Математические методы анализа текстов. Тематическое моделирование (часть 1)

К. В. Воронцов, А. А. Потапенко, А. С. Попов, М. А. Апишев,
Р. Ю. Дербаносов, Н. А. Шаталов

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Математические методы анализа текстов
(курс лекций, К.В.Воронцов, А.А.Потапенко)»

31 октября 2018

Содержание

Что такое «тема» в коллекции текстовых документов?

- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся терминов
- *тема* — семантически однородный кластер текстов

Тематическая модель выявляет латентные темы
по наблюдаемым распределениям слов $p(w|d)$ в документах.

Имея коллекцию текстовых документов, хотим узнать,

- из каких тем состоит коллекция,
- из каких тем состоит каждый документ,
  $p(t|d)$ — вероятность темы t в документе d.
- из каких терминов состоит каждая тема,
  $p(w|t)$ — вероятность термина w в теме t;

Пример. Мультиязычная модель Википедии

216 175 русско-английских пар статей.
Первые 10 слов и их вероятности $p(w|t)$ в %:

Тема №68				Тема №79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример. Мультиязычная модель Википедии

216 175 русско-английских пар статей.

Первые 10 слов и их вероятности $p(w|t)$ в %:

Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

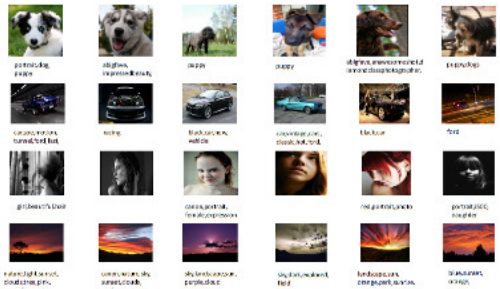
Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Приложения тематического моделирования

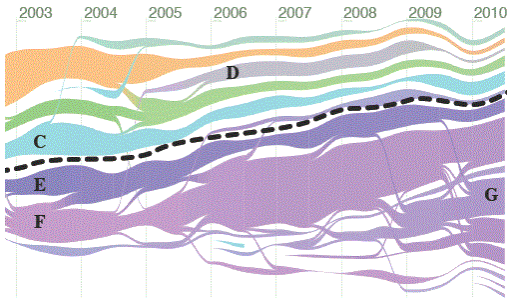
разведочный поиск в электронных библиотеках

рекомендательные с-мы и поиск в соцсетях

мультимодальный поиск текстов и изображений

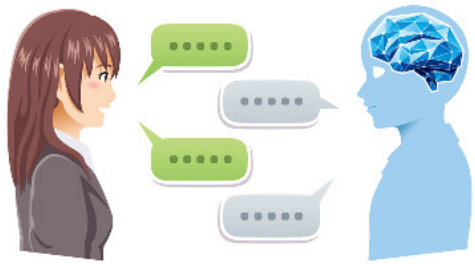


детектирование и трекинг новостных сюжетов



навигация по большим текстовым коллекциям

управлением диалогом в разговорном интеллекте



Пусть

- W — конечное множество слов (терминов, токенов)
- D — конечное множество текстовых документов
- T — конечное множество тем
- каждое слово w в документе d связано с некоторой темой t
- $D \times W \times T$ — дискретное вероятностное пространство
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- коллекция — это i.i.d. 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: $p(w|d,t) = p(w|t)$

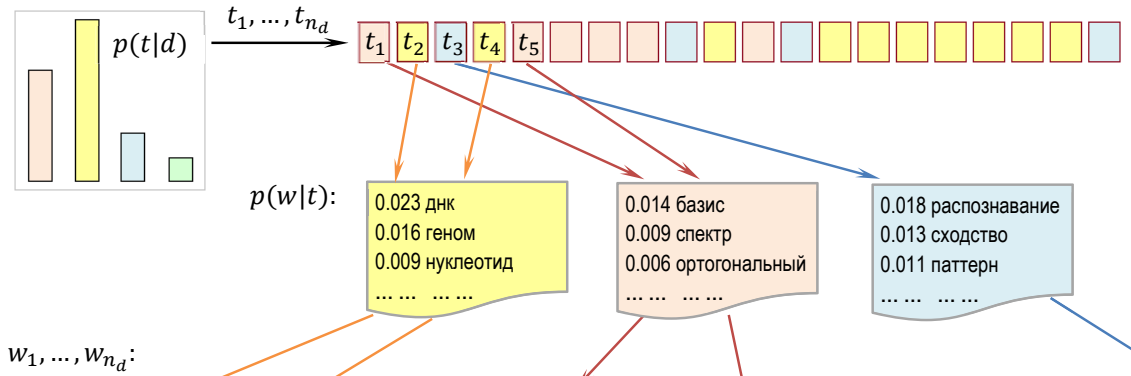
Тематическая модель, по формуле полной вероятности:

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

Прямая задача — порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D описывает появление терминов w в документах d темами t:

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



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).

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

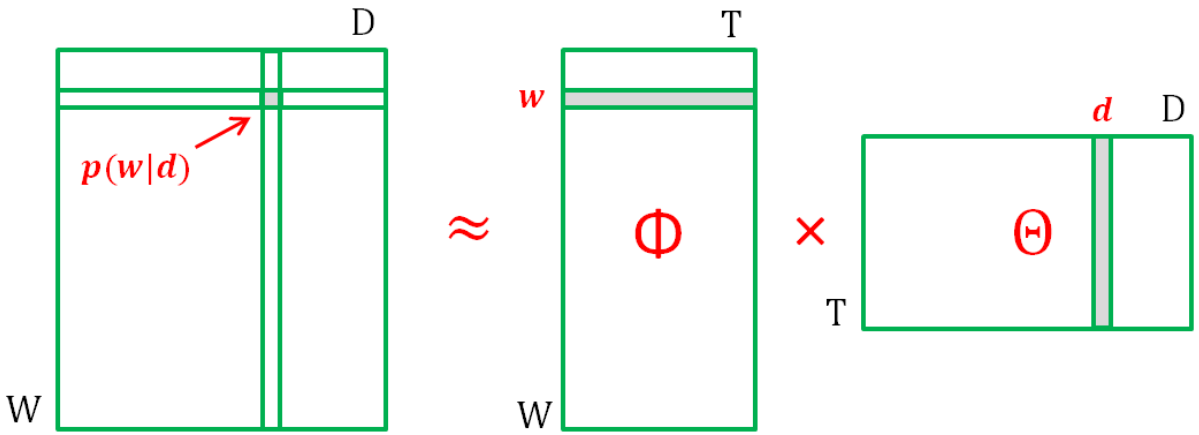
Дано: коллекция текстовых документов

- n_{dw} — частоты терминов в документах, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической 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

Это задача стохастического матричного разложения:



Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

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

Максимизация логарифма правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) \underbrace{p(d)}_{\text{const}} \to \max_{\Phi, \Theta}$$

эквивалентна максимизации функционала

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

при ограничениях неотрицательности и нормировки

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

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Задача максимизации 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 $p_{tdw} = p(t|d,w)$:

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

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.

Элементарная интерпретация EM-алгоритма

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:

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

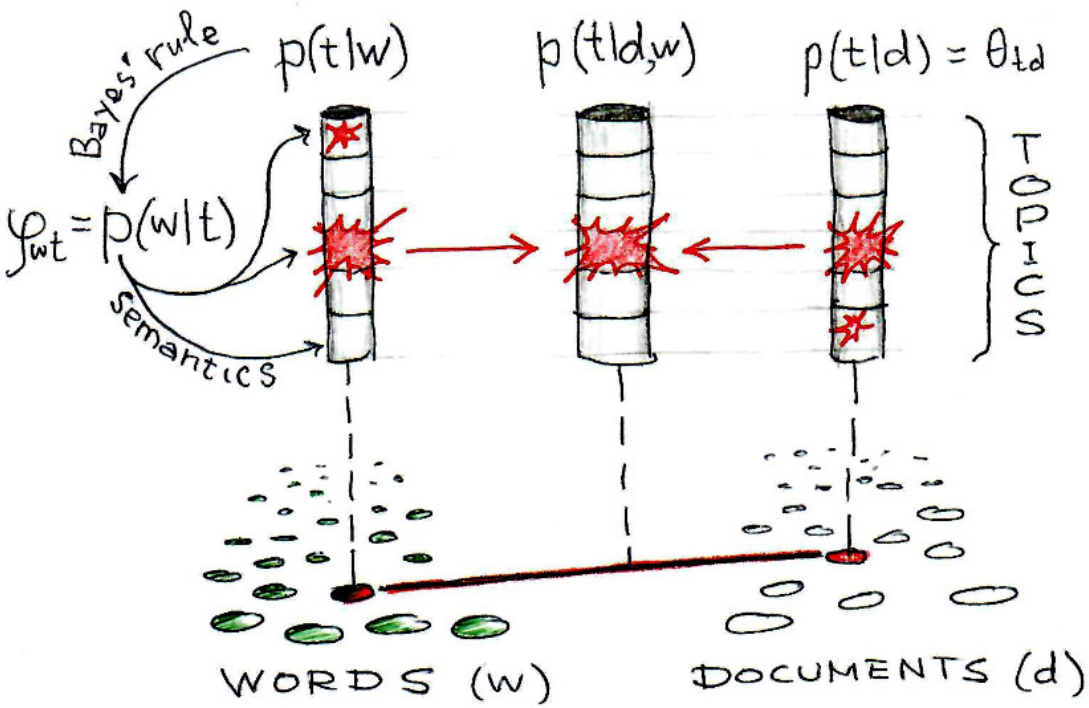
M-шаг: частотные оценки условных вероятностей вычисляются суммированием счётчика $n_{tdw} = n_{dw}p(t|d,w)$:

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

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

Интерпретируемые эмбединги слов и документов

- Коллекция текстов — двудольный граф с рёбрами (d, w)
- Слово w встречается в d, когда у них есть общие темы
- Интерпретируемость тем возникает благодаря $p(w|t)$



Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар (1865–1963)

Наша задача матричного разложения *некорректно поставлена*: если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, $\operatorname{rank} S = |T|$
- $\mathscr{L}(\Phi', \Theta') = \mathscr{L}(\Phi, \Theta)$
- $\mathscr{L}(\Phi', \Theta') \leqslant \mathscr{L}(\Phi, \Theta) + \varepsilon$ — приближённые решения

Регуляризация — стандартный приём доопределения решения с помощью дополнительных критериев.

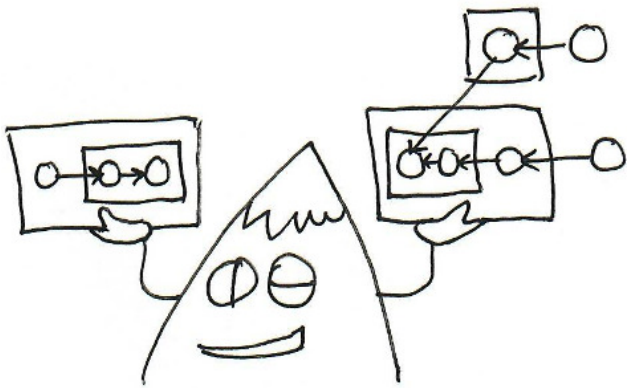
Обобщение №1: модель LDA

Проблема

Неединственность влечёт неустойчивость и переобучение.
Надо наложить ограничения на столбцы матриц Φ и Θ.
Желательно так, чтобы они стали более разреженными.

Решение

Модель латентного размещения Дирихле (2003).



Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

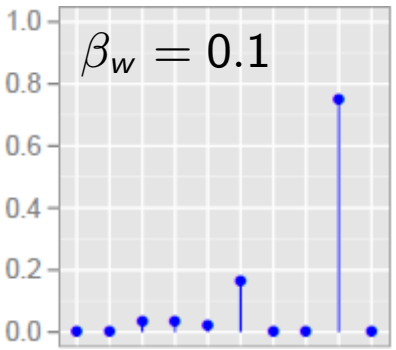
Вероятностная байесовская интерпретация LDA [Blei, 2003]

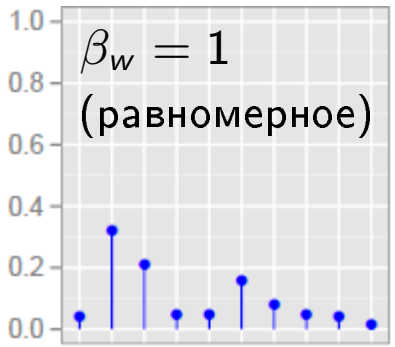
Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

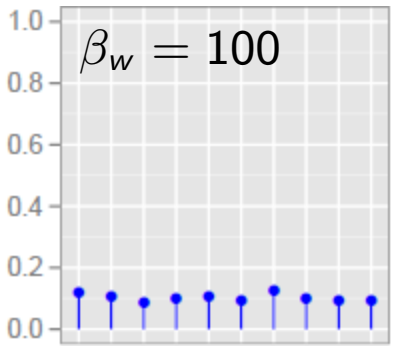
$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \ \beta_w > 0;$$

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

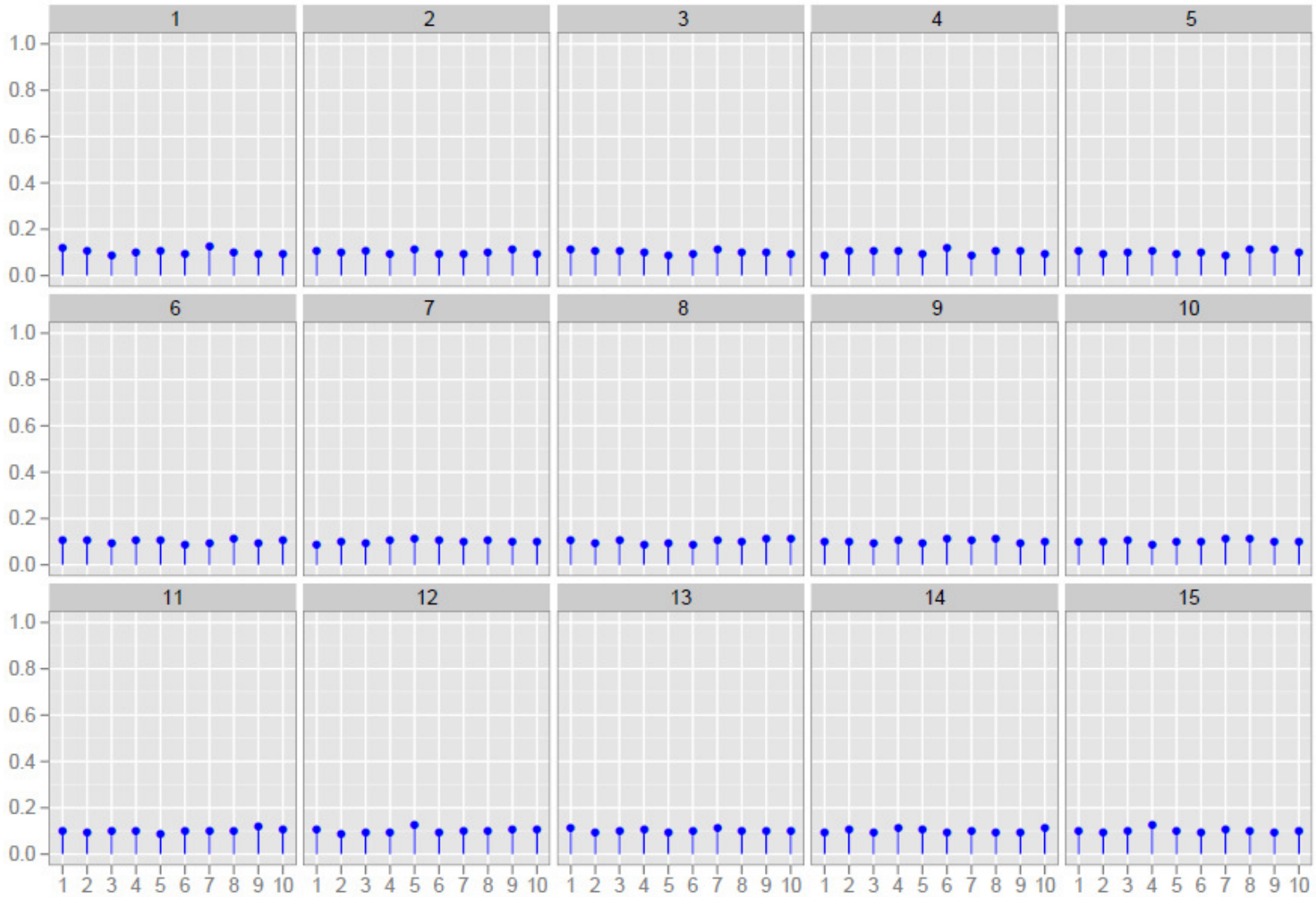
Пример. Распределение $\mathrm{Dir}(\phi|\beta)$ при $|W| = 10$, $\phi, \beta \in \mathbb{R}^{10}$:



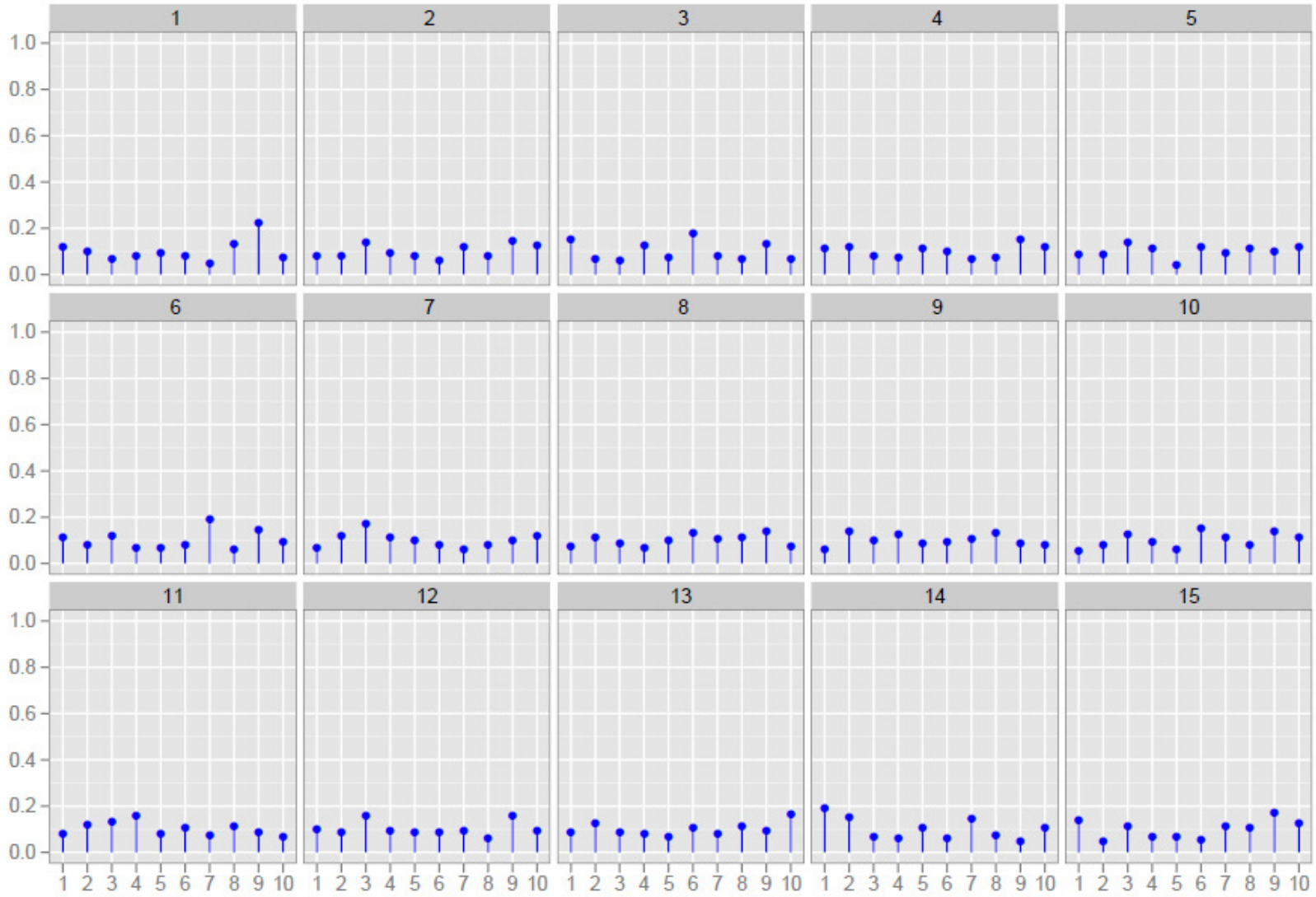




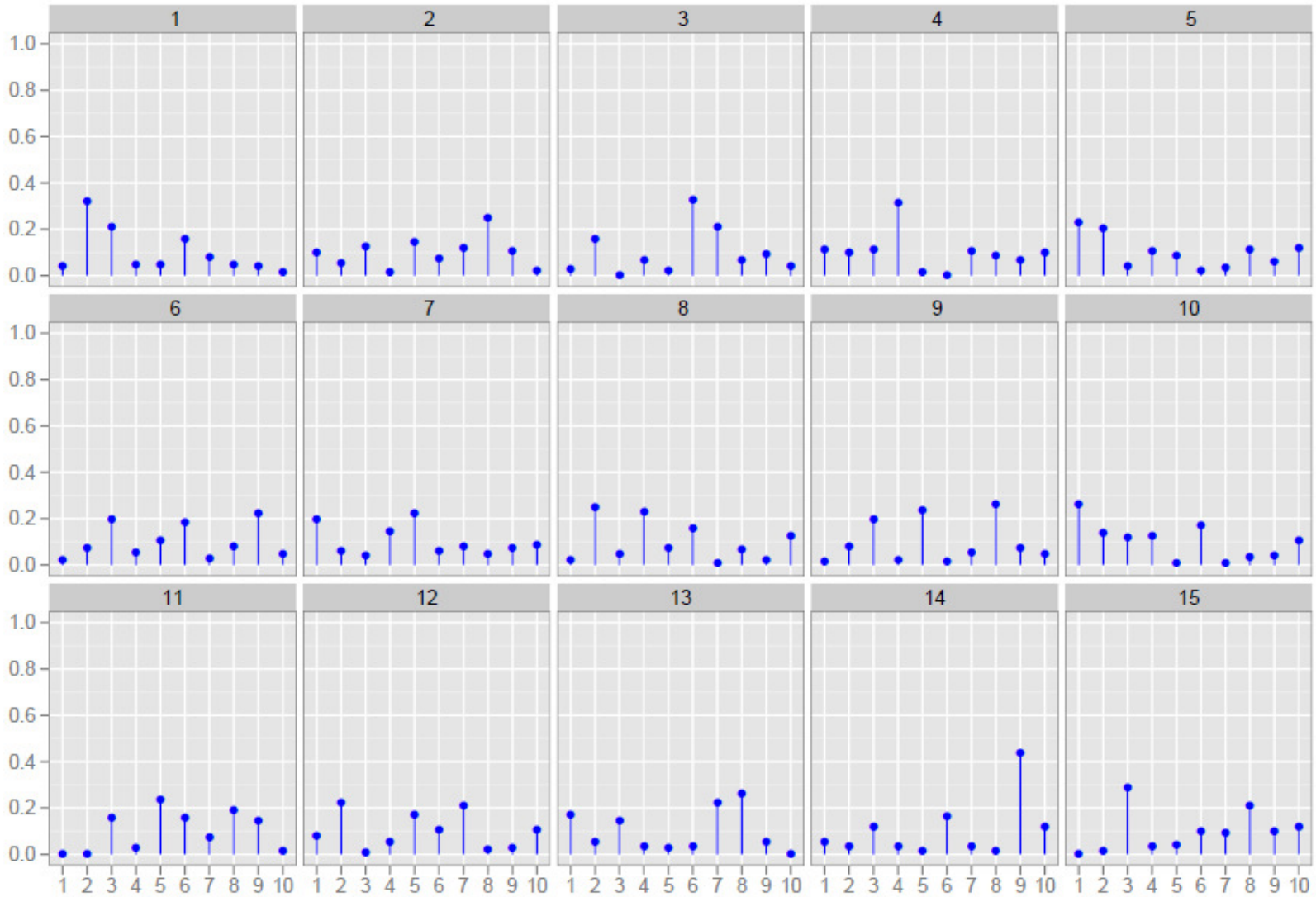
Распределение Дирихле при $\alpha_t \equiv 100$, 10 тем, 15 документов



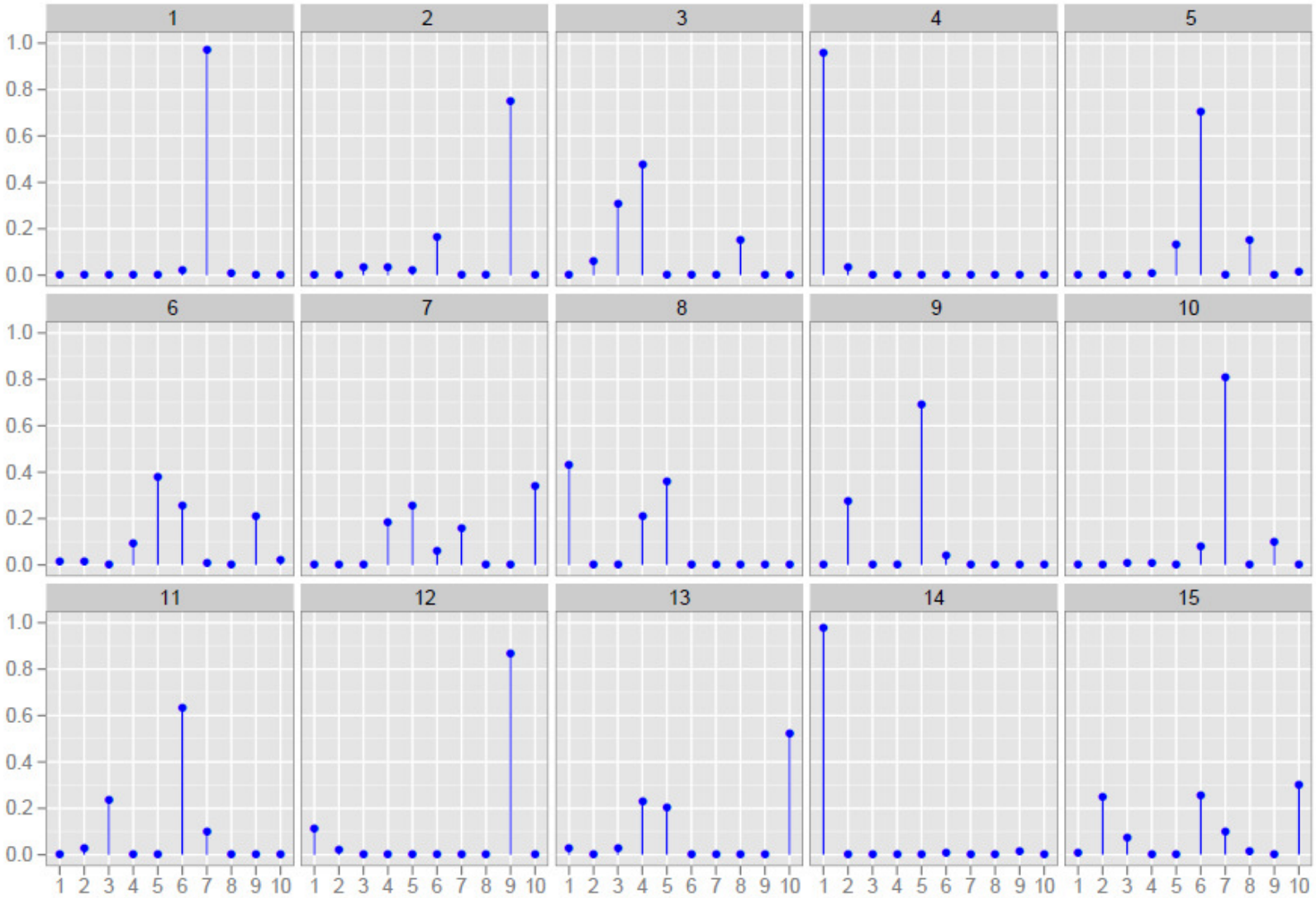
Распределение Дирихле при $\alpha_t \equiv 10$, 10 тем, 15 документов



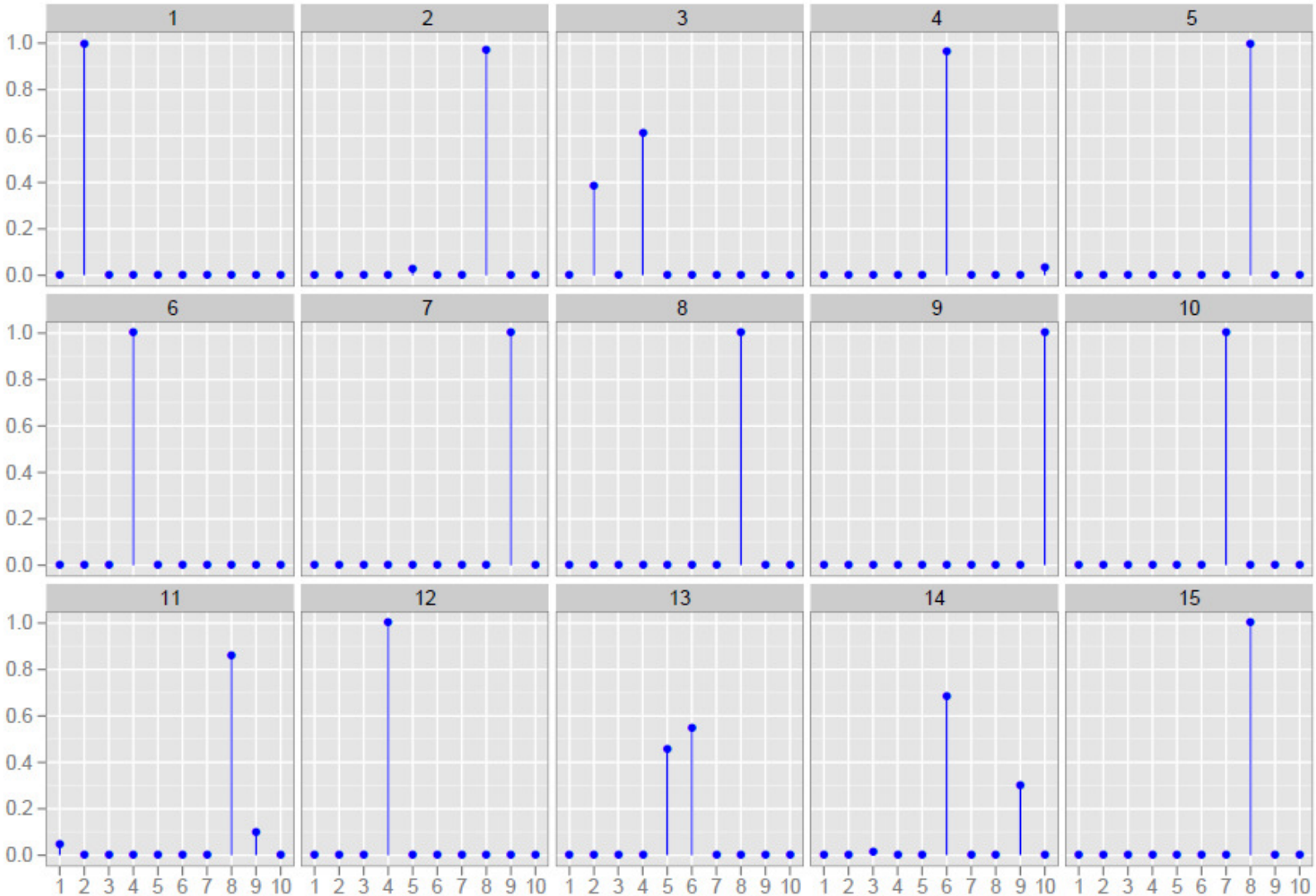
Распределение Дирихле при $\alpha_t \equiv 1$, 10 тем, 15 документов



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



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



Регуляризованный EM-алгоритм: модель LDA

Задача максимизации апостериорной вероятности:

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

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td}}_{\ln\text{ правдоподобия } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{t,w}(\beta_w - 1)\ln\phi_{wt} + \sum_{d,t}(\alpha_t - 1)\ln\theta_{td}}_{\text{критерий регуляризации } R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\mathrm{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\mathrm{norm}}\left(\sum\limits_{d\in D} n_{dw}p_{tdw} + \beta_w - 1\right) \\ & \theta_{td} = \underset{t\in T}{\mathrm{norm}}\left(\sum\limits_{w\in d} n_{dw}p_{tdw} + \alpha_t - 1\right) \end{cases}$$

Почему именно распределение Дирихле?

Плюсы:

- удобно для байесовского вывода, т. к. является сопряжённым к мультиномиальному распределению
- описывает широкий класс распределений на симплексе
- позволяет управлять разреженностью ϕ_{wt} и θ_{td}
- при малых n_{wt}, n_{td} уменьшает переобучение

Минусы:

- не имеет лингвистических обоснований
- не даёт выигрыша против PLSA на больших коллекциях
- слабый разреживатель: запрещены $\beta_w \leqslant 0$, $\alpha_t \leqslant 0$
- слабый регуляризатор: проблема неединственности остаётся

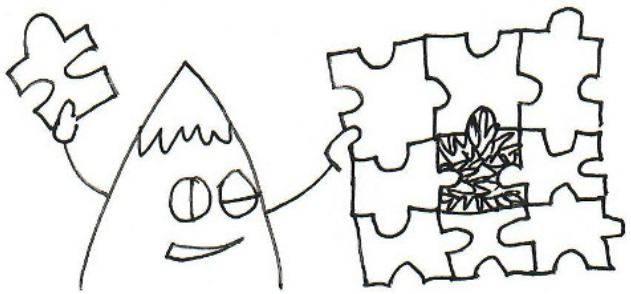
Обобщение №2: аддитивная регуляризация

Проблема

LDA — слишком простой и слабый регуляризатор.
LDA не позволяет комбинировать разные регуляризаторы.

Решение

Ввести произвольный регуляризатор $R(\Phi, \Theta)$
или сумму регуляризаторов $R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$.
Аддитивность $\rightarrow$ модульный подход к моделированию.



ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta)$$

при $\phi_{wt} \geqslant 0$, $\theta_{td} \geqslant 0$, $\sum_{w \in W} \phi_{wt} \in \{0,1\}$, $\sum_{t \in T} \theta_{td} \in \{0,1\}$.

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

Условия вырожденности модели для тем и документов

Решение может быть вырожденным для некоторых тем (столбцов матриц Φ) и документов (столбцов матрицы Θ).

Тема t вырождена, если для всех терминов $w \in W$

$$n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \leqslant 0.$$

Если тема t вырождена, то $p(w|t) = \phi_{wt} \equiv 0$; это означает, что тема исключается из модели (происходит отбор тем).

Документ d вырожден, если для всех тем $t \in T$

$$n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \leqslant 0.$$

Если документ d вырожден, то $p(t|d) = \theta_{td} \equiv 0$; это означает, что модель не в состоянии описать данный документ.

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, $j = 1, \ldots, k$:

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

Вывод системы уравнений из условий Каруша–Куна–Таккера

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

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

2. Умножим обе части равенства на ϕ_{wt} и выделим p_{tdw}:

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

3. Если $\lambda_t \leqslant 0$, то тема t вырождена, $\phi_{wt} \equiv 0$ для всех w.
4. Если $\lambda_t > 0$, то либо $\phi_{wt} = 0$, либо $n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} > 0$:

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

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

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

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

Напоминание. Дивергенция Кульбака–Лейблера

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

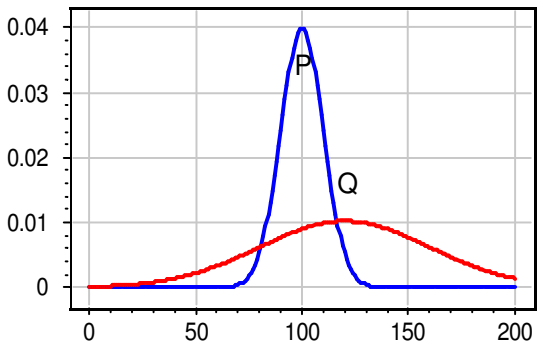
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

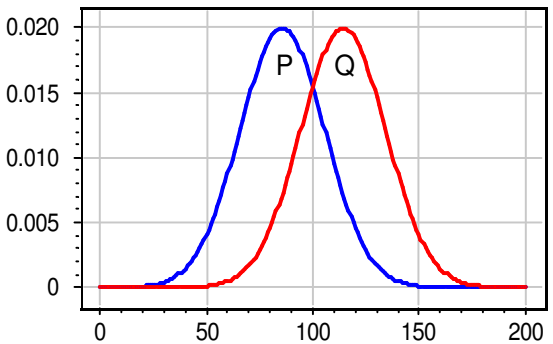
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

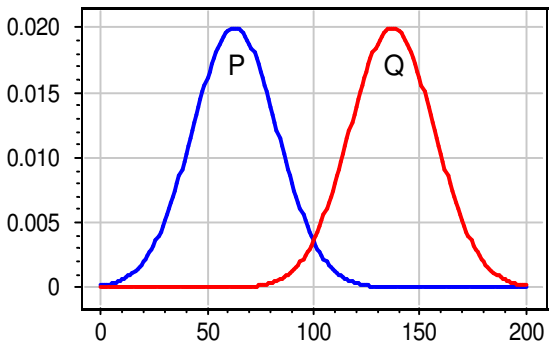
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



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



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



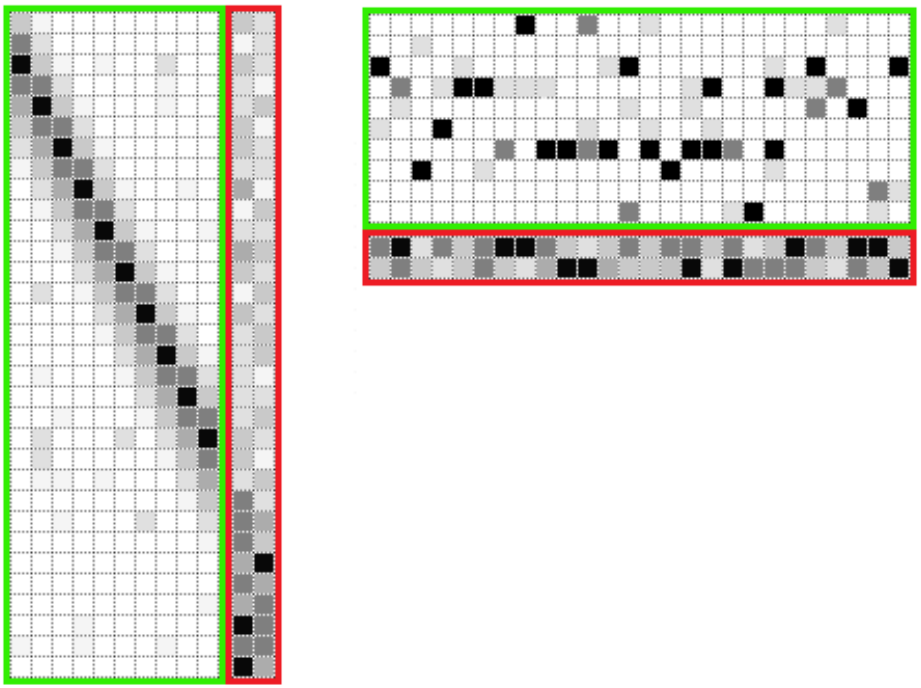
$\mathrm{KL}(P\|Q) = 2.969$
$\mathrm{KL}(Q\|P) = 2.969$

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

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



Регуляризаторы сглаживания и разреживания

Сглаживание фоновых тем $B \subset T$:

Распределения ϕ_{wt} близки к заданному распределению β_w

Распределения θ_{td} близки к заданному распределению α_t

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

где β_0, α_0 — коэффициенты регуляризации

Разреживание предметных тем $S = T \setminus B$:

Распределения ϕ_{wt} далеки от заданного распределения β_w

Распределения θ_{td} далеки от заданного распределения α_t

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

где β_0, α_0 — коэффициенты регуляризации.

Регуляризатор декоррелирования тем

Цель: усилить различность тем; выделить в каждой теме лексическое ядро, отличающее её от других тем;
вывести слова общей лексики из предметных тем в фоновые.

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:

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

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

Вероятностная порождающая модель

D — конечное множество документов
W — конечное множество терминов
T — конечное множество тем
$D \times W \times T$ — вероятностное пространство
$p(d, w, t)$ — распределение в этом пространстве

$X = (d_i, w_i)_{i=1}^n$ — *наблюдаемые переменные*
$Z = (t_i)_{i=1}^n$ — *скрытые переменные*

Вероятностная модель порождения данных:

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

$\Omega = (\Phi, \Theta)$, $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ — *параметры модели*

Задача: по X найти Ω

Принцип максимума правдоподобия

Пусть скрытые переменные Z известны: $\ln p(X, Z|\Omega) \to \max_{\Omega}$.

Тогда известны и все частоты, связанные с темами:

$$n_{dwt} = \sum_{i=1}^{n} [d_i = d]\,[w_i = w]\,[t_i = t], \quad n_{wt} = \sum_{d} n_{dwt}, \quad n_{td} = \sum_{w} n_{dwt}.$$

Воспользуемся независимостью элементов выборки (d_i, w_i, t_i):

$$p(X, Z|\Omega) = \prod_{i=1}^{n} p(d_i, w_i, t_i|\Omega) = \prod_{d,w,t} p(d, w, t|\Omega)^{n_{dwt}} =$$

$$= \prod_{d,w,t} \big(p(w|t,\Omega)\, p(t|d,\Omega)\, p(d)\big)^{n_{dwt}} = \prod_{d,w,t} \big(\phi_{wt}\theta_{td} p_d\big)^{n_{dwt}} =$$

$$= \prod_{d} p_d^{n_d} \prod_{w,t} \phi_{wt}^{n_{wt}} \prod_{d,t} \theta_{td}^{n_{td}} = C \prod_{w,t} \phi_{wt}^{n_{wt}} \prod_{d,t} \theta_{td}^{n_{td}},$$

где константа C не зависит от параметров модели.

Решение задачи максимизации правдоподобия

Максимизация логарифма правдоподобия

$$\ln p(X, Z|\Phi, \Theta) = \sum_{w,t} n_{wt} \ln \phi_{wt} + \sum_{d,t} n_{td} \ln \theta_{td} \;\to\; \max_{\Phi,\Theta}$$

при ограничениях нормировки и неотрицательности

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

Решение — частотные оценки условных вероятностей:

$$\phi_{wt} = \frac{n_{wt}}{n_t} = \underset{w \in W}{\operatorname{norm}}(n_{wt}), \qquad n_t = \sum_w n_{wt};$$

$$\theta_{td} = \frac{n_{td}}{n_d} = \underset{t \in T}{\operatorname{norm}}(n_{td}), \qquad n_d = \sum_t n_{td}.$$

Теперь перейдём к случаю, когда Z не известны.

Максимизация неполного правдоподобия

Проблема — возникает сумма под логарифмом:

$$\ln p(X|\Omega) = \ln \sum_Z p(X,Z|\Omega) \;\to\; \max_\Omega$$

Формула условной вероятности:

$$p(X,Z|\Omega) = p(Z|X,\Omega)p(X|\Omega) \;\Rightarrow\; p(X|\Omega) = \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)}$$

Для произвольного распределения $q(Z)$

$$\ln p(X|\Omega) = \sum_Z q(Z) \ln p(X|\Omega) = \sum_Z q(Z) \ln \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)} =$$

$$= \underbrace{\sum_Z q(Z) \ln p(X,Z|\Omega) - \sum_Z q(Z) \ln q(Z)}_{L(q,\Omega)\ \text{— нижняя оценка}\ \ln p(X|\Omega)} + \underbrace{\sum_Z q(Z) \ln \frac{q(Z)}{p(Z|X,\Omega)}}_{\mathrm{KL}\left(q(Z)\,\middle\|\,p(Z|X,\Omega)\right) \geqslant 0}$$

Идея EM-алгоритма. Задача E-шага

Максимизировать нижнюю оценку $L(q, \Omega)$ то по q, то по Ω:

$$\text{E-шаг:} \quad L(q, \Omega) \to \max_q$$
$$\text{M-шаг:} \quad L(q, \Omega) \to \max_\Omega$$

Задача E-шага.

Подставим $p(X, Z|\Omega) = p(Z|X, \Omega)p(X|\Omega)$ в формулу $L(q, \Omega)$:

$$\sum_Z q(Z) \ln p(Z|X, \Omega) + \underbrace{\sum_Z q(Z)}_{=1} \underbrace{\ln p(X|\Omega)}_{\text{const по } q} - \sum_Z q(Z) \ln q(Z) \to \max_q$$

$$\mathrm{KL}\big(q(Z) \,\big\|\, p(Z|X, \Omega)\big) \to \min_q$$

Утв. 1. $q(Z) = p(Z|X, \Omega)$ — точное решение задачи E-шага.

Утв. 2. $L(q, \Omega)$ — достигаемая нижняя оценка $\ln p(X|\Omega)$.

EM-алгоритм. Обоснование сходимости

Мы вывели EM-алгоритм для Z и Ω общего вида:

$$\text{E-шаг:} \quad q(Z) = p(Z|X,\Omega)$$

$$\text{M-шаг:} \quad \sum_Z q(Z) \ln p(X,Z|\Omega) \;\to\; \max_\Omega$$

и доказали его *сходимость в слабом смысле:*

- на каждом шаге правдоподобие $\ln p(X|\Omega)$ увеличивается;
- не гарантируется достижение max с заданной точностью;
- не гарантируется глобальная сходимость, так как задача в общем случае многоэкстремальная (на практике важен выбор начального приближения).

N.B. Если скрытая переменная Z не дискретна, а непрерывна, то суммирование $\sum_Z$ заменяется интегрированием $\int_Z$.

Максимизация регуляризованного правдоподобия

Пусть $p(\Omega)$ — априорное распределение параметров модели

Принцип максимума апостериорной вероятности:

$$\ln p(X,\Omega) = \ln p(X|\Omega) + \underbrace{\ln p(\Omega)}_{R(\Omega)} \;\to\; \max_{\Omega}$$

Регуляризатор $R(\Omega)$ может даже и не иметь вероятностной интерпретации, тем не менее, все выкладки остаются в силе!

$$\text{E-шаг:}\quad q(Z) = p(Z|X,\Omega)$$

$$\text{M-шаг:}\quad \sum_{Z} q(Z)\ln p(X,Z|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

Регуляризаторы используются для формализации дополнительных требований к вероятностной модели.

Регуляризованный EM-алгоритм для тематической модели

Напоминание: $\Omega = (\Phi, \Theta)$, $X = (d_i, w_i)_{i=1}^n$, $Z = (t_i)_{i=1}^n$.

E-шаг: в силу независимости элементов выборки

$$q(Z) = p(Z|X,\Omega) = \prod_{i=1}^{n} p(t_i|d_i,w_i) = \prod_{i=1}^{n} \underset{t_i}{\operatorname{norm}}\left(\phi_{w_i t_i}\theta_{t_i d_i}\right)$$

M-шаг:

$$\sum_{Z\in T^n} q(Z)\ln p(X,Z|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{(t_1,\ldots,t_n)\in T^n} \prod_{k=1}^{n} p(t_k|d_k,w_k) \sum_{i=1}^{n} \ln p(d_i,w_i,t_i|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{i=1}^{n} \sum_{t_1\in T} \cdots \sum_{t_n\in T} \prod_{k=1}^{n} p(t_k|d_k,w_k) \ln p(d_i,w_i,t_i|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{i=1}^{n} \sum_{t\in T} p(t|d_i,w_i) \ln p(d_i,w_i,t|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

Регуляризованный EM-алгоритм для тематической модели

... продолжаем вывод формулы M-шага:

$$\sum_{i=1}^{n}\sum_{t\in T} p(t|d_i,w_i)\ln p(d_i,w_i,t|\Omega) + R(\Omega) \to \max_{\Omega}$$

$$\sum_{d\in D}\sum_{w\in W}\sum_{t\in T} \underbrace{n_{dw}p(t|d,w)}_{\text{обозначим } n_{dwt}} \ln\left(\phi_{wt}\theta_{td}\right) + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

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

Чтобы применить условия KKT, выписываем лагранжиан:

$$\mathscr{L}(\Phi,\Theta) = \sum_{w,t} n_{wt}\ln\phi_{wt} - \sum_{t}\lambda_t\left(\sum_{w}\phi_{wt} - 1\right) +$$

$$+ \sum_{d,t} n_{td}\ln\theta_{td} - \sum_{d}\mu_d\left(\sum_{t}\theta_{td} - 1\right) + R(\Phi,\Theta)$$

Регуляризованный EM-алгоритм для тематической модели

Условия ККТ для стационарной точки лагранжиана:

$$\frac{\partial \mathscr{L}}{\partial \phi_{wt}} = \frac{n_{wt}}{\phi_{wt}} + \frac{\partial R}{\partial \phi_{wt}} - \lambda_t = 0 \qquad \frac{\partial \mathscr{L}}{\partial \theta_{td}} = \frac{n_{td}}{\theta_{td}} + \frac{\partial R}{\partial \theta_{td}} - \mu_d = 0$$

$$\Big(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big)_+ = \lambda_t \phi_{wt} \qquad \Big(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big)_+ = \mu_d \theta_{td}$$

$$\phi_{wt} = \underset{w \in W}{\text{norm}}\Big(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big) \qquad \theta_{td} = \underset{t \in T}{\text{norm}}\Big(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big)$$

Ещё раз вывели формулы ARTM, теперь из общего EM-алгоритма.

Частные случаи:

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

LDA: $R(\Phi, \Theta) = \ln \prod_{t \in T} \text{Dir}(\phi_t | \beta) \prod_{d \in D} \text{Dir}(\theta_d | \alpha).$

Промежуточный итог

Мы узнали более общий вариант EM-алгоритма:

- также снабжённый возможностью регуляризации,
- для которого имеется доказательство слабой сходимости,
- используемый также в методах байесовского вывода.

Байесовский вывод в тематическом моделировании:

- даёт апостериорные распределения $p(\Omega|X)$,
  хотя в BTM используются только точечные оценки Ω.
- намного более громоздкий по сравнению с ARTM,
  хотя в литературе именно он в основном и используется.
- претендует на то, чтобы оценивать меньше параметров,
  хотя на деле оценивает те же Φ и Θ, плюс гиперпараметры.

Байесовское обучение — доминирующий подход в TM

Основа подхода — байесовский вывод:

$$\text{Posterior}(\Phi, \Theta|\text{data}) \;\propto\; \text{Prior}(\Phi, \Theta)\, \text{P}(\text{data}|\Phi, \Theta)$$

В модели LDA Prior и Posterior — распределения Дирихле.

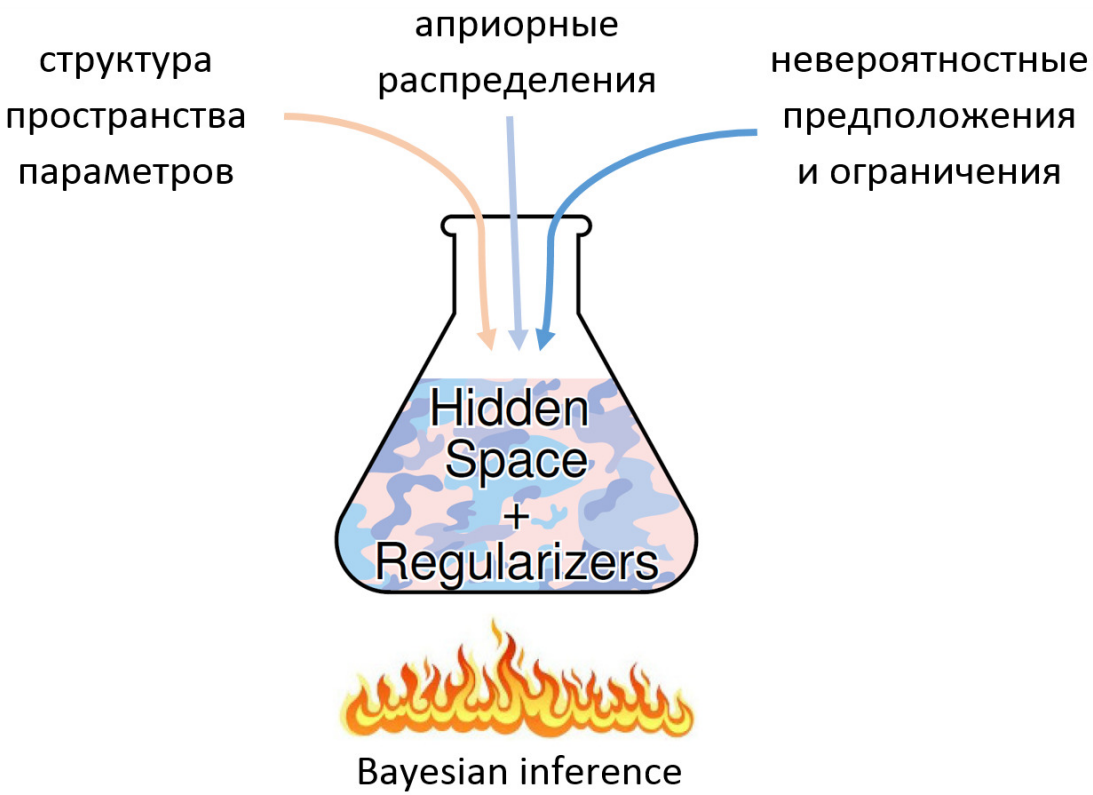
Проблемы:

- Нам нужны лишь значения Φ, Θ, а не их распределения
- Prior Дирихле имеет слабые лингвистические обоснования
- Задача сильно усложняется для несопряжённых Prior
- Байесовский вывод уникален для каждой модели
- Технически трудно обобщать и комбинировать модели

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

Байесовское обучение в тематическом моделировании

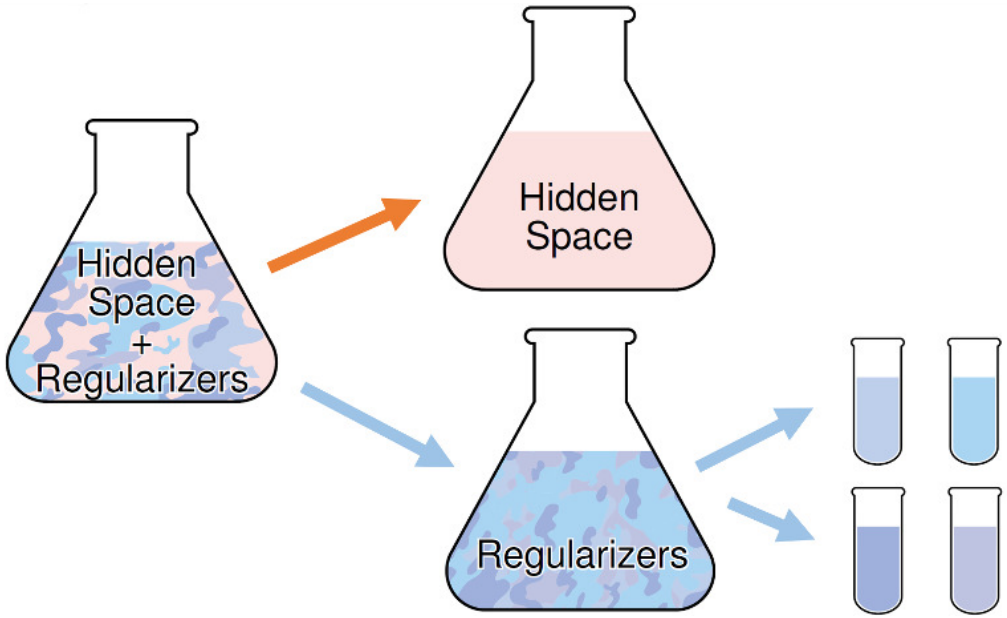
Вероятностная модель порождения данных объединяет в едином описании структуру пространства параметров, априорные распределения, дополнительные ограничения.



Не-байесовская регуляризация в тематическом моделировании

Простая *порождающая модель* описывает структуру пространства.
Регуляризаторы суммируются с весами, в любых сочетаниях,
и каждый описывает только одно дополнительное требование.

Декомпозиция — классический способ упрощения задачи



BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый параллельный мультимодальный ARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов и мер качества

Сообщество:

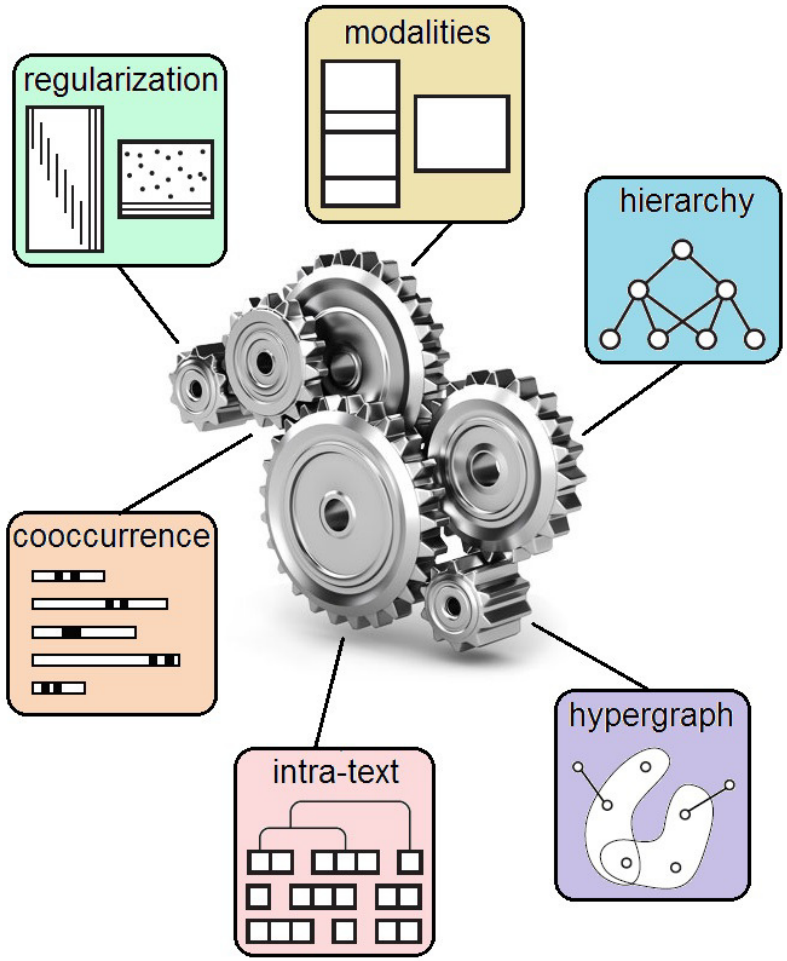
- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

Ключевые механизмы BigARTM

1. регуляризация
2. модальности
3. иерархия тем
4. совстречаемость термов
5. гиперграфы транзакций
6. внутри-текстовая регуляризация



BigARTM упрощает разработку тематических моделей

Для построения сложных моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».

Этапы моделирования	Bayesian TM	ARTM	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная порождающая модель данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Общий регуляризованный EM-алгоритм для любых моделей	
Реализация:	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)	
Оценивание:	Исследовательские метрики, исследовательский код	Стандартные метрики	Свои метрики
	Внедрение	Внедрение	

-- *нестандартизуемые этапы, уникальная разработка для каждой задачи*

-- *стандартизуемые этапы*

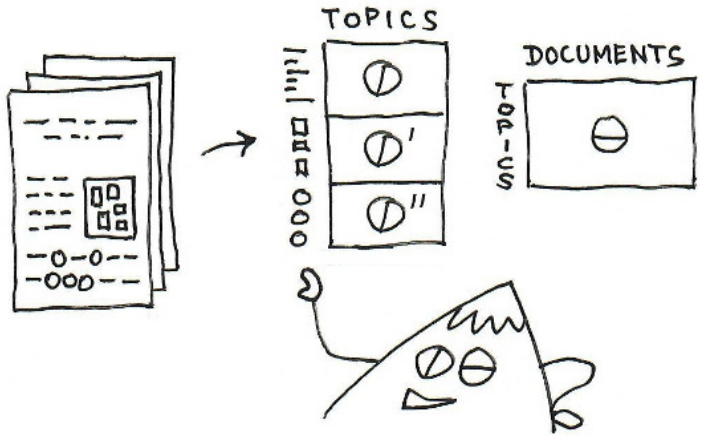
Обобщение №3: мультимодальные модели

Проблема

Есть много задач, в которых документы содержат не только слова, но и элементы других модальностей

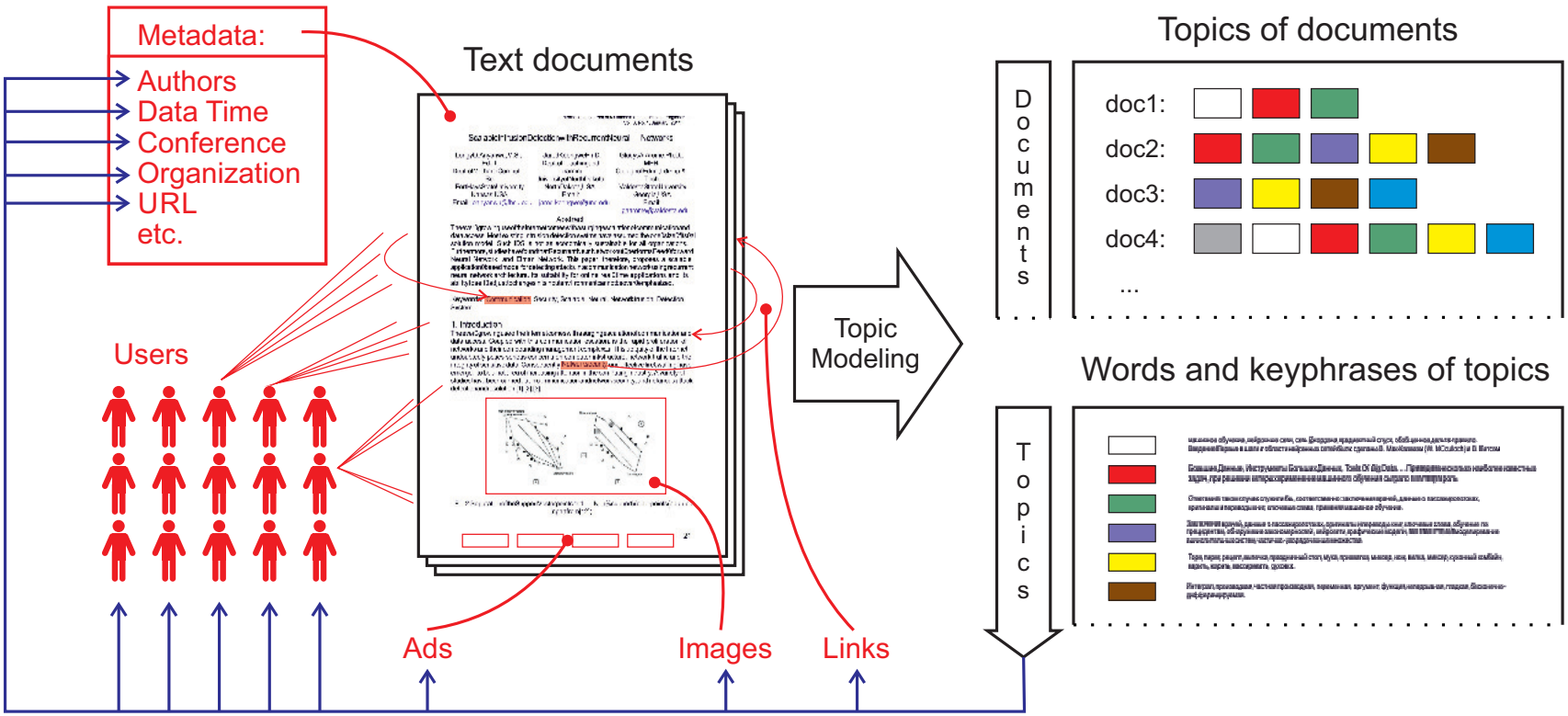
Решение

Ввести для каждой модальности свою матрицу Φ
и максимизировать свой критерий лог-правдоподобия



Задачи мультимодального тематического моделирования

Темы определяют распределения не только терминов $p(w|t)$, но и других *модальностей*: $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{ссылка}|t)$, $p(n\text{-грамма}|t)$, $p(w_{\text{язык}}|t)$, $p(\text{пользователь}|\text{t})$, $p(\text{баннер}|t), \ldots$



Мультимодальная ARTM

W^m — словарь токенов m-й модальности, $m \in M$

Максимизация суммы log правдоподобий с регуляризацией:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

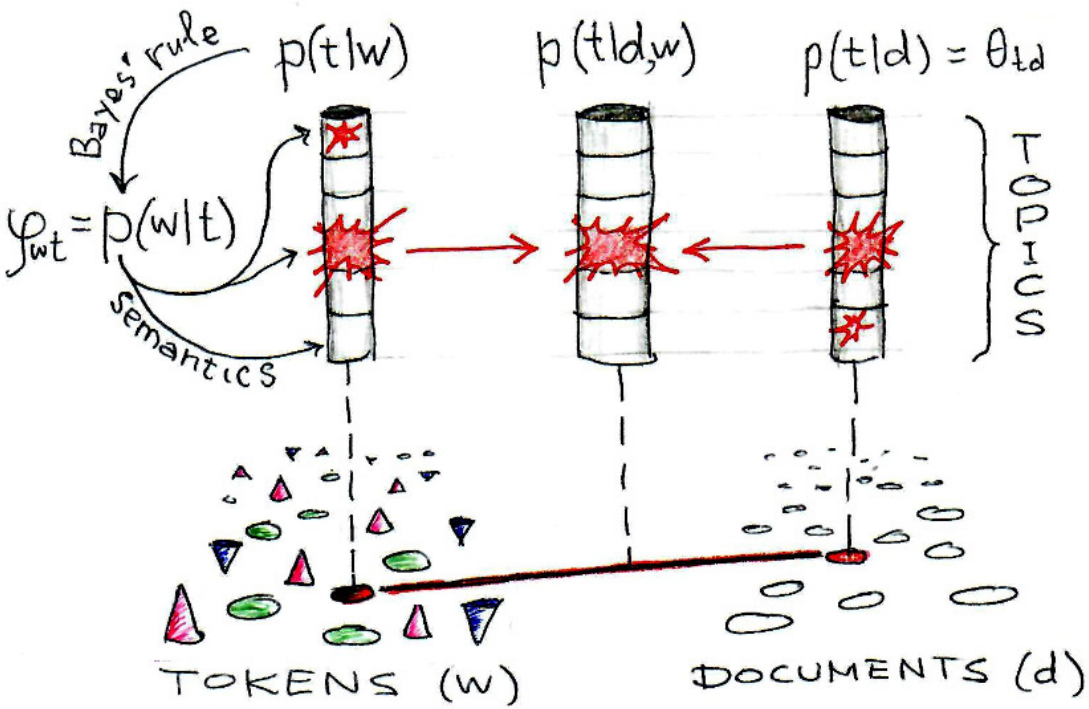
$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right)$$

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

K.Vorontsov, O.Frei, M.Apishev et al. Non-bayesian additive regularization for multimodal topic modeling of large collections. CIKM TM workshop, 2015.

Интерпретируемые эмбединги мультимодальных документов

- Документы содержат слова и токены других *модальностей*
- Примеры модальностей: авторы, время, теги, пользователи,...
- Через темы смыслы слов передаются другим модальностям



Регуляризатор для классификации и категоризации текстов

Обучающие данные: C — множество классов (категорий);
$C_d \subseteq C$ — классы, к которым d относится;
$C'_d \subseteq C$ — классы, к которым d не относится.

$p(c|d) = \sum_{t \in T} \phi_{ct}\theta_{td}$ — линейная модель классификации

Правдоподобие вероятностной модели бинарных данных:

$$R(\Phi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C_d} \ln \sum_{t \in T} \phi_{ct}\theta_{td} + \\ + \tau \sum_{d \in D} \sum_{c \in C'_d} \ln\Big(1 - \sum_{t \in T} \phi_{ct}\theta_{td}\Big) \;\to\; \max$$

При $C'_d = \varnothing$, $n_{dc} = [c \in C_d]$ это правдоподобие модальности C.

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 88 (1–2).

Регуляризатор для задач регрессии

$y_d \in \mathbb{R}$ для всех документов d — обучающие данные.

$E(y|d) = \sum_{t \in T} v_t \theta_{td}$ — линейная модель регрессии, $v \in \mathbb{R}^{|T|}$.

Регуляризатор — среднеквадратичная ошибка (МНК):

$$R(\Theta, v) = -\tau \sum_{d \in D} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right)^2 \to \max$$

Подставляем, получаем формулы М-шага:

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \left(n_{td} + \tau v_t \theta_{td} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right) \right);$$

$$v = (\Theta \Theta^{\mathsf{T}})^{-1} \Theta y.$$

Sokolov E., Bogolubsky L. Topic Models Regularization and Initialization for Regression Problems // CIKM-2015 Workshop on Topic Models. ACM, pp. 21–27.

Примеры задач регрессии на текстах

MovieReview [Pang, Lee, 2005]

d — текст отзыва на фильм

y_d — рейтинг фильма (1..5), поставленный автором отзыва

Salary (kaggle.com: *Adzuna Job Salary Prediction*)

d — описание вакансии, предлагаемой работодателем

y_d — годовая зарплата

Yelp (kaggle.com: *Yelp Recruiting Competition*)

d — отзыв (на ресторан, отель, сервис и т.п.)

y_d — число голосов «useful», которые получит отзыв

Прогнозирование скачков цен на финансовых рынках

d — текст новости

y_d — изменение цены в последующие 10–60 минут

B. Pang, L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales // ACL, 2005.

Модальность биграмм улучшает интерпретируемость тем

Коллекция 850 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Стенин С. С. Мультиграммные аддитивно регуляризованные тематические модели. Магистерская диссертация, МФТИ, 2015.

Параллельные и сравнимые корпуса текстов

Parallel — точный перевод (с выравниванием предложений), пример: EuroParl, протоколы европарламента, 21 язык.

Comparable — не перевод, а пересказ на другом языке, пример: Википедия.

W^ℓ — словарь языка ℓ из множества языков L.

Модель ML-P (MultiLingual Parallel)

- каждый язык — отдельная модальность
- $\theta_{td} = p(t|d)$ общее для всех связных документов $d = \bigsqcup\limits_{\ell \in L} d^\ell$

Дополнительные данные — двуязычные словари:

- $\Pi_k(w) \subset W^k$ — все переводы слова $w \in W^\ell$ в языке k

---

I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. 2015

Пример тем. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их вероятности $p(w|t)$ в %:

Тема №68				Тема №79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример тем. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их вероятности $p(w|t)$ в %:

Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Регуляризация по двуязычным словарям. Модель ML-TD

Гипотеза. Если $u \in \Pi_k(w)$, то тематика слов w и u близка:

$$\mathrm{KL}\big(\hat{p}(t|u) \,\big\|\, p(t|w)\big) \to \min,$$

где $\hat{p}(t|u) = \frac{n_{ut}}{n_u}$, $\quad p(t|w) = p(w|t)\frac{p(t)}{p(w)} = \phi_{wt}\frac{n_t}{n_w}$.

Модель ML-TD (MultiLingual Translation Dictionary)

$$R(\Phi) = \tau \sum_{\ell,k \in L} \sum_{w \in W^\ell} \sum_{u \in \Pi_k(w)} \sum_{t \in T} n_{ut} \ln \phi_{wt} \;\to\; \max_{\Phi}.$$

Недостатки. Модель ML-TD не учитывает два обстоятельства:

- тематику омонимов сближать не нужно,
- слово может иметь разные переводы в разных темах.

Дударенко М. А. Регуляризация многоязычных тематических моделей // Вычислительные методы и программирование. 2015. Т. 16. С. 26–36.

Матрица вероятностей переводов. Модель ML-TDP

Гипотеза. Переводы слов зависят от тем: $\pi_{uwt}^{k\ell} = p(u|w,t)$, темы согласуются в разных языках через переводы слов:

$$\mathrm{KL}\big(\hat{p}(u|t) \,\big\|\, p(u|t)\big) \to \min;$$

$\hat{p}(u|t) = \frac{n_{ut}}{n_t}$ — частотная оценка по модальности (языку) k,
$p(u|t)$ — модель темы t в языке k по языку ℓ:

$$p(u|t) = \sum_{w \in \Pi_\ell(u)} p(u|w,t)p(w|t) = \sum_{w \in \Pi_\ell(u)} \pi_{uwt}^{k\ell}\phi_{wt}.$$

Модель ML-TDP (MultiLingual Translation Dictionary Probability)

$$R(\Phi,\Pi) = \tau \sum_{\ell,k \in L} \sum_{u \in W^k} \sum_{t \in T} n_{ut} \ln \sum_{w \in \Pi_\ell(u)} \pi_{uwt}^{k\ell}\phi_{wt} \to \max_{\Phi,\Pi}.$$

Дударенко М. А. Регуляризация многоязычных тематических моделей // Вычислительные методы и программирование. 2015. Т. 16. С. 26–36.

Формулы M-шага для моделей ML-TD и ML-TDP

ML-TD (MultiLingual Translation Dictionary):

$$\phi_{wt} = \underset{w \in W^\ell}{\text{norm}} \left(n_{wt} + \tau \sum_{k \in L \setminus \ell} \sum_{u \in \Pi_k(w)} n_{ut} \right)$$

ML-TDP (MultiLingual Translation Dictionary Probability):

$$\phi_{wt} = \underset{w \in W^\ell}{\text{norm}} \left(n_{wt} + \tau \sum_{k \in L \setminus \ell} \sum_{u \in \Pi_k(w)} \pi^{\ell k}_{wut} n_{ut} \right)$$

$$\pi^{k\ell}_{uwt} = \underset{u \in W^k}{\text{norm}} \left(\pi^{\ell k}_{wut} n_{ut} \right)$$

Смысл регуляризации:

условные вероятности $\phi_{wt} = p(w|t)$ согласуются с их частотными оценками по словам других языков

Тематические переводы слов $\pi^{k\ell}_{uwt} = p(u|w,t)$

Темы, в которых $p(\text{«sum»}|\text{«сумма»},t) > 0.9$

Тема №6		Тема №12		Тема №20	
множество	set	математика	triangle	вектор	vector
пространство	space	треугольник	square	координата	coordinate
группа	point	теорема	number	пространство	field
точка	left	точка	point	преобразование	tensor
элемент	limit	математический	theorem	базис	transform
функция	symmetry	угол	angle	тензор	basis
предел	function	координата	mathematics	сила	space
отображение	open	экономика	real	векторный	force
симметрия	property	число	theory	точка	rotation
открытый	topology	квадрат	geometry	система	thermometer

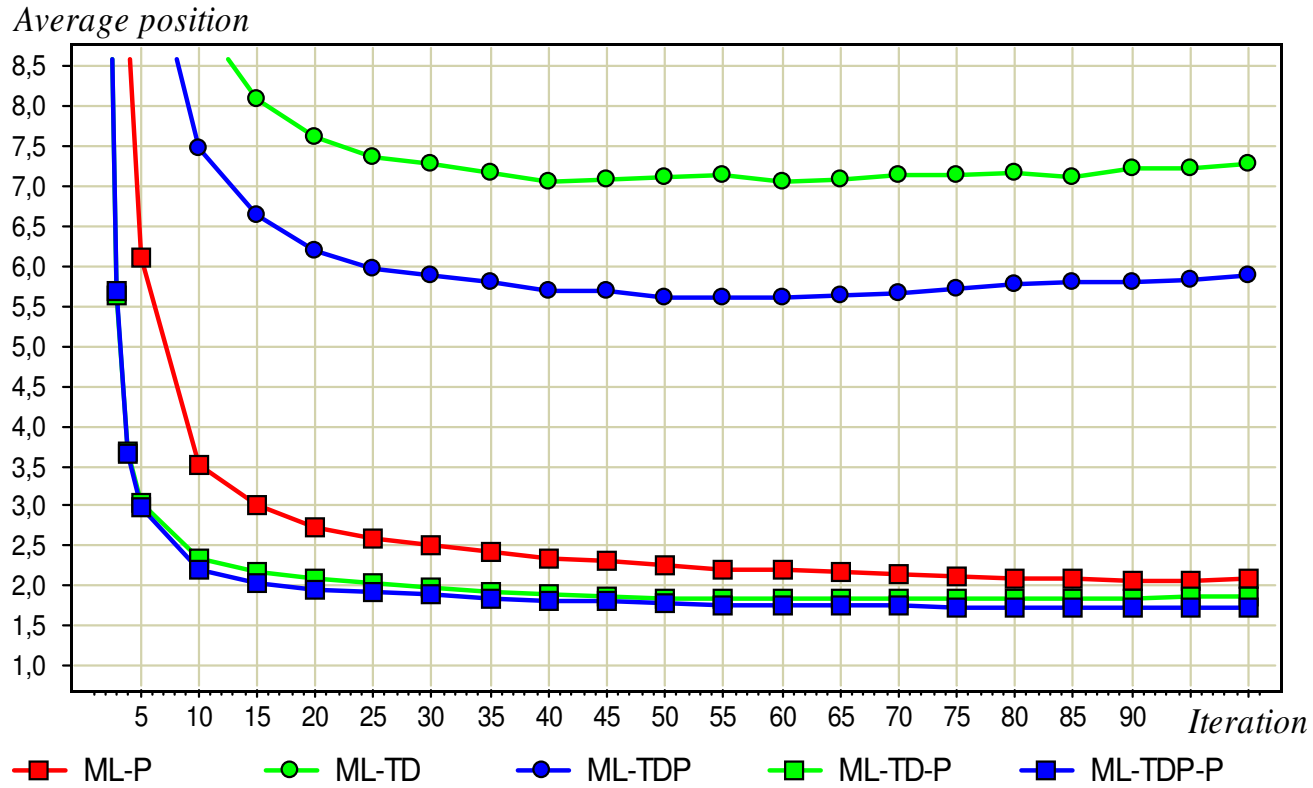
Темы, в которых $p(\text{«total»}|\text{«сумма»},t) > 0.9$

Тема №5		Тема №19		Тема №22	
орбита	space	программный	software	игра	game
аппарат	nasum	версия	version	видеосигнал	character
космический	orbit	работа	news	игрок	video
земля	instrument	компания	company	фильм	player
поверхность	earth	анонимный	work	головоломка	series
солнечный	surface	примечание	note	серия	puzzle
станция	solar	терминатор	release	качество	movie
запуск	system	журнал	support	шахматы	jason
система	landing	рей	terminator	джейсон	world
атмосфера	camera	персонаж	anonymous	буква	chess

Кросс-язычный поиск: ищем документ по его переводу

Wiki: $|D| = 586$, категория «Математика», $|T| = 100$,
$|W^{\text{рус}}| = 19\,305$, $|W^{\text{eng}}| = 23\,413$, переводов 82 642 пар.

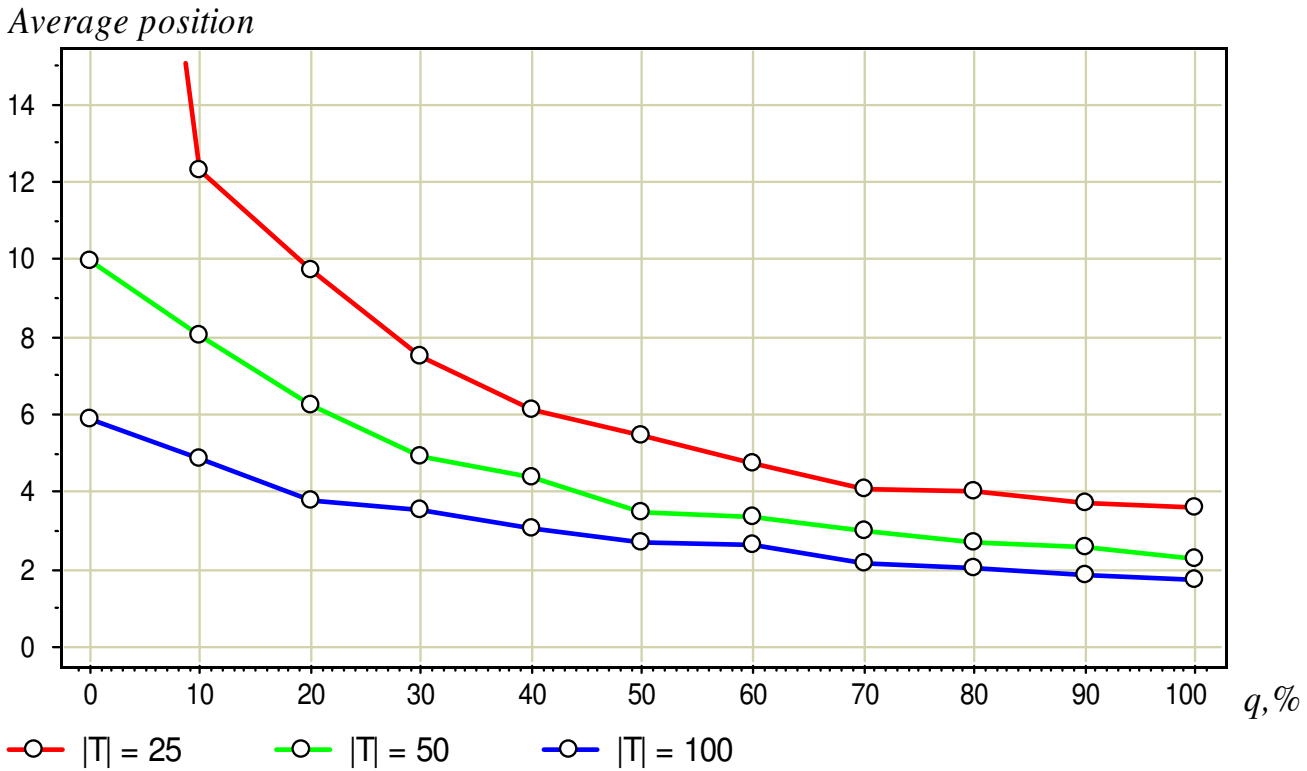
Качество поиска — средняя позиция перевода в выдаче:



Кросс-язычный поиск: ищем документ по его переводу

Wiki: $|D| = 586$, категория «Математика», $|T| = 25, 50, 100$, $|W^{\text{рус}}| = 19\,305$, $|W^{\text{eng}}| = 23\,413$, переводов 82 642 пар.

Зависимость средней позиции перевода в поисковой выдаче от числа тем $|T|$ и доли q параллельных текстов в коллекции:



Резюме по мультиязычным моделям

- Главное чудо: для построения мультиязычных тем достаточно иметь сравнимые корпуса.
- Сравнимая коллекция является более сильным источником многоязычной информации, чем словарь переводов (!)
- Модель с вероятностями переводов — самая сильная
- Не обязательно, чтобы все документы имели параллельные
- Главное применение — по запросу на одном языке ищем:
  — тексты на другом языке — *кросс-язычный поиск*,
  — тексты на всех языках — *мульти-язычный поиск*.
- Применение в статистическом машинном переводе:
  выбор варианта перевода согласно тематике документа.