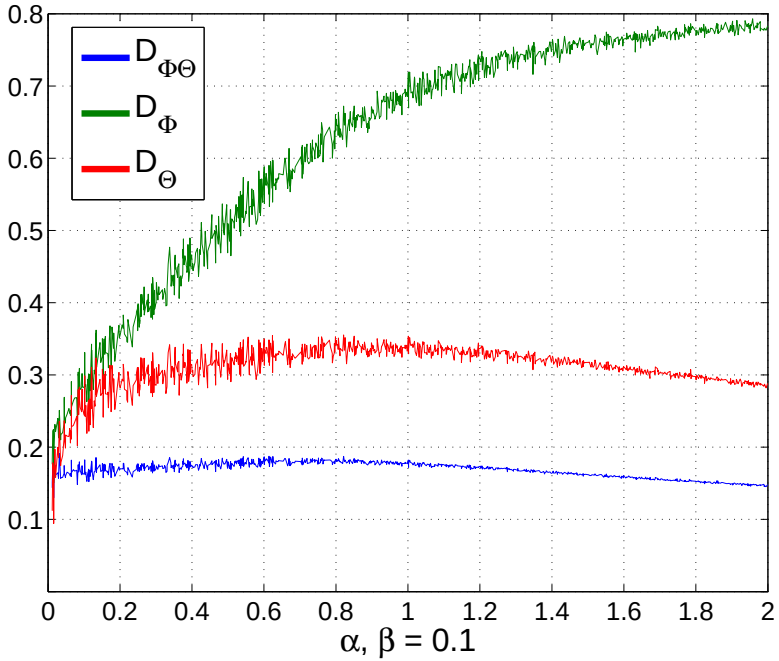
Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от параметра α, т.е. от НЕразреженности матрицы Θ_0

PLSA



LDA



Выводы

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. Задача некорректно поставлена, нет единственности: для любых $S_{T\times T}$ таких, что Φ', Θ' — стохастические,
$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'.$$
4. Поэтому необходима регуляризация, однако распределение Дирихле — слишком слабый регуляризатор

Реализация экспериментов:
Виталий Глушаченков. Магистерская диссертация. МФТИ, 2013.

ARTM — аддитивная регуляризация тематической модели

Пусть, наряду с правдоподобием, требуется максимизировать ещё n критериев $R_i(\Phi, \Theta)$, $i = 1, \ldots, n$ — *регуляризаторов*.

Метод многокритериальной оптимизации — *скаляризация*.

Задача: максимизировать регуляризованное правдоподобие

$$\underbrace{\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\text{log-likelihood } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

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

где $\tau_i > 0$ — *коэффициенты регуляризации*.

Регуляризованный EM-алгоритм

Теорема

Если Φ, Θ — решение задачи максимизации регуляризованного правдоподобия, то оно удовлетворяет системе уравнений

$$\begin{cases} n_{dwt} = n_{dw} \dfrac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}}; \\ \phi_{wt} = \dfrac{n_{wt}}{n_t}; \quad n_{wt} = \left(\sum\limits_{d \in D} n_{dwt} + \phi_{wt} \dfrac{\partial R}{\partial \phi_{wt}} \right)_+; \quad n_t = \sum\limits_{w \in W} n_{wt}; \\ \theta_{td} = \dfrac{n_{td}}{n_d}; \quad n_{td} = \left(\sum\limits_{w \in d} n_{dwt} + \theta_{td} \dfrac{\partial R}{\partial \theta_{td}} \right)_+; \quad n_d = \sum\limits_{t \in T} n_{td} \end{cases}$$

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

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

Напоминания. Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_x; \\ g_i(x) \leqslant 0, \quad i = 1, \ldots, m; \\ h_j(x) = 0, \quad j = 1, \ldots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители μ_i, $i = 1, \ldots, m$, λ_j, $j = 1, \ldots, k$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0; \ h_j(x) = 0; \ \text{(исходные ограничения)} \\ \mu_i \geqslant 0; \ \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; \ \text{(условие дополняющей нежёсткости)} \end{cases}$$

Доказательство Теоремы о регуляризации M-шага

1. Условия ККТ для ϕ_{wt} (для θ_{td} всё аналогично):

$$\sum_d n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \lambda_{wt}; \quad \lambda_{wt} \geqslant 0; \quad \lambda_{wt}\phi_{wt} = 0.$$

2. Умножим обе части равенства на ϕ_{wt} и выделим n_{dwt}:

$$\phi_{wt}\lambda_t = \sum_d n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}.$$

3. Предположим, что тема t *регулярна*: $\exists w$: $n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} > 0$

$$\phi_{wt}\lambda_t = \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

4. Суммируем обе части равенства по $w \in W$:

$$\lambda_t = \sum_{w \in W} \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

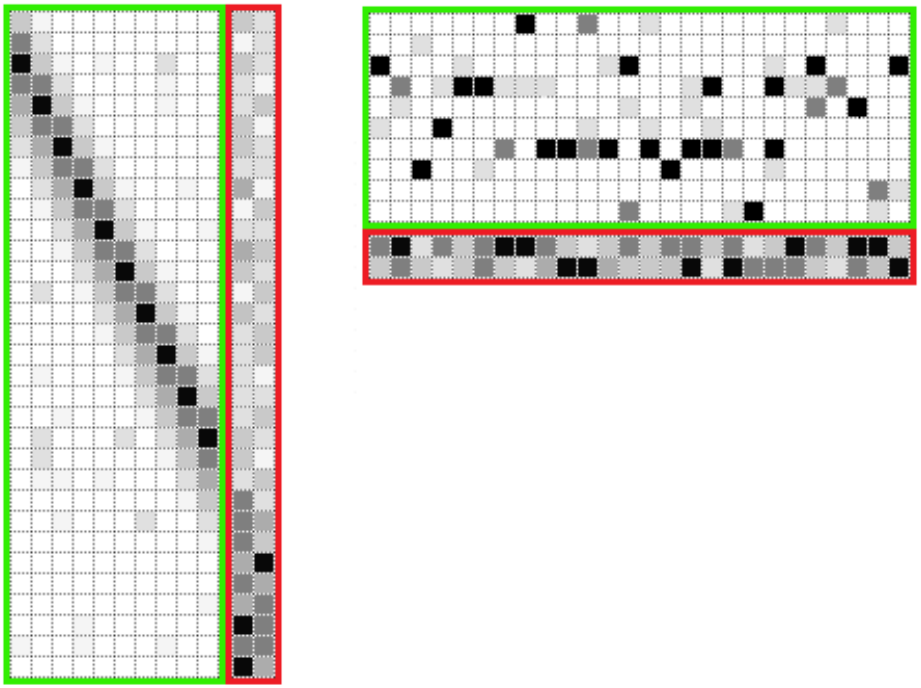
5. Подставим λ_t из (4) в (3), получим требуемое. ■

Требования интерпретируемости и гипотезы о структуре тем

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

Фоновые темы B содержат слова общей лексики, $p(w|t)$ и $p(t|d)$ не разреженные в этих темах

$\Phi_{W \times T}$ $\Theta_{T \times D}$



Напоминания. Дивергенция Кульбака–Лейблера

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

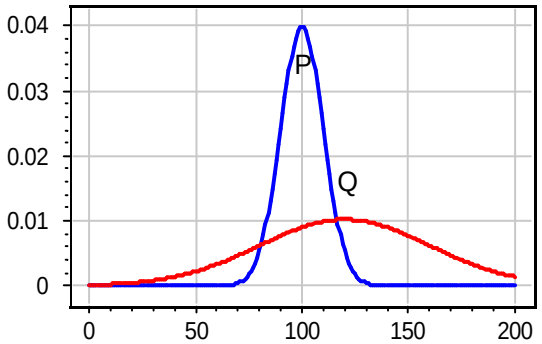
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

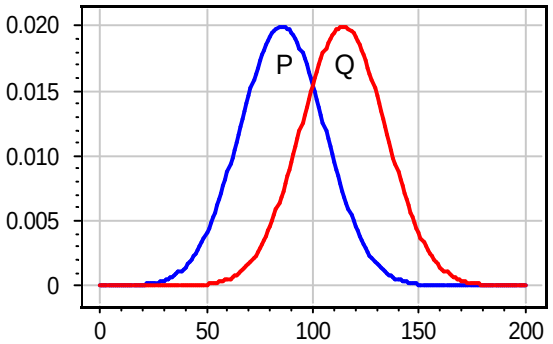
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

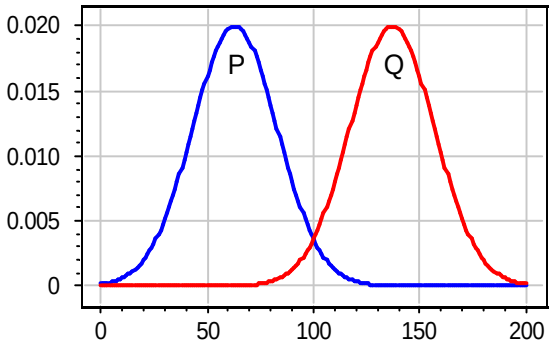
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



$\mathrm{KL}(P\|Q) = 0.442$
$\mathrm{KL}(Q\|P) = 2.966$



$\mathrm{KL}(P\|Q) = 0.444$
$\mathrm{KL}(Q\|P) = 0.444$



$\mathrm{KL}(P\|Q) = 2.969$
$\mathrm{KL}(Q\|P) = 2.969$

Регуляризатор сглаживания (почти совпадает с LDA)

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

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

Максимизируем сумму регуляризаторов:

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

Подставляем, получаем формулы M-шага LDA, для всех $t \in B$:

$$\phi_{wt} \propto n_{wt} + \beta_0\beta_w, \qquad \theta_{td} \propto n_{td} + \alpha_0\alpha_t.$$

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.

Чем сильнее разрежено распределение, тем ниже его энтропия.
Максимальной энтропией обладает равномерное распределение.

Максимизируем дивергенцию между распределениями β_w, α_t (равномерными?) и искомыми распределениями ϕ_{wt}, θ_{td}:

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

Подставляем, получаем «анти-LDA», для всех $t \in S$:

$$\phi_{wt} \propto \left(n_{wt} - \beta_0\beta_w\right)_+, \qquad \theta_{td} \propto \left(n_{td} - \alpha_0\alpha_t\right)_+.$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.

Регуляризатор декоррелирования тем

Гипотеза: в каждой теме должно быть своё лексическое ядро, отличающее 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:

$$\phi_{wt} \propto \left(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right)_+ .$$

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

Регуляризатор удаления незначимых тем

Гипотеза: если тема собрала мало слов, то она не нужна.

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

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

Подставляем, получаем:

$$\theta_{td} \propto \left(n_{td} - \tau \frac{n_d}{n_t}\theta_{td}\right)_+.$$

Строки матрицы Θ могут целиком обнуляться для тем t, собравших мало слов по коллекции, $n_t = \sum_d \sum_w n_{dwt}$.

Регуляризатор для максимизации когерентности тем

Гипотеза: тема лучше интерпретируется, если она содержит *когерентные* (часто встречающиеся рядом) слова $u, w \in W$.

Пусть C_{uw} — оценка когерентности, например $\hat{p}(w|u) = \frac{N_{uw}}{N_u}$.
Согласуем ϕ_{wt} с оценками $\hat{p}(w|t)$ по когерентным словам,

$$\hat{p}(w|t) = \sum_u p(w|u)p(u|t) = \frac{1}{n_t}\sum_u C_{uw} n_{ut};$$

$$R(\Phi,\Theta) = \tau \sum_{t\in T} n_t \sum_{w\in W} \hat{p}(w|t) \ln \phi_{wt} \to \max.$$

Подставляем, получаем ещё один вариант сглаживания:

$$\phi_{wt} \propto n_{wt} + \tau \sum_{u\in W\setminus w} C_{uw} n_{ut}.$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Empirical Methods in Natural Language Processing, EMNLP-2011. — Pp. 262–272.

Регуляризатор для учёта связей между документами

Гипотеза: чем больше n_{dc} — число ссылок из d на c, тем более близки тематики документов d и c.

Максимизируем ковариации между вектор-столбцами связанных документов θ_d, θ_c:

$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \text{cov}(\theta_d, \theta_c) \to \max,$$

Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} \propto n_{td} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc}.$$

Dietz L., Bickel S., Scheffer T. Unsupervised prediction of citation influences // ICML 2007. — Pp. 233–240.

Обзор регуляризаторов (сглаживание и разреживание)

1. разреживание предметных тем $S \subset T$:

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

2. сглаживание фоновых тем $B \subset T$:

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

3. частичное обучение по подмножествам $W_t \subset W$, $T_d \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T_d} \theta_{td} \to \max$$

Обзор регуляризаторов (корреляции и декорреляции)

4. декоррелирование тем как столбцов Φ:

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

5. максимизация когерентности тем:

$$R(\Phi) = \tau \sum_{t \in T} \sum_{u,w \in W} C_{uw} n_{ut} \ln \phi_{wt} \to \max$$

6. учёт связей между документами $n_{dd'}$:

$$R(\Theta) = \tau \sum_{d,d'} n_{dd'} \sum_{t \in T} \theta_{td} \theta_{td'} \to \max$$

7. учёт корреляций между темами как строками Θ:

$$R(\Theta) = -\frac{\tau}{2} \sum_{d \in D} (\ln \theta_d - \mu)^{\mathsf{T}} \Sigma^{-1} (\ln \theta_d - \mu) \to \max$$

Обзор регуляризаторов (определение числа тем)

8. удаление неинформативных тем:

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

9. разреживание тем во времени:

$$R(\Theta) = -\tau \sum_{y \in Y} \sum_{t \in T} \ln p(t|y) \to \max, \quad p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$$

10. сглаживание тем во времени:

$$R(\Theta) = -\tau \sum_{y \in Y} \sum_{t \in T} \left| p(y|t) - p(y-1|t) \right| \to \max$$

Обзор регуляризаторов (классификация)

11. классификация документов по классам $c \in C$, $\psi_{ct} = p(c|t)$:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct} \theta_{td} \to \max$$

12. категоризация документов по классам $c \in C$:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \sum_{t \in T} \psi_{ct} \theta_{td} \to \max$$

13. оптимизация AUC (D_c — множество документов класса c):

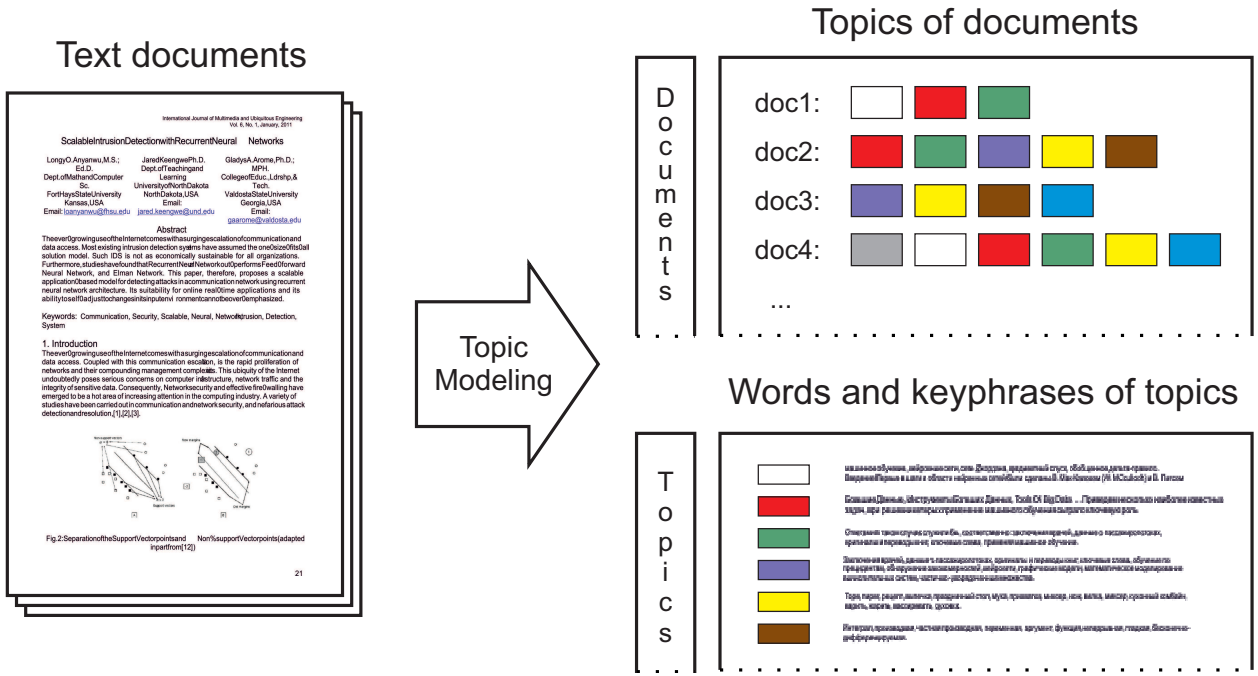
$$R(\Psi, \Theta) = -\tau \sum_{c \in C} \sum_{d \in D_c} \sum_{d' \notin D_c} \mathscr{L}\left(\sum_{t \in T} \psi_{ct} (\theta_{td} - \theta_{td'}) \right) \to \max$$

14. мультимодальная классификация документов:

$$R(\Phi, \Theta) = \sum_{j=1}^{m} \tau_j \sum_{d \in D} \sum_{x \in X_j} n_{dx} \ln \sum_{t \in T} \phi_{xt} \theta_{td} \to \max$$

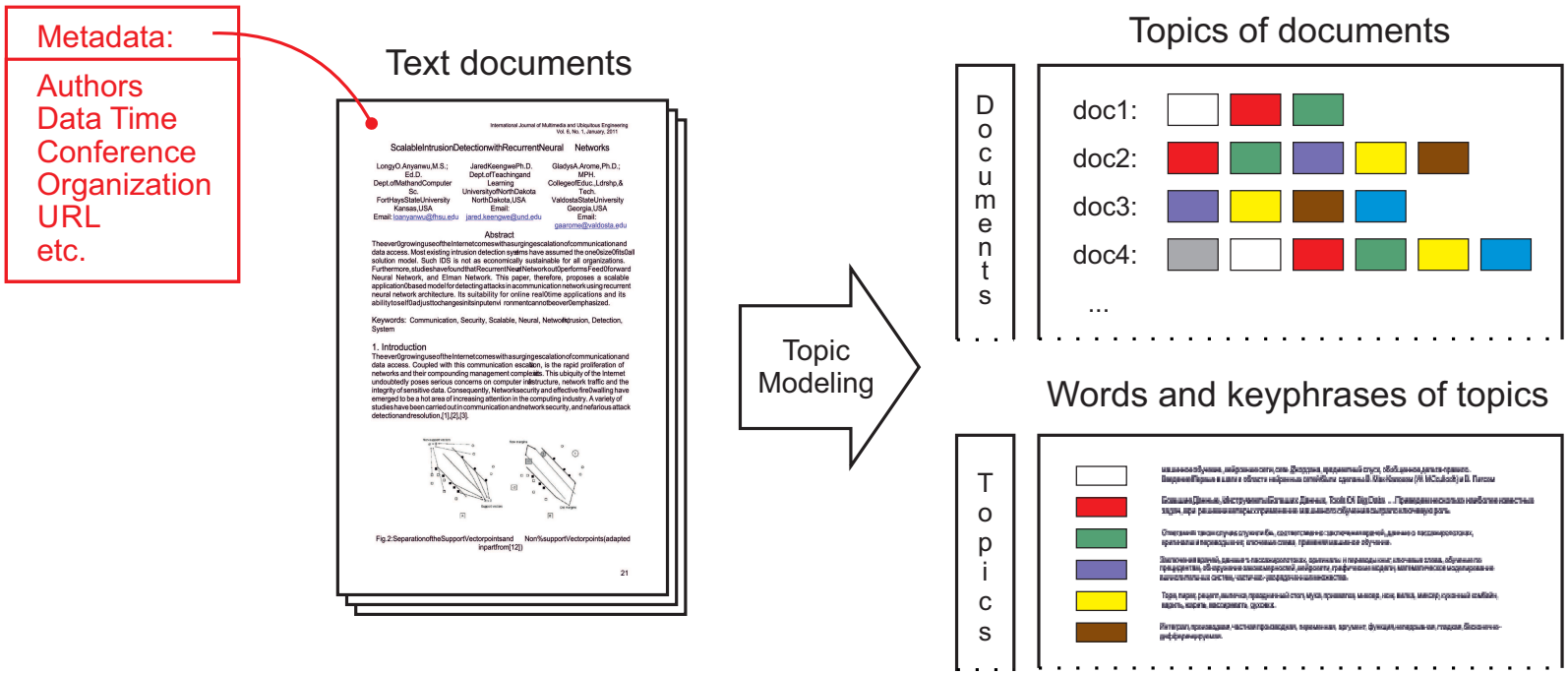
Примеры модальностей в текстах

Модальностями называются термины, но не только...



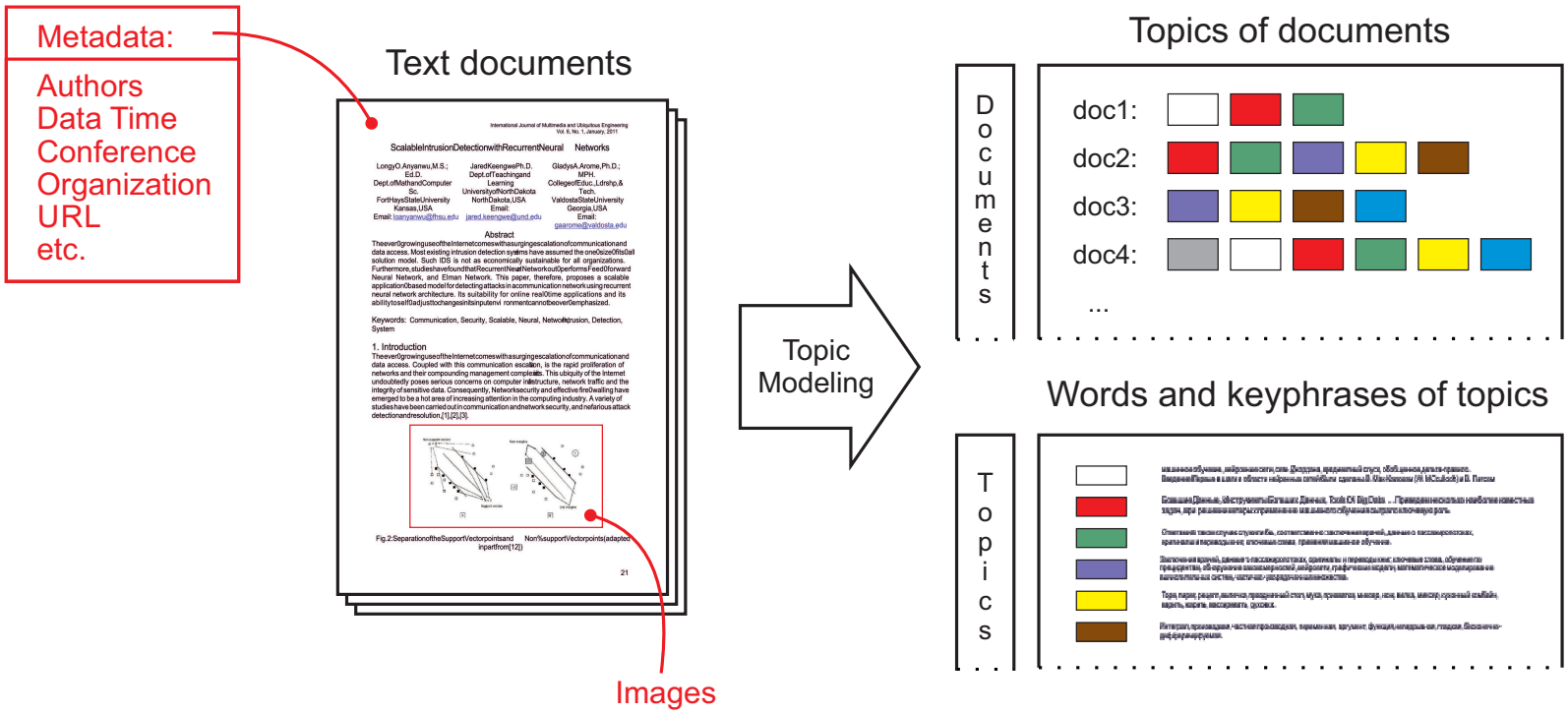
Примеры модальностей в текстах

Модальности — это термины,
метаданные (авторы, метки времени, организации, URL-ы и т. д.)



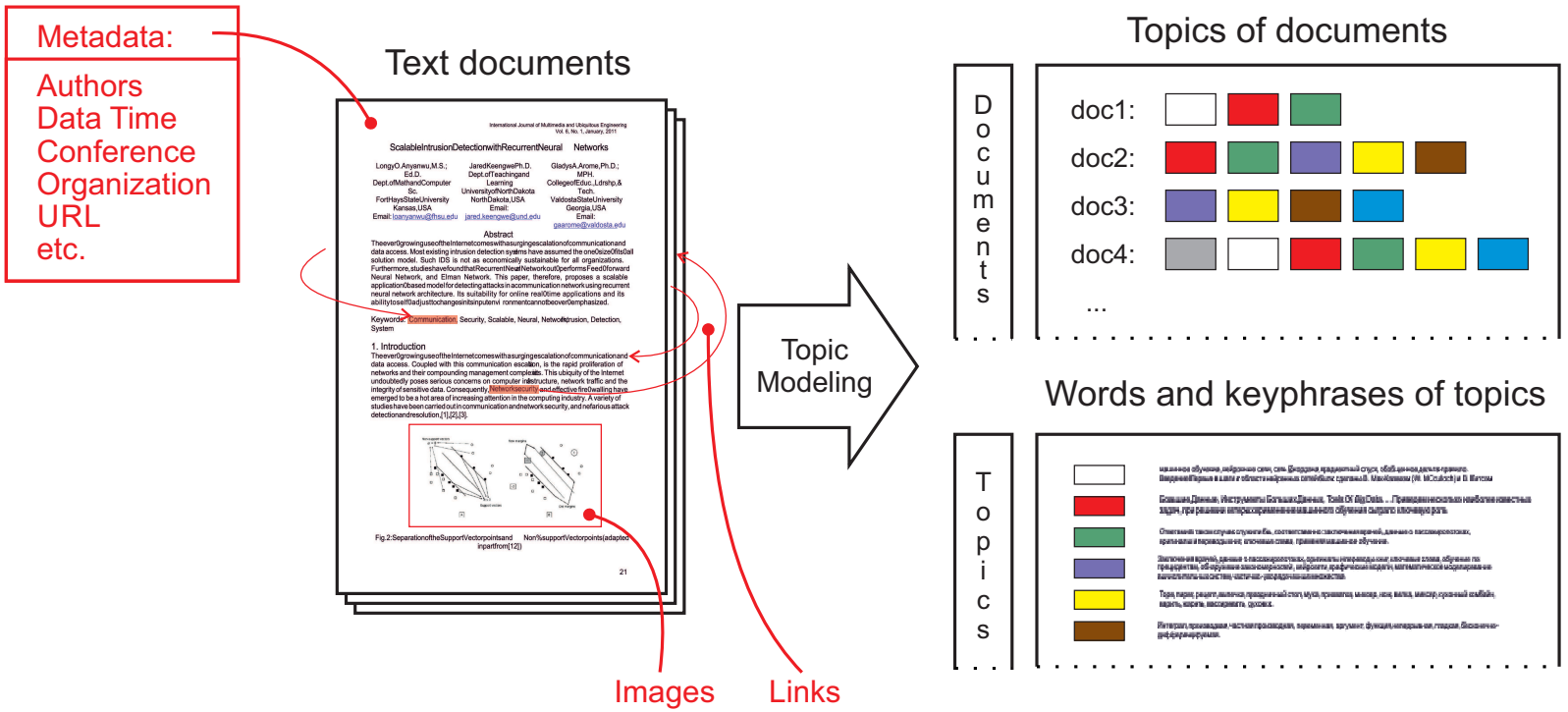
Примеры модальностей в текстах

Модальности — это термины,
метаданные (авторы, метки времени, организации, URL-ы и т. д.)
элементы изображений,...



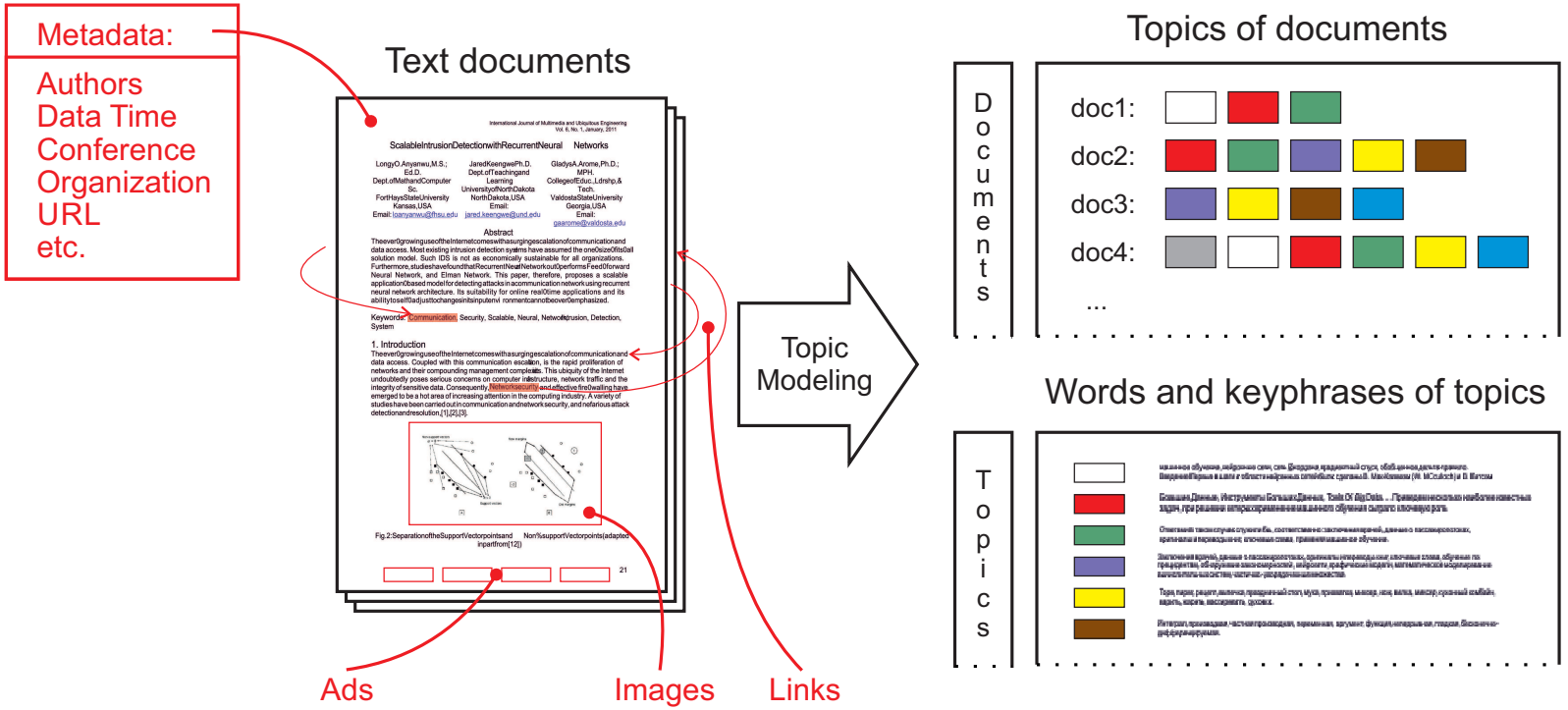
Примеры модальностей в текстах

Модальности — это термины,
метаданные (авторы, метки времени, организации, URL-ы и т. д.)
элементы изображений, ссылки,...



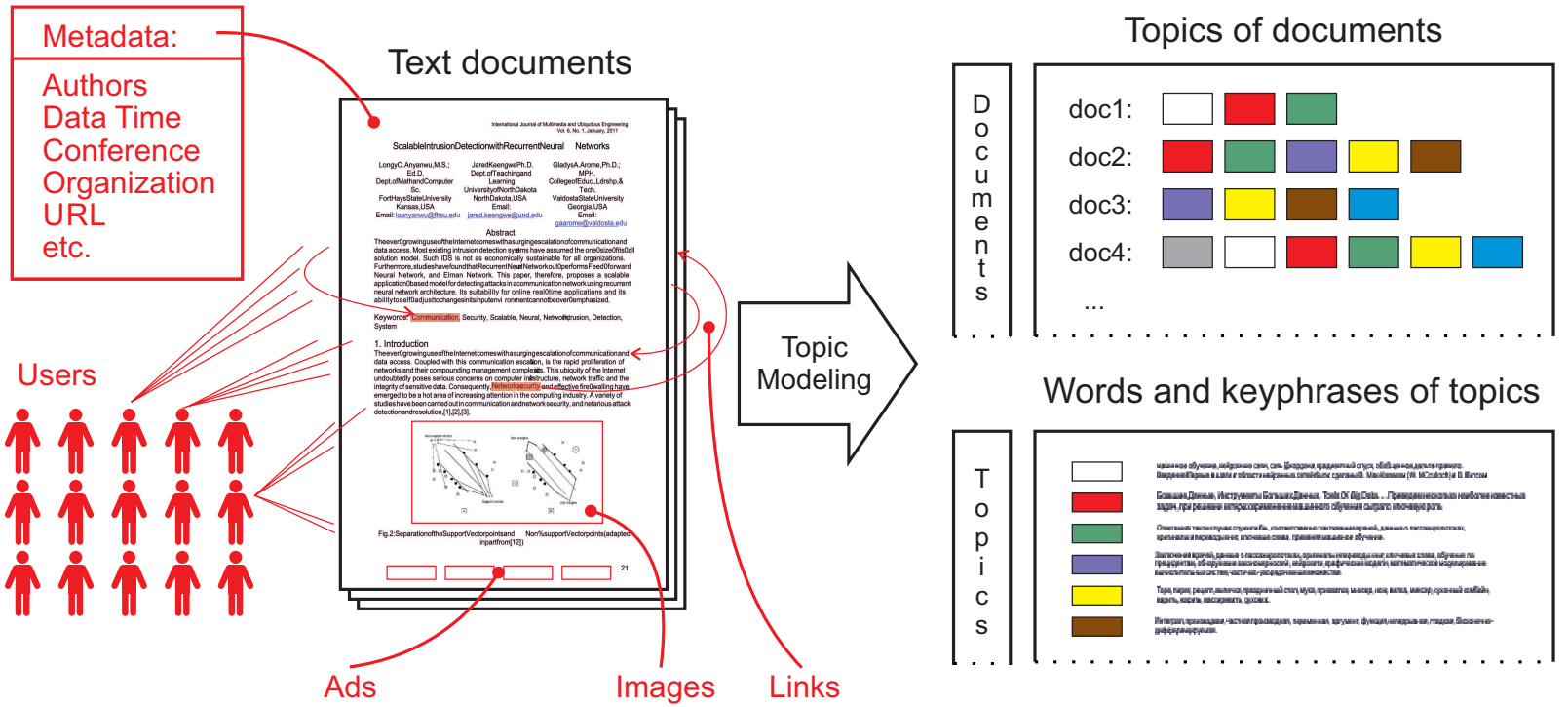
Примеры модальностей в текстах

Модальности — это термины,
метаданные (авторы, метки времени, организации, URL-ы и т. д.)
элементы изображений, ссылки, баннеры,...



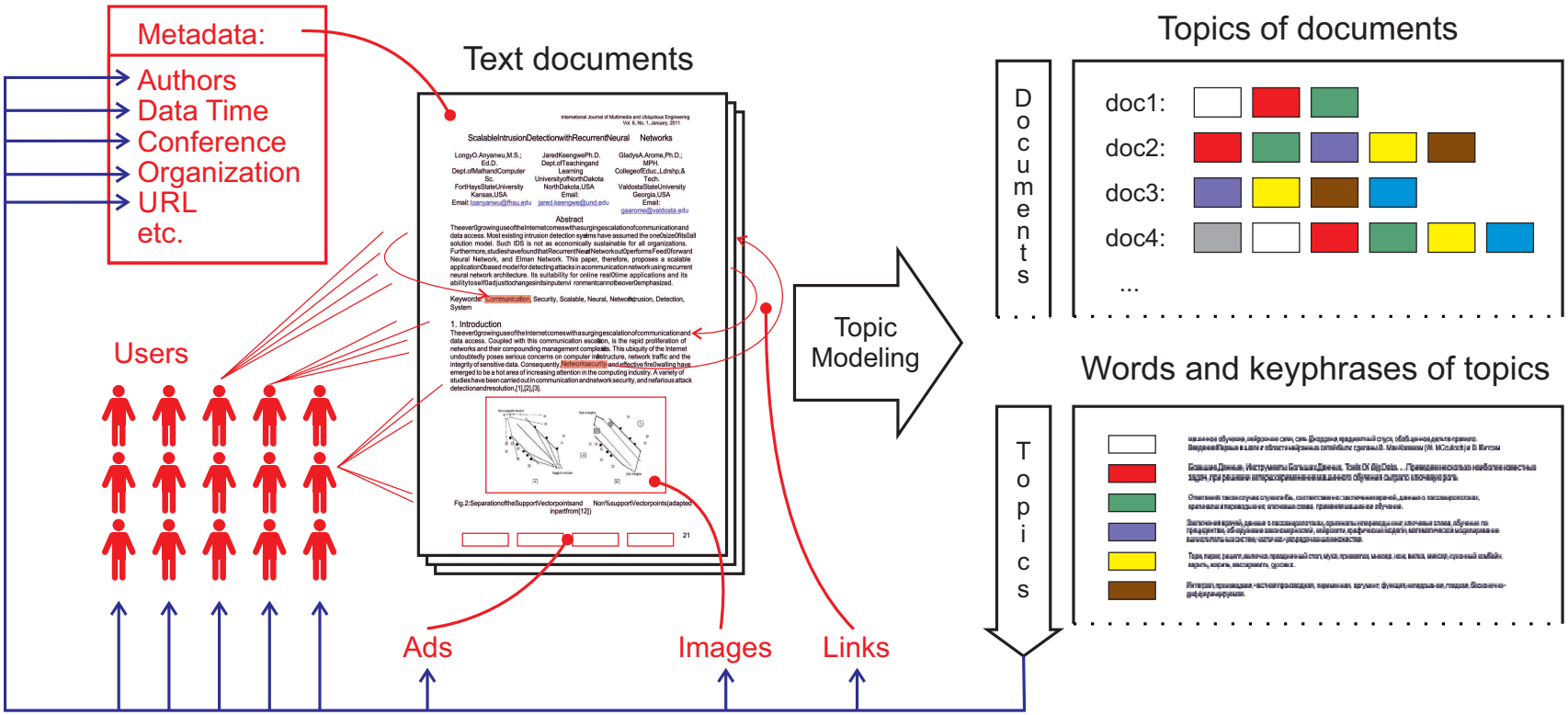
Примеры модальностей в текстах

Модальности — это термины,
метаданные (авторы, метки времени, организации, URL-ы и т. д.)
элементы изображений, ссылки, баннеры, пользователи,...



Примеры модальностей в текстах

Каждая модальность j описывается своим словарём X_j.
документы могут содержать элементы разных модальностей
каждая тема имеет своё распределение $p(x|t)$, $x \in X_j$



Ещё примеры модальностей и общее определение

- слова каждого языка (для параллельных коллекций)
- категории рубрикатора
- entity — предметы реального мира, упоминаемые в текстах: изделия, вещества, болезни, виды животных, небесные тела, фирмы, города, страны, персоны, и т. д.

Определение

Модальность — это конечное множество, элементы которого могут встречаться в документах или иным образом быть связанными с документами.

Мультимодальные тематические модели

Произвольное число модальностей X_j, $j = 1, \ldots, m$.
Вероятностное пространство $D \times T \times X$, $X = X_1 \sqcup \cdots \sqcup X_m$.
Каждый документ d состоит из токенов $x_1, \ldots, x_{n_d} \in X$.
Тематическая модель j-й модальности:

$$p(x|d) = \sum_{t \in T} p(x|t)\, p(t|d) = \sum_{t \in T} \phi_{xt}\theta_{td}, \quad x \in X_j, \quad d \in D$$

Задача максимизации взвешенного правдоподобия:

$$\mathscr{L}(\Phi, \Theta) = \sum_{j=1}^{m} \tau_j \sum_{d \in D} \sum_{x \in X_j} n_{dx} \ln \sum_{t \in T} \phi_{xt}\theta_{td} \to \max,$$

при ограничениях нормировки и неотрицательности

$$\phi_{xt} \geqslant 0; \quad \sum_{x \in X_j} \phi_{xt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t} \theta_{td} = 1$$

Модифицированный EM-алгоритм

Теорема

Точка максимума $\mathscr{L}(\Phi,\Theta)+R(\Phi,\Theta)$ удовлетворяет системе уравнений со вспомогательными переменными n_{dxt}, n_{xt}, n_{tdj}

$$\text{E-шаг:}\quad n_{dxt} = n_{dx}\frac{\phi_{xt}\theta_{td}}{\sum\limits_{s\in T}\phi_{xs}\theta_{sd}};$$

$$\text{M-шаг:}\quad \phi_{xt} \propto \left(n_{xt} + \phi_{xt}\frac{\partial R}{\partial \phi_{xt}}\right)_+; \quad n_{xt} = \sum_{d\in D} n_{dxt};$$

$$\theta_{td} \propto \left(\sum_{j=1}^{m}\tau_j n_{tdj} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+; \quad n_{tdj} = \sum_{x\in X_j} n_{dxt}.$$

Критерии качества модели

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

- Перплексия контрольной коллекции: $\mathcal{P} = \exp\left(-\frac{1}{n'} L(D')\right)$
- Разреженность — доля нулевых элементов в Φ и Θ
- Характеристики интерпретируемости тем:
  - когерентность темы [Newman, 2010]
  - размер ядра темы: $|W_t|$, ядро $W_t = \{w : p(t|w) > 0.25\}$
  - 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:
  - число тем: $|T|$
  - доля фоновых слов: $\frac{1}{n} \sum_{d \in D} \sum_{w \in d} \sum_{t \in B} p(t|d, w)$

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}.

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

N_{uv} — число документов, в которых термины 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.

Разреживание + Сглаживание + Декорреляция + Отбор тем

М-шаг при комбинировании 6 регуляризаторов:

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

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

Траектория регуляризации (*regularization path*) в пространстве $\tau = (\tau_1, \ldots, \tau_6)$ подбирается экспериментально в ходе итераций.

Воронцов К. В., Потапенко А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Межд. конф. по компьютерной лингвистике Диалог-2014.

Эксперимент на коллекции NIPS

Данные: NIPS (Neural Information Processing System)

- $|D| = 1566$ статей конференции NIPS на английском языке;
- суммарной длины $n \approx 2.3 \cdot 10^6$,
- словарь $|W| \approx 1.3 \cdot 10^4$.
- контрольная коллекция: $|D'| = 174$.

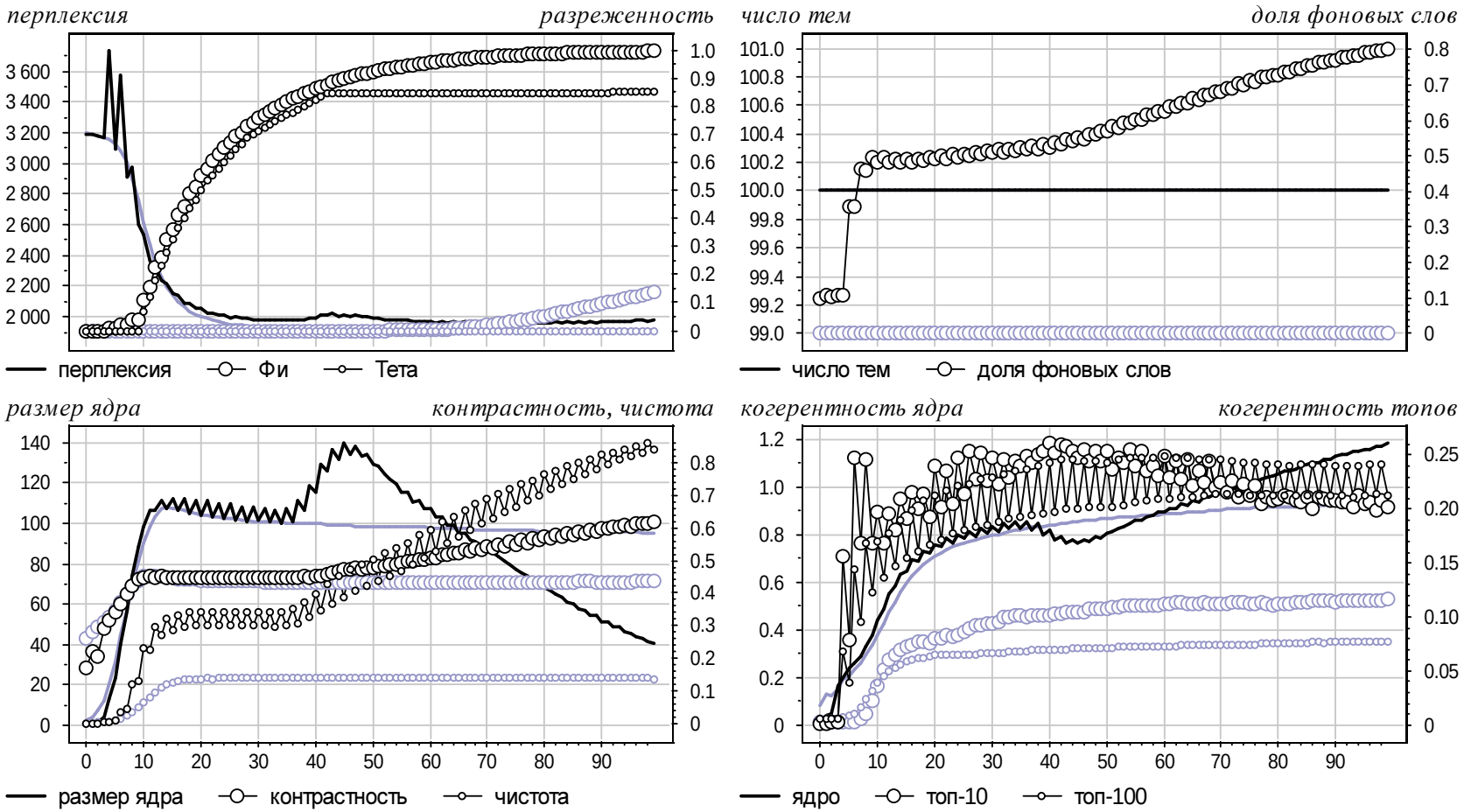
Измерение качества модели:

- перплексия контрольной коллекции: $\mathcal{P} = \exp(-\frac{1}{n'}\mathscr{L}(D'))$
- разреженность — доля нулевых элементов в Φ и Θ
- чистота, контрастность, когерентность, размер ядра тем

Воронцов К. В., Потапенко А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Межд. конф. по компьютерной лингвистике Диалог-2014.

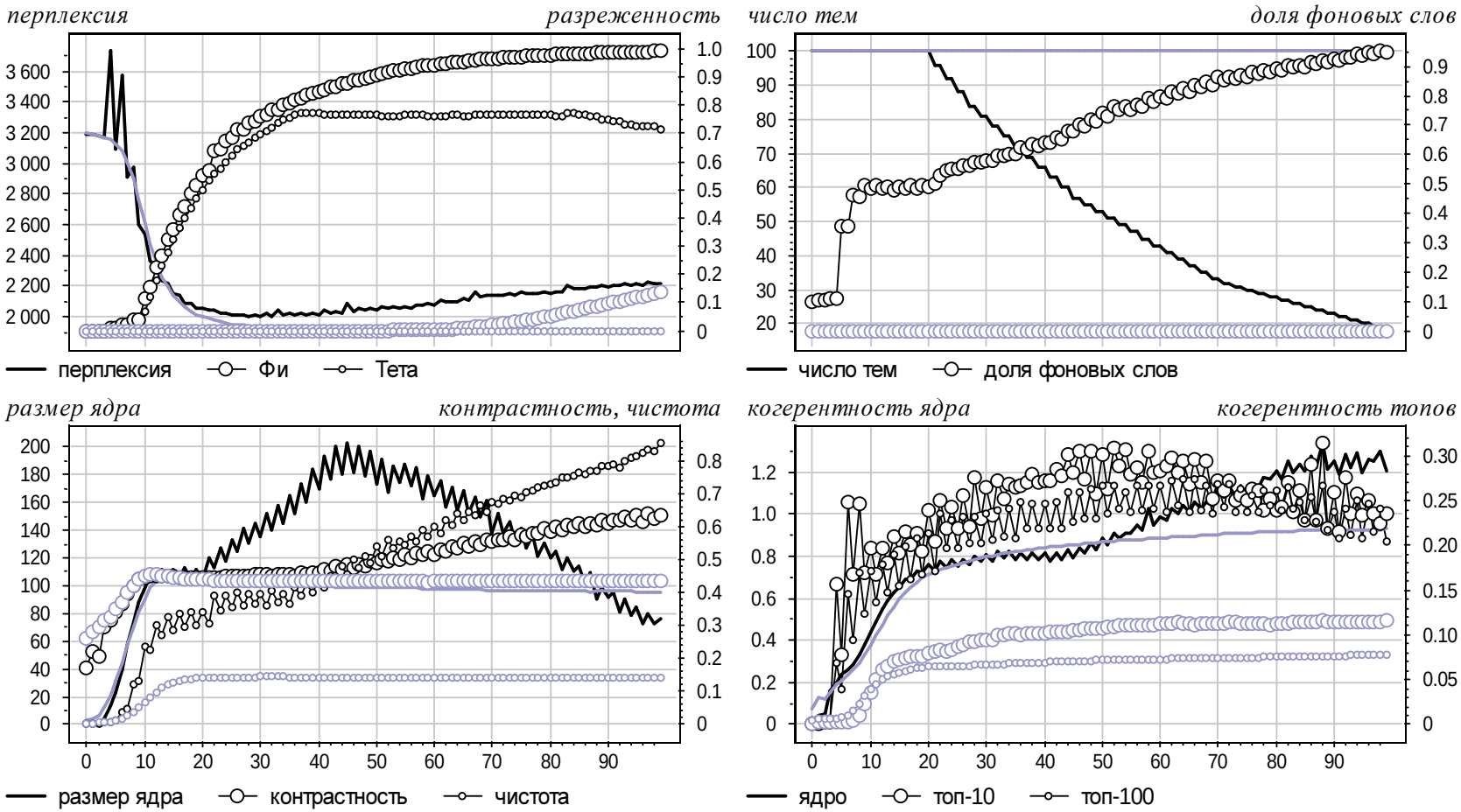
Комбинирование разреживания, сглаживания и декорреляции

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



Все те же, с удалением незначимых тем

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



Выводы по результатам экспериментов

Одновременное улучшение многих показателей:

- *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
- *размер ядер тем* вырос от 0 до 150 терминов
- почти без потери *перплексии* (правдоподобия) модели

Рекомендации по подбору 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

Резюме

- Тематическое моделирование — это восстановление латентных тем по коллекции текстовых документов.
- Задача сводится к стохастическому матричному разложению.
- Стандартные методы — PLSA и LDA.
- Задача является некорректно поставленной, так как множество её решений в общем случае бесконечно.
- Уточнение постановки задачи с помощью регуляризации приводит к многокритериальной оптимизации.
- Регуляризаторы тематических моделей разнообразны, аддитивная регуляризация позволяет их комбинировать, не сильно изменяя EM-алгоритм.

От теории к практике

Модели

- PLSA (Probabilistic Latent Semantic Analisys) — базовая идея
- LDA (Latent Dirichlet Allocation) — вероятностная регуляризация
- ARTM — адитивная регуляризация тематической модели

Инструменты

- nltk — пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования Python.
- gensim — Gensim - библиотека Python для моделирования, тематическое моделирование документов и извлечения подобия с больших корпусов.
- artm — BigARTM — Параллельная распределённая реализация методов вероятностного тематического моделирования на основе аддитивной регуляризации.

Полезные ссылки

- Описание библиотеки, пример, tutorial: http://docs.bigartm.org/en/stable/
- Пошаговое построение тематической модели https://docs.microsoft.com/ru-ru/azure/machine-learning/preview/scenario-document-collection-analysis
- Еще один пошаговый пример построения тематической модели https://github.com/piskvorky/topic_modeling_tutorial
- Сравнение и создание морфологических анализаторов в NLTK: https://habrahabr.ru/post/340404/