Математические методы анализа текстов. Тематическое моделирование

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, 7 ноября 2018

Содержание

1. **Теория тематического моделирования**
   - Постановка задачи тематического моделирования
   - Теория аддитивной регуляризации
   - Альтернативный вывод EM-алгоритма
2. **Дальнейшие обобщения**
   - Модальности, дистрибутивная семантика, иерархии
   - Модели транзакционных данных
   - Модели с регуляризацией E-шага
3. **Оценивание качества и визуализация**
   - Внутренние (intrinsic) критерии качества
   - Внешние (extrinsic) критерии качества
   - Визуализация тематических моделей

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.

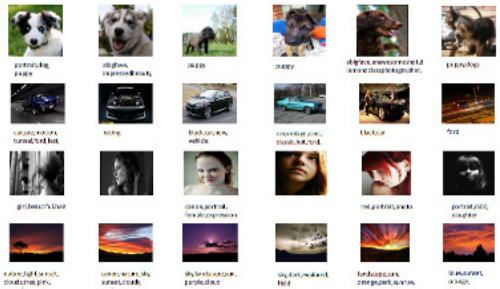
Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Приложения тематического моделирования

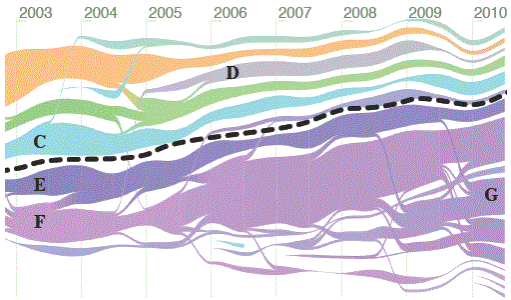
разведочный поиск в электронных библиотеках

рекомендательные с-мы и поиск в соцсетях

мультимодальный поиск текстов и изображений

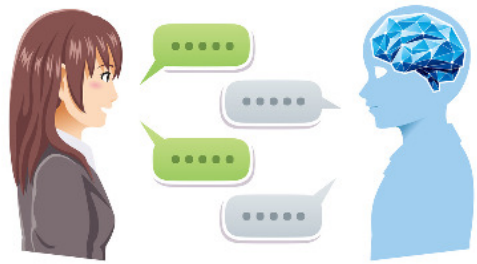


детектирование и трекинг новостных сюжетов



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: $p(w|d,t) = p(w|t)$

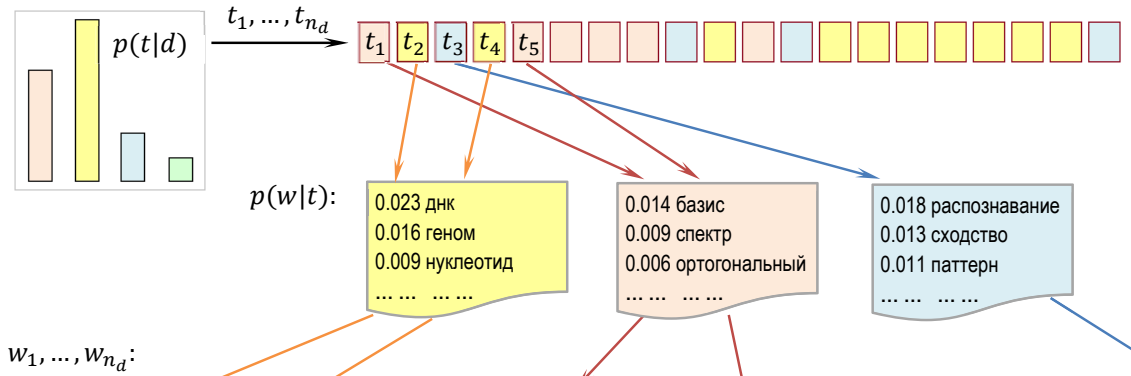
Тематическая модель, по формуле полной вероятности:

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

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

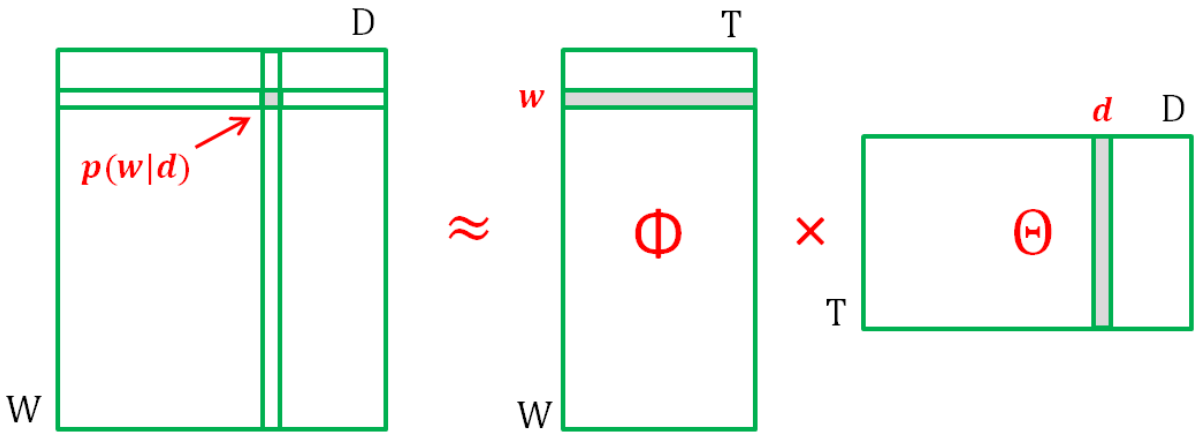
Дано: коллекция текстовых документов

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

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

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

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.

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:

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

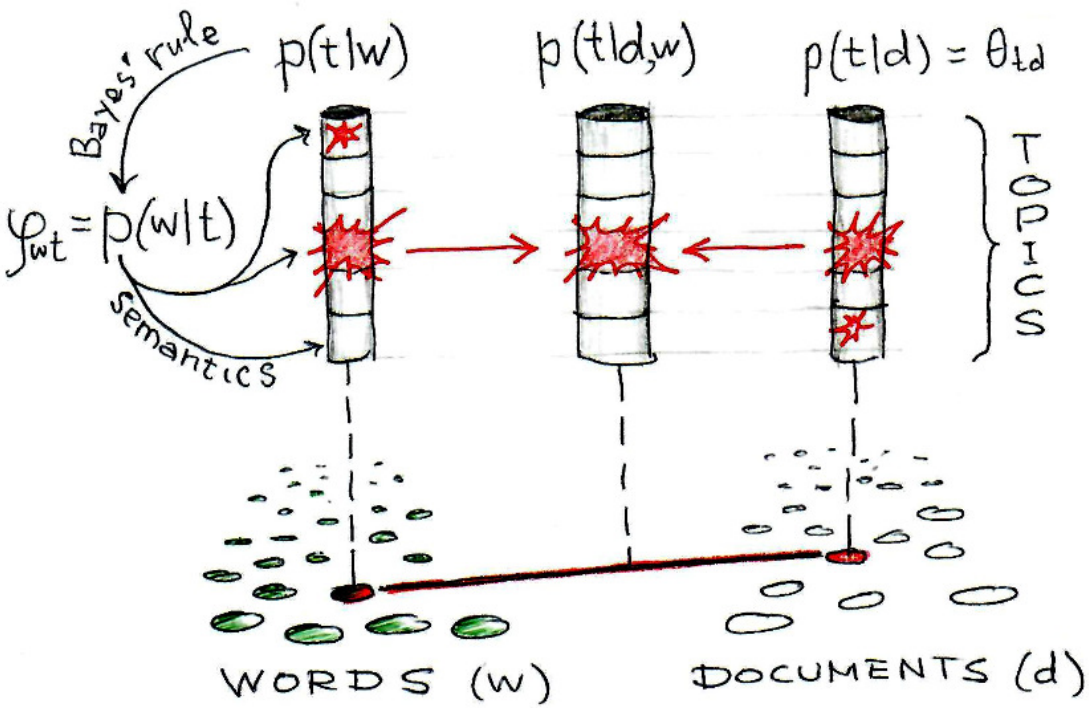
M-шаг: частотные оценки условных вероятностей вычисляются суммированием счётчика $n_{tdw} = n_{dw}p(t|d,w)$:

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

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

Интерпретируемые эмбеддинги 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.

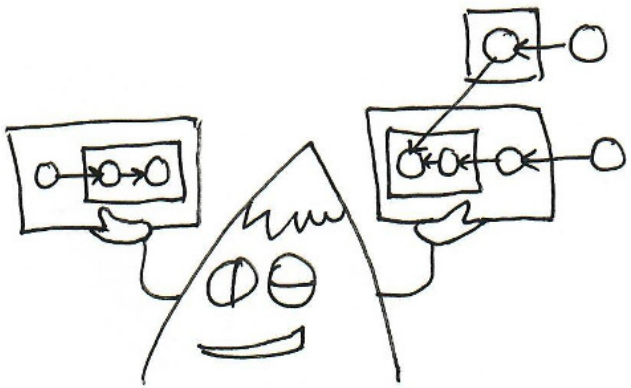
Обобщение №1: модель LDA

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.

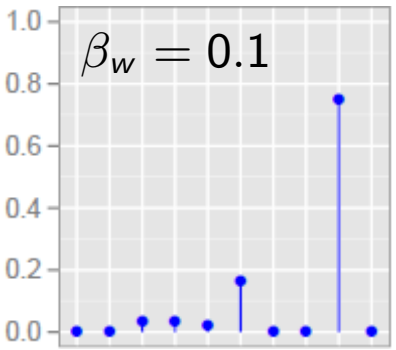
Вероятностная байесовская интерпретация LDA [Blei, 2003]

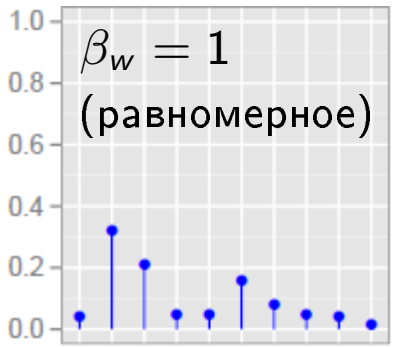
Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

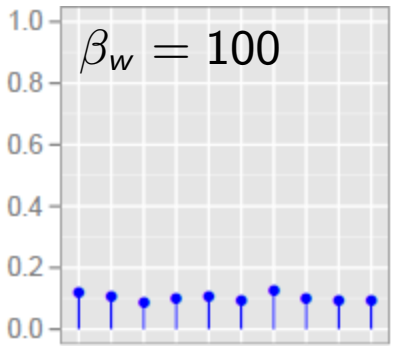
$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \ \beta_w > 0;$$

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

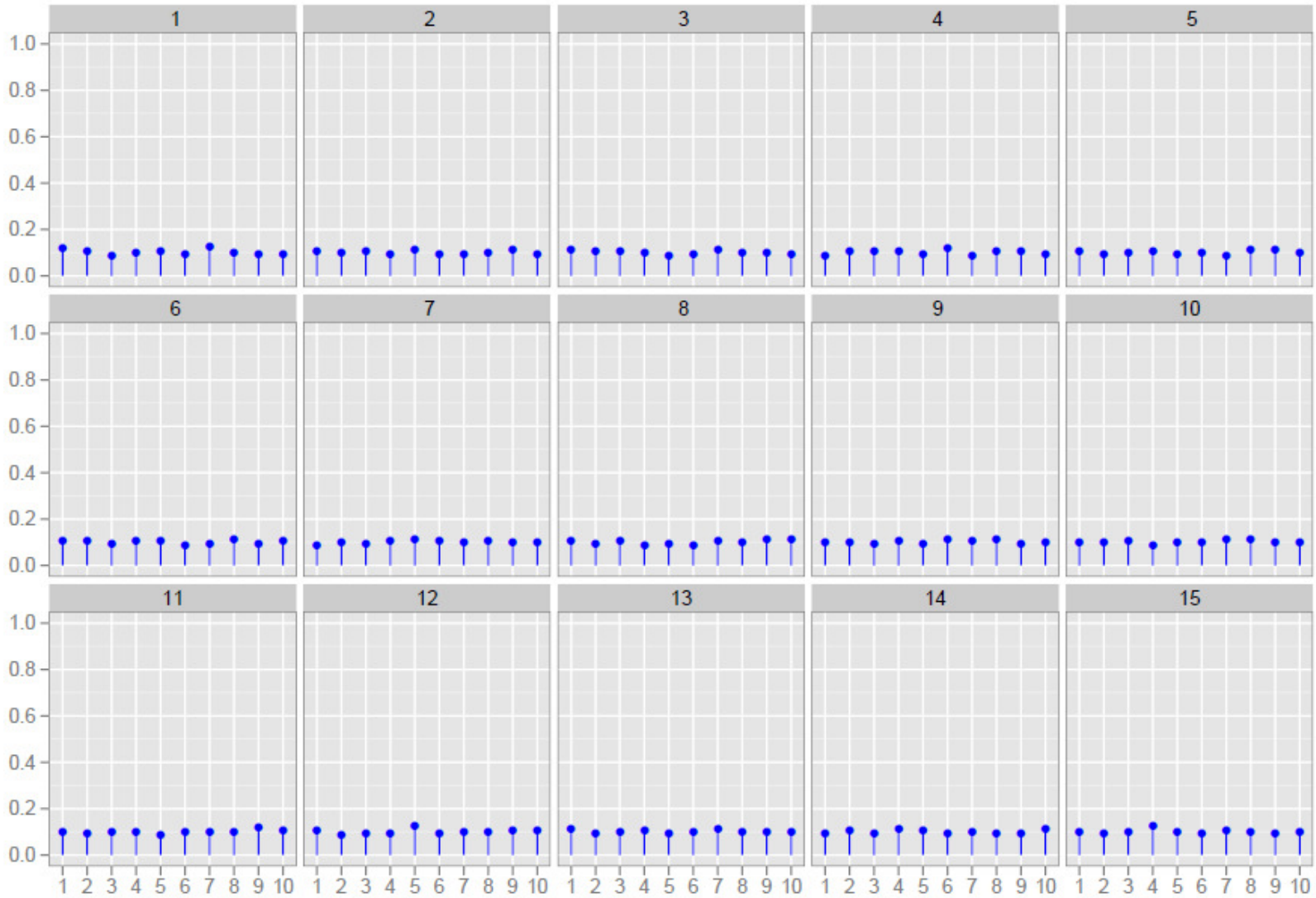
Пример. Распределение $\mathrm{Dir}(\phi|\beta)$ при $|W| = 10$, $\phi, \beta \in \mathbb{R}^{10}$:



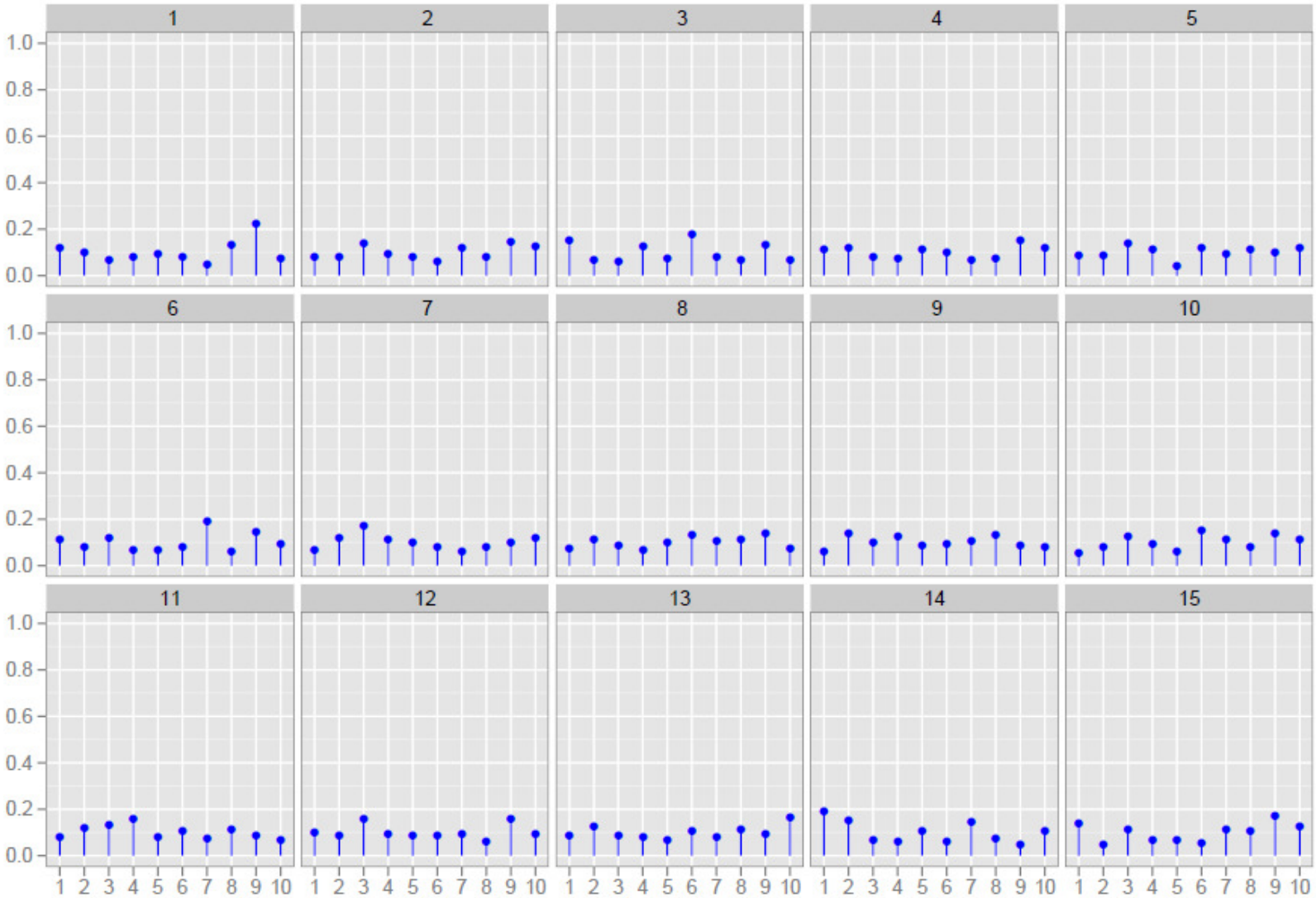




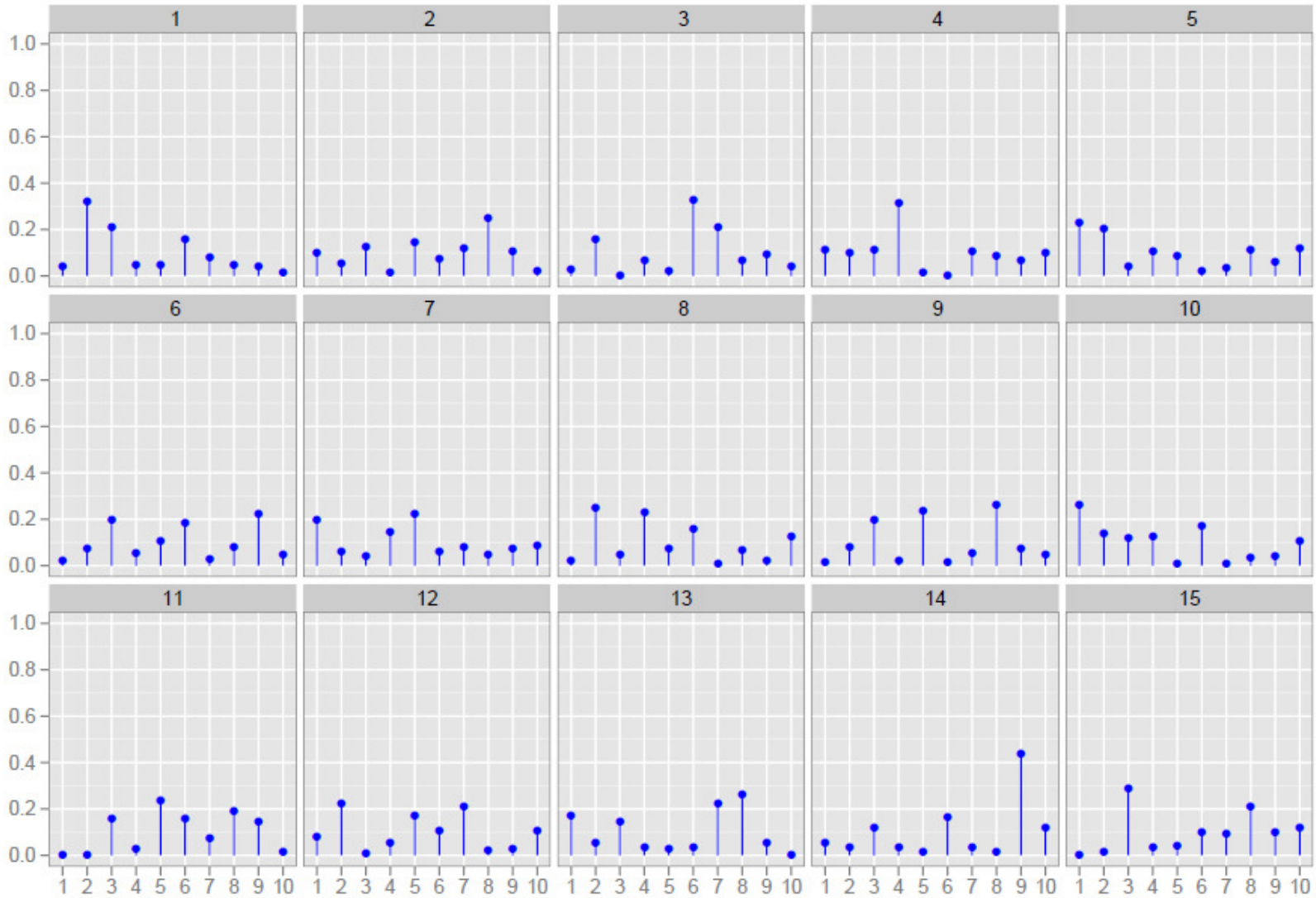
Распределение Дирихле при $\alpha_t \equiv 100$, 10 тем, 15 документов



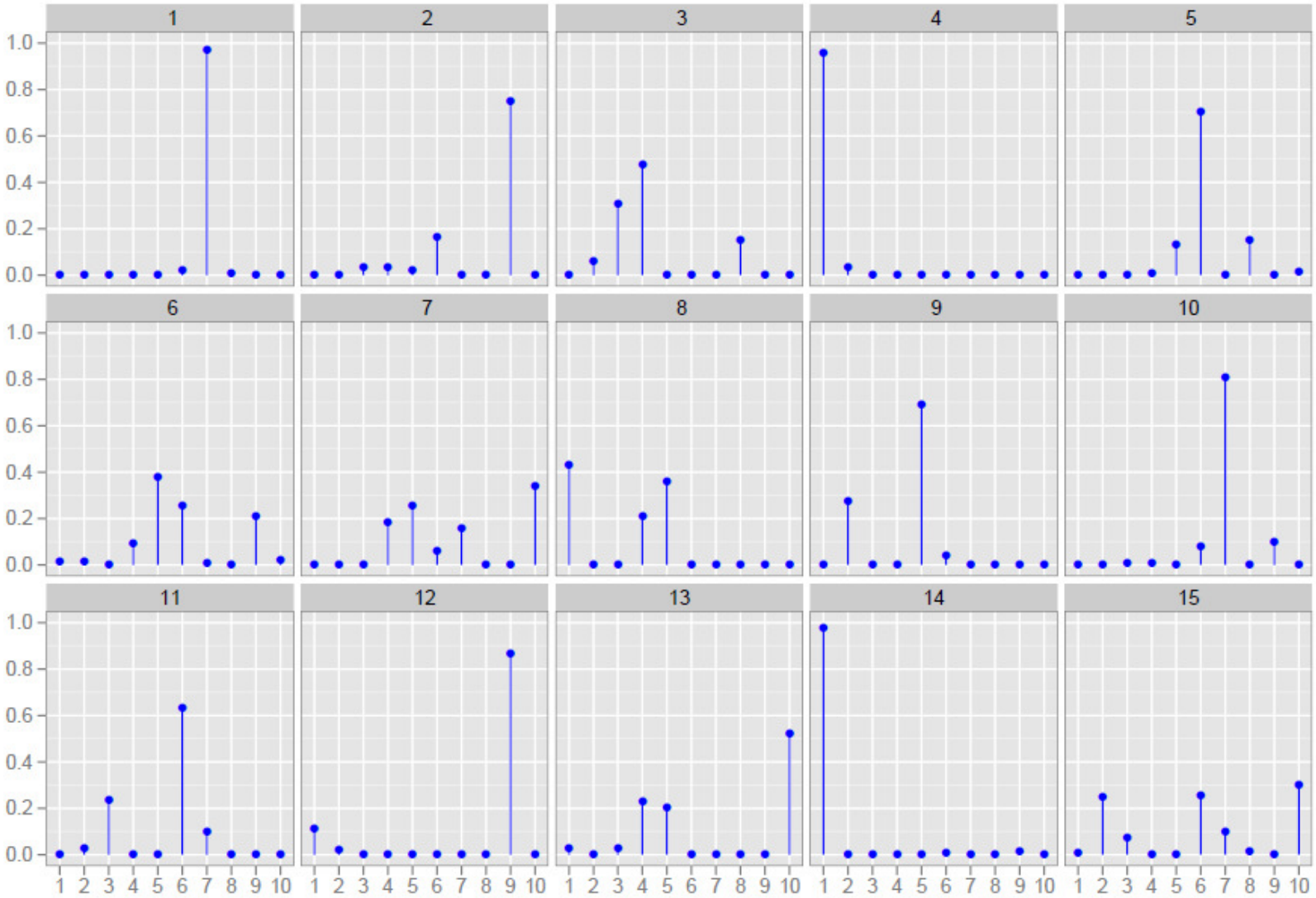
Распределение Дирихле при $\alpha_t \equiv 10$, 10 тем, 15 документов



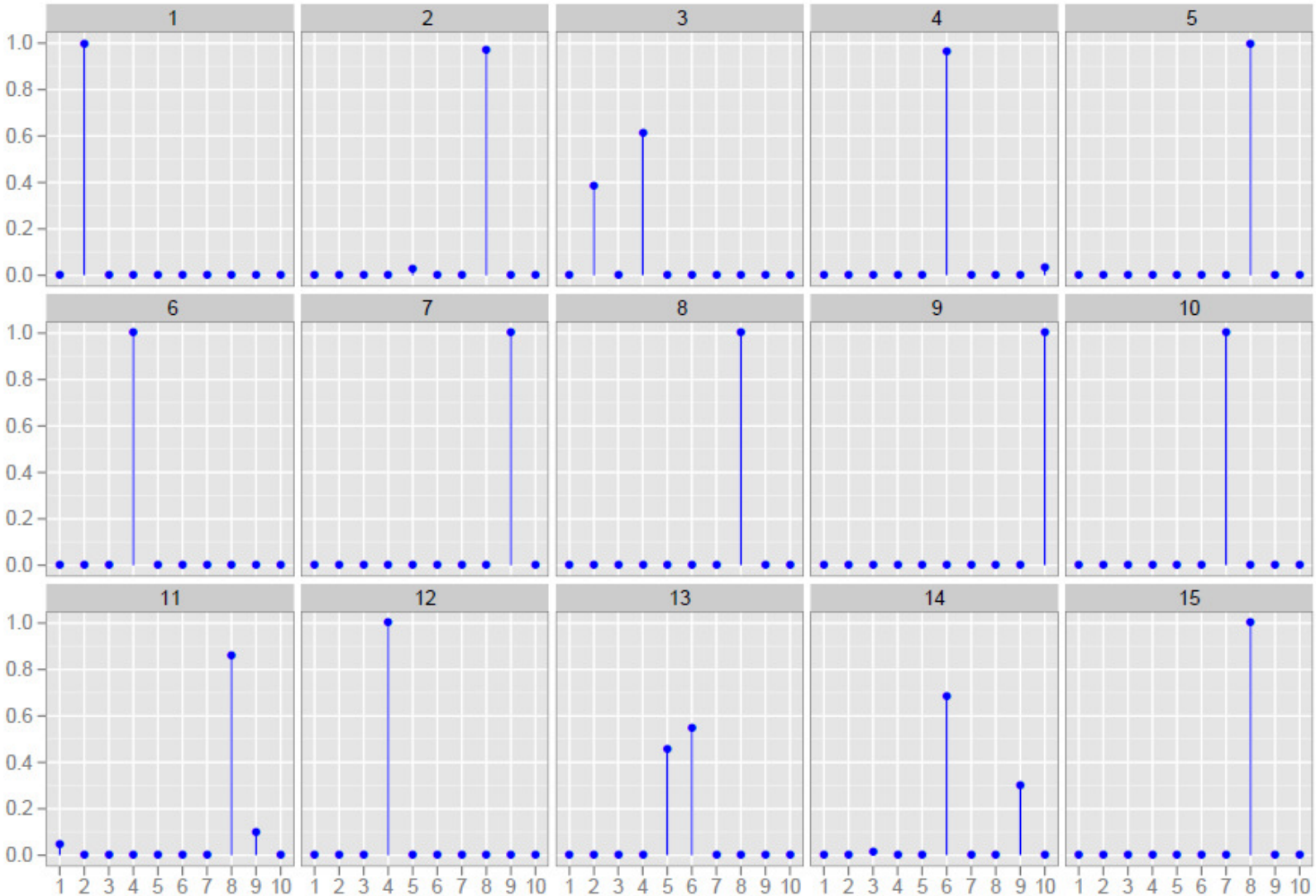
Распределение Дирихле при $\alpha_t \equiv 1$, 10 тем, 15 документов



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

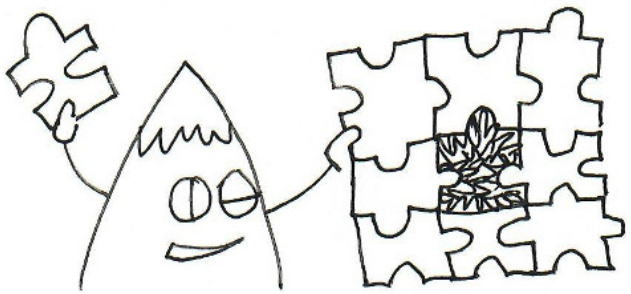
Обобщение №2: аддитивная регуляризация

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:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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.

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 $Q = (q_i)_{i=1}^n$:

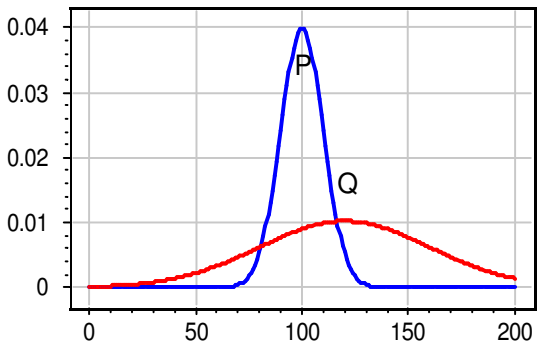
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

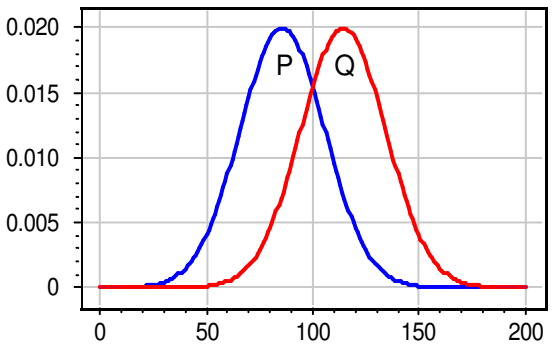
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

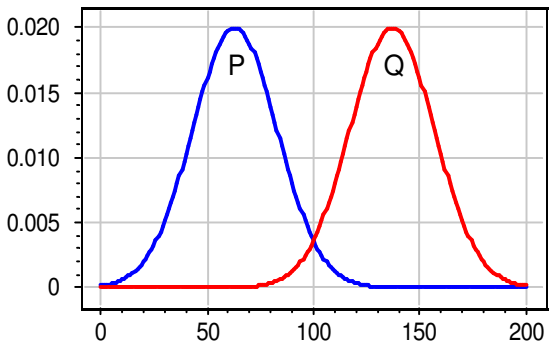
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



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



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



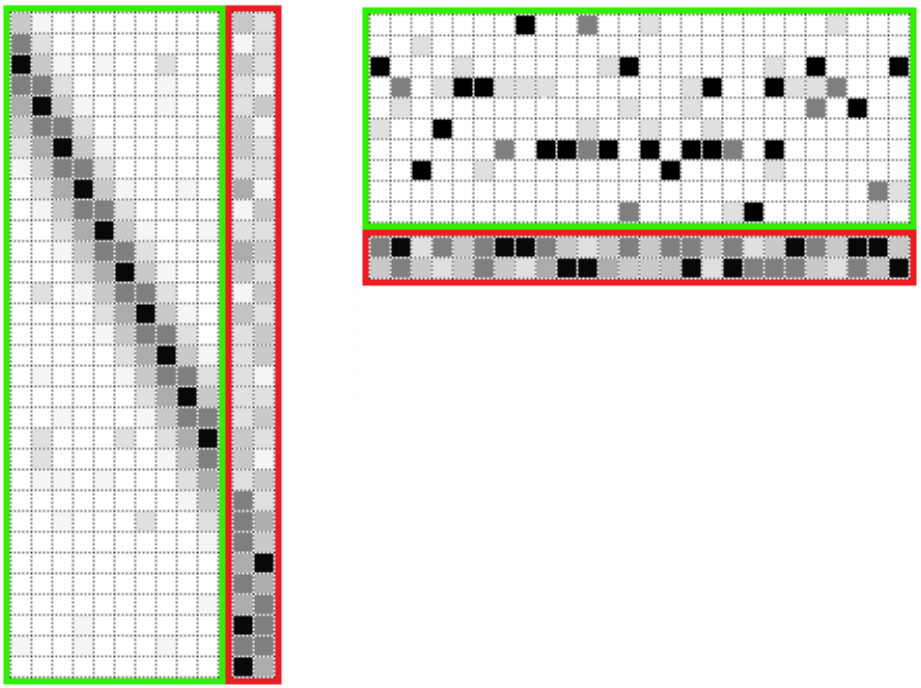
$\mathrm{KL}(P\|Q) = 2.969$
$\mathrm{KL}(Q\|P) = 2.969$

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

Наблюдаемые переменные описывают исходные данные:
$X = (d_i, w_i)_{i=1}^n$

Скрытые переменные объясняют появление данных:
$Z = (t_i)_{i=1}^n$

Вероятностная модель порождения данных:
$$p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$$
Параметры модели: $\Omega = (\Phi, \Theta)$, $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$

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:

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

Задача E-шага.

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

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

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

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

$D \times W \times T \times \{\Omega\}$ — вероятностное пространство
$p(\Omega)$ — 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.

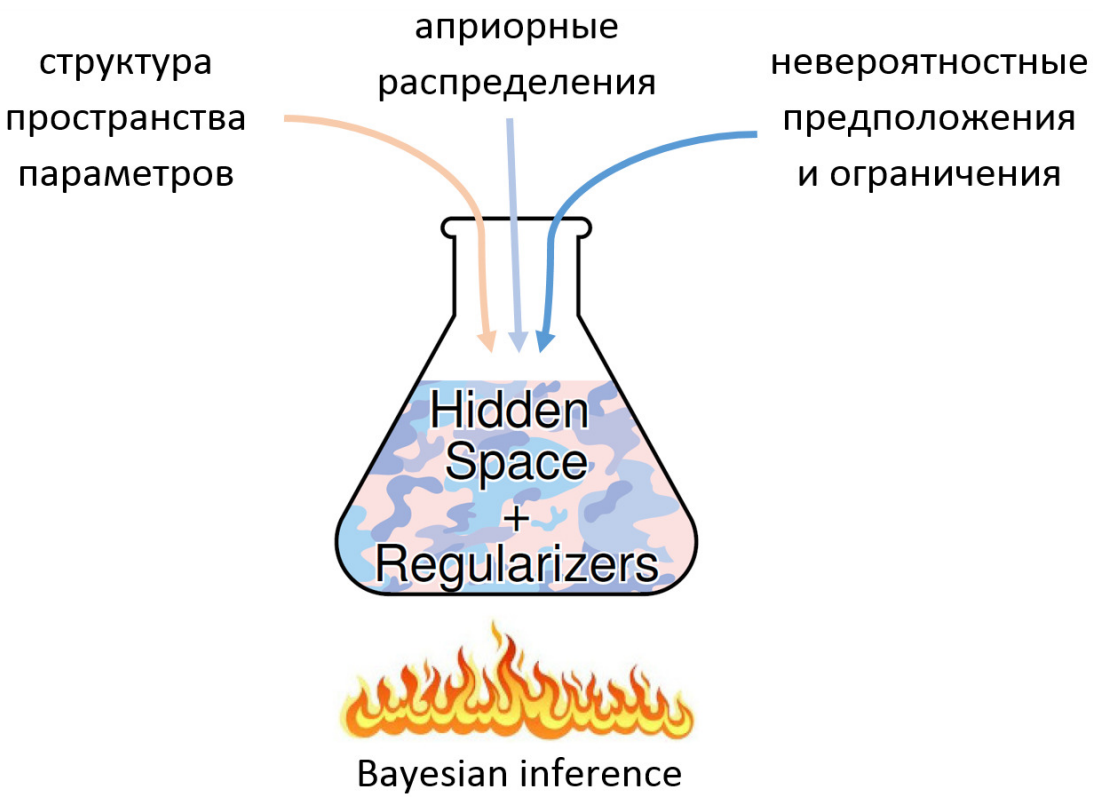
Проблемы:

- 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

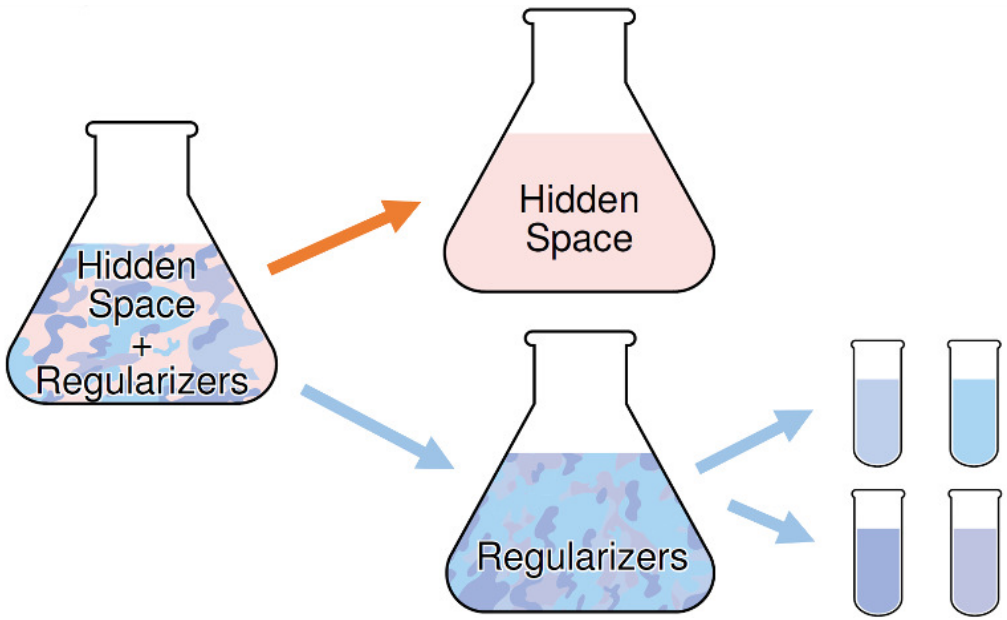
Вероятностная модель порождения данных объединяет в едином описании структуру пространства параметров, априорные распределения, дополнительные ограничения.



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:

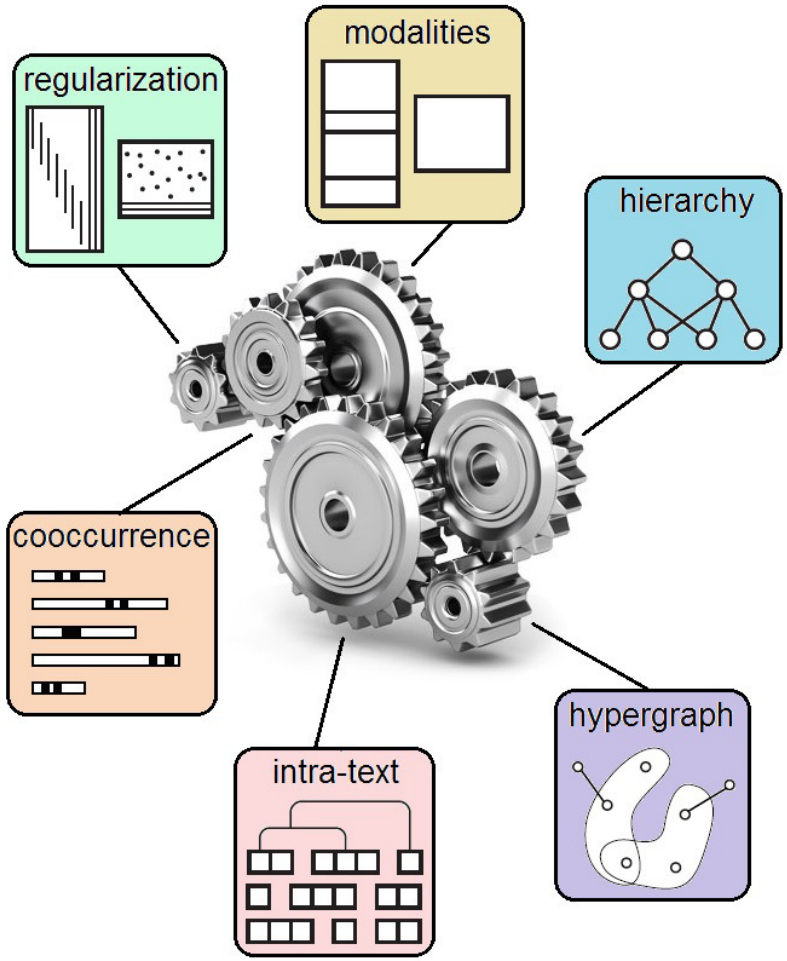
- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

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*

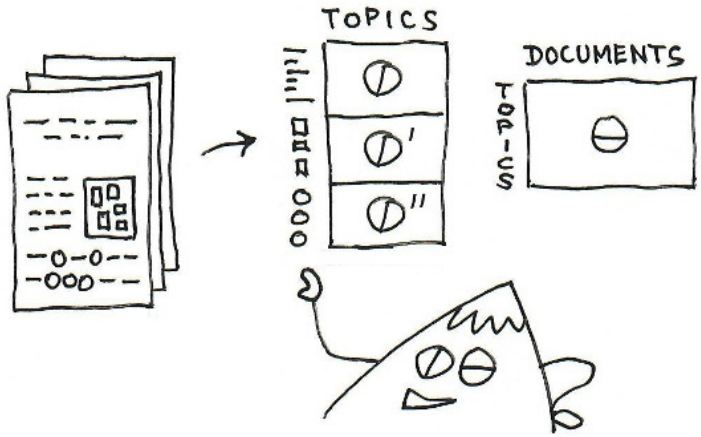
Обобщение №3: мультимодальные модели

Проблема

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

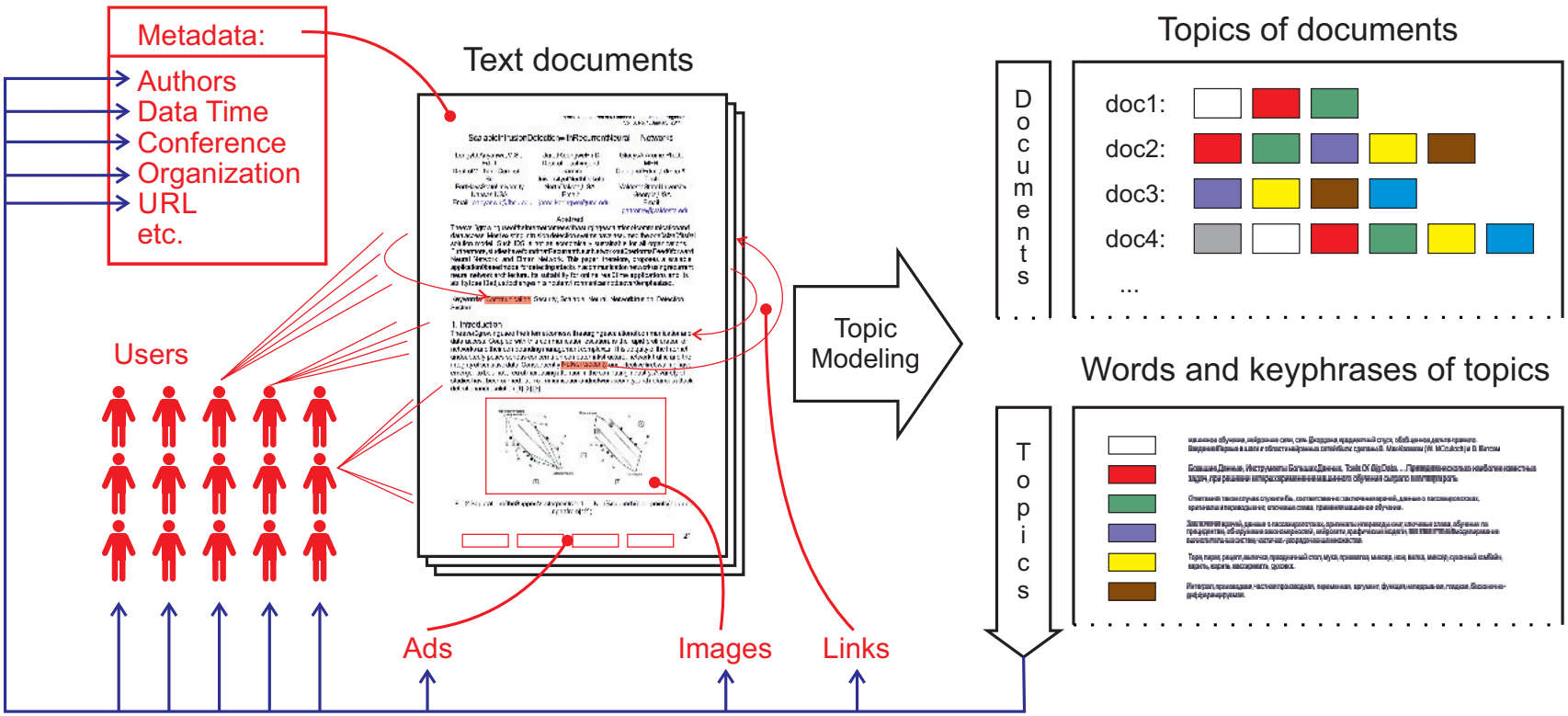
Решение

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

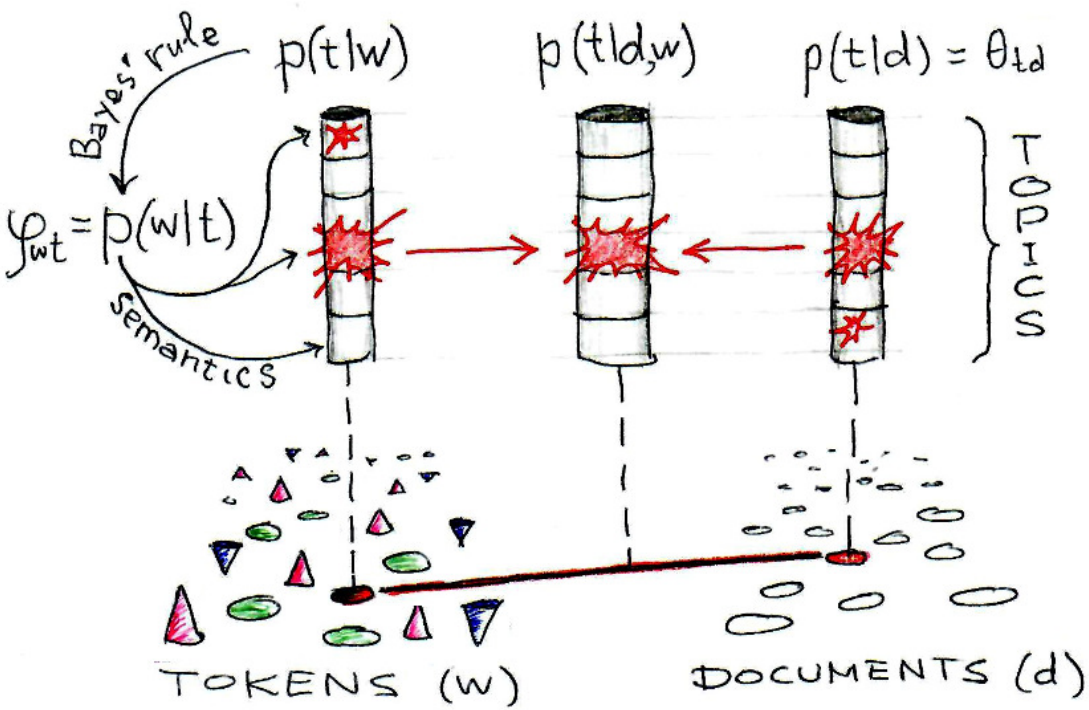
$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right)$$

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

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



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.

Регуляризатор для классификации и категоризации текстов

Y — множество классов;

$n_{dy} = [\text{документ } d \text{ относится к классу } y]$ — обучающие данные;

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

Регуляризатор — правдоподобие модальности классов:

$$R(\Phi, \Theta) = \tau \sum_{d \in D} \sum_{y \in Y} n_{dy} \ln \sum_{t \in T} \phi_{yt}\theta_{td} \ \to \ \max,$$

это тематическая модель с двумя модальностями, W и Y.
TM превосходит SVM в случае несбалансированных классов.

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification. Machine Learning, 2012.

Vorontsov, Frei, Apishev, Romov, Suvorova, Yanina. Non-Bayesian additive regularization for multimodal topic modeling of large collections. CIKM-2015 WTM.

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 M-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.

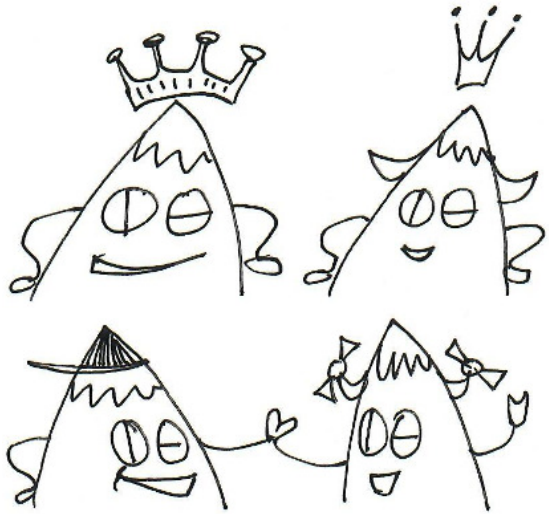
Обобщение №4: модели совстречаемости слов

Проблема

Тематические модели формируют векторные представления (эмбединги) слов, но почему-то они не способны решать задачи семантической близости слов, как word2vec.

Решение

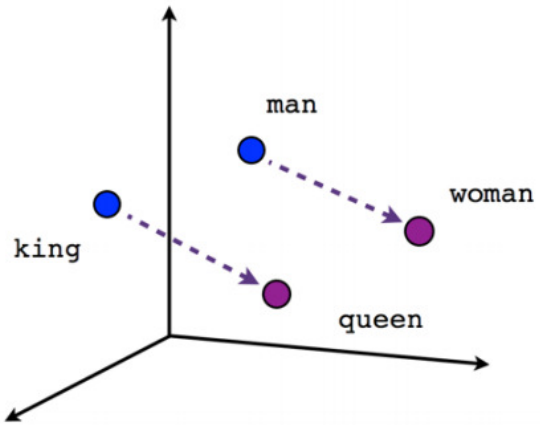
Понять, что такого есть в word2vec, и ввести это в TM.



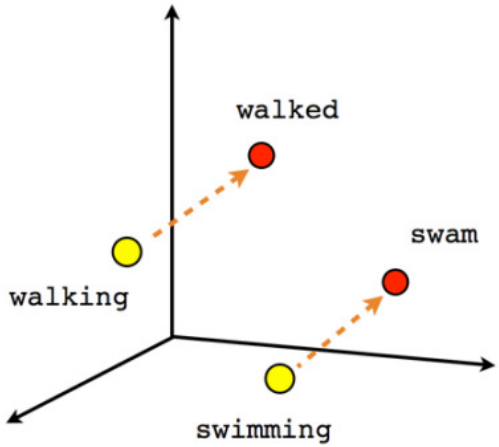
Задача семантического векторного представления слов

Найти для каждого слова w вектор $x_w \in \mathbb{R}^T$, чтобы близкие по смыслу слова имели близкие векторы.

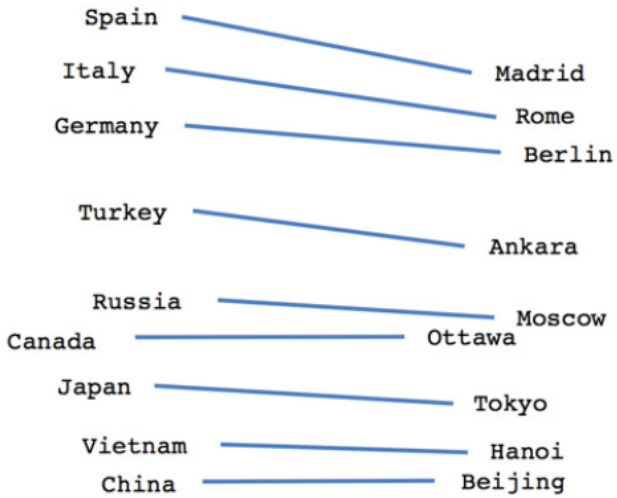
Задача семантической аналогии слов:
по трём словам угадать четвёртое.



Male-Female



Verb tense



Country-Capital

Дистрибутивная гипотеза

- Words that occur in the same contexts tend to have similar meanings [Harris, 1954].
- You shall know a word by the company it keeps [Firth, 1957].

Синтагматическая близость слов:
со-встречаемость слов в одном контексте.

здание–строитель, кран–вода, функция–точка

Парадигматическая близость слов:
взаимозаменяемость слов в одном контексте.

здание–дом, кран–смеситель, функция–отображение

Z.Harris. Distributional structure. 1954.
J.R.Firth. A synopsis of linguistic theory 1930-1955. Oxford, 1957.
P.D.Turney, P.Pantel. From frequency to meaning: Vector space models of semantics // Journal of Artificial Intelligence Research (JAIR). 2010.

Модели векторных представлений для текстов и графов

word2vec: эмбединги слов

T.Mikolov et al. Efficient estimation of word representations in vector space. 2013.

paragraph2vec: эмбединги фрагментов или документов

Q.Le, T.Mikolov. Distributed representations of sentences and documents. 2014.

sent2vec: эмбединги предложений

M.Pagliardini et al. Unsupervised learning of sentence embeddings using compositional n-gram features. 2017.

FastText: эмбединги символьных n-грамм

`https://github.com/facebookresearch/fastText`

node2vec: эмбединги вершин графа

A.Grover, J.Leskovec. Node2vec: scalable feature learning for networks. 2016.

graph2vec: более общие эмбединги на графах

A.Narayanan et al. Graph2vec: learning distributed representations of graphs. 2017.

StarSpace: эмбединги чего угодно от Facebook AI Research

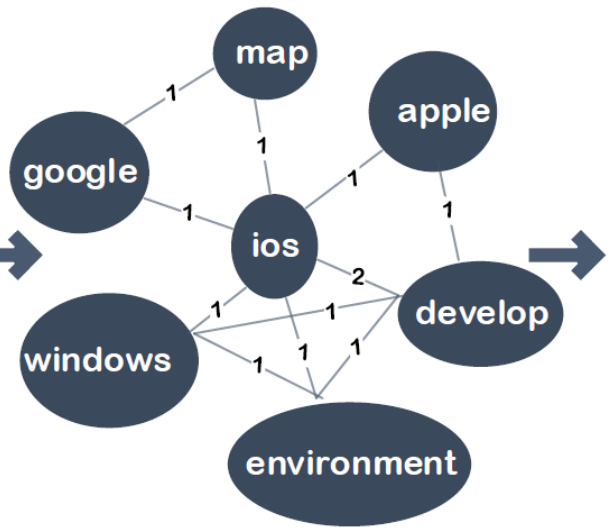
L.Wu, A.Fisch, S.Chopra, K.Adams, A.B.J.Weston. StarSpace: embed all the things! 2018.

Недостаток: координаты векторов не интерпретируемы

Модель сети слов WNTM для коротких текстов

Идея: моделировать не документы, а связи между словами.
d_u — псевдо-документ, объединение всех контекстов слова u.
n_{uw} — число вхождений слова w в псевдо-документ d_u.
Контекст — короткое сообщение / предложение / окно $\pm h$ слов.

doc1	Google Map for iOS
doc2	iOS is developed by Apple
doc3	iOS develop environment for Windows



word	pseudo-document
map	google ios
google	map ios
apple	develop ios
ios	google map apple windows environment develop develop
windows	environment ios develop
develop	apple ios ios windows environment
environment	ios windows develop

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Модели WNTM и WTM (Word Topic Model)

Тематическая модель контекстов, разложение $W \times W$-матрицы:

$$p(w|d_u) = \sum_{t \in T} p(w|t)p(t|d_u) = \sum_{t \in T} \phi_{wt}\theta_{tu},$$

где d_u — псевдо-документ слова u.

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{uw} \log \sum_{t \in T} \phi_{wt}\theta_{tu} \to \max_{\Phi,\Theta},$$

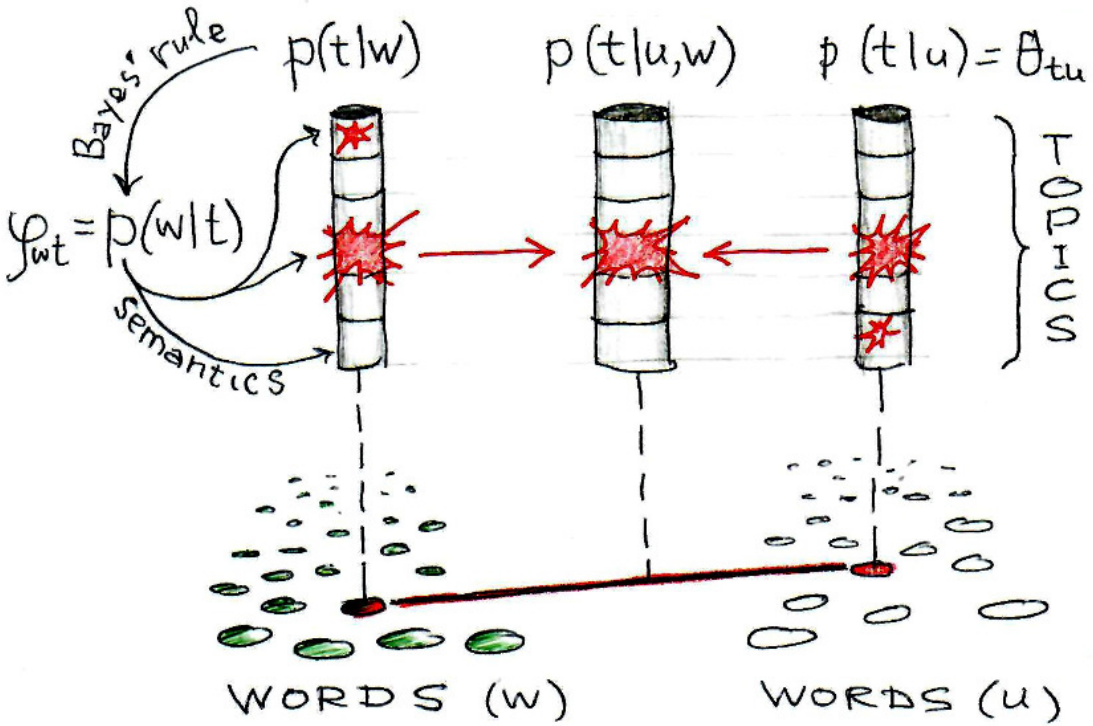
где n_{uw} — совстречаемость слов u, w (кстати, $n_{uw} = n_{wu}$).

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Berlin Chen. Word Topic Models for spoken document retrieval and transcription. ACM Trans., 2009.

Интерпретируемые эмбединги совстречаемости слов

- Идея *дистрибутивной семантики*: “Words that occur in the same contexts tend to have similar meanings” [Harris, 1954].
- Слово индуцирует псевдо-документ всех его контекстов



word2vec и ARTM на задачах аналогии слов

Два подхода к синтезу векторных представлений слов:

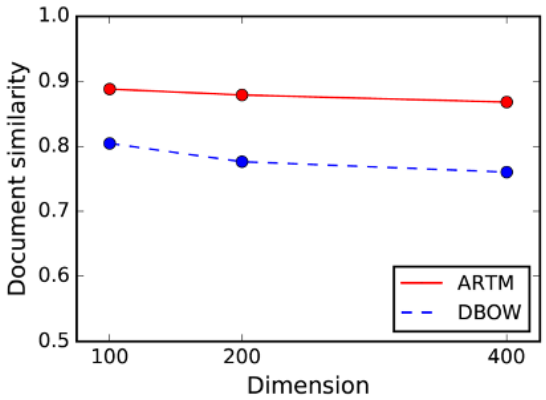
- **ARTM**: интерпретируемые разреженные компоненты
- **word2vec**: интерпретируемые векторные операции

Операция	Результат ARTM	Результат word2vec
king – boy + girl	*queen*, princess, lord, prince	*queen*, princess, regnant, kings
moscow – russia + spain	*madrid*, barcelona, aires, buenos	*madrid*, barcelona, valladolid, malaga
india – russia + ruble	*rupee*, birbhum, pradesh, madhaya	*rupee*, rupiah, devalued, debased
cars – car + computer	*computers*, software, servers, implementations	*computers*, software, hardware, microcomputers

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

word2vec и ARTM в задаче семантической близости документов

ArXiv triplets dataset: 20K троек статей:
⟨ статья A, схожая статья B, непохожая статья C ⟩



- обучение по 1M текстов статей ArXiv
- тестирование на триплетах ArXiv
- Конкурент DBOW: paragraph2vec [Dai et. al, 2015]

ARTM превосходит модель DBOW (distributed bag-of-words).

Andrew Dai, Cristopher Olah, Quoc Le. Document Embedding with Paragraph Vectors, CoRR, 2015

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

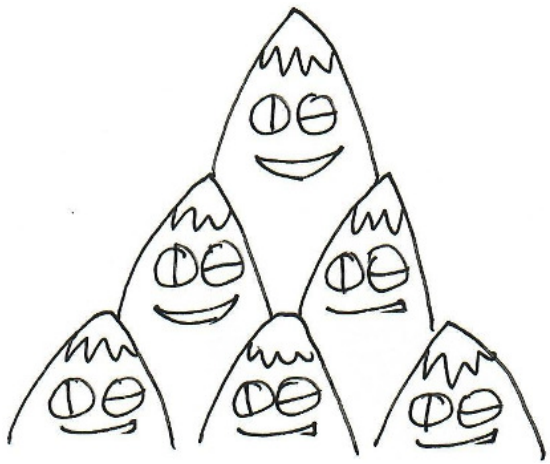
Обобщение №5: иерархические модели

Проблема

Невозможно определить оптимальное число тем.
Хотелось бы разделять темы на подтемы иерархически.
Придумано много иерархических моделей, но они либо ограниченные, либо тормозные, либо замороченные.

Решение

Придумать что-то радикально простое



Послойное построение уровней тематической иерархии

Шаг 1. Строим модель с небольшим числом тем.

Шаг *k*. Пусть модель с множеством тем T уже построена. Строим множество дочерних тем S (subtopics), $|S| > |T|$.

Родительские темы приближаются смесями дочерних тем:

$$\sum_{t\in T} n_t \,\mathrm{KL}_w\Big(p(w|t) \,\Big\|\, \sum_{s\in S} p(w|s)p(s|t)\Big) \;\to\; \min_{\Phi,\Psi},$$

где $p(s|t) = \psi_{st}$, $\Psi = (\psi_{st})_{S\times T}$ — *матрица связей*.

Родительская $\Phi^p \approx \Phi\Psi$, отсюда регуляризатор матрицы Φ:

$$R(\Phi,\Psi) = \tau \sum_{t\in T}\sum_{w\in W} n_{wt} \ln \sum_{s\in S} \phi_{ws}\psi_{st} \;\to\; \max.$$

Родительские темы t — *псевдо-документы* с частотами слов n_{wt}.

Обобщение №6: модели транзакционных данных

Проблема

Исходные данные могут быть сложнее, чем парные взаимодействия (транзакции) между объектами

Решение

Тематическая модель должна описывать транзакции, состоящие из любых подмножеств объектов

Транзакционные данные

Выборка может содержать не только пары (d, w), но также тройки, четвёрки, ..., n-ки элементов разных модальностей.

Примеры:

- **Данные социальной сети:**
  (d, u, w) — пользователь u записал слово w в блоге d
- **Данные сети интернет-рекламы:**
  (u, d, b) — пользователь u кликнул баннер b на странице d
- **Данные рекомендательной системы:**
  (u, f, s) — пользователь u оценил фильм f в ситуации s
- **Данные финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g

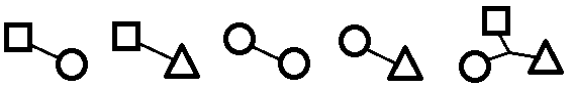
Задача: по наблюдаемой выборке рёбер гиперграфа выявить латентные темы его вершин.

Тематическая модель гиперграфа: определения и обозначения

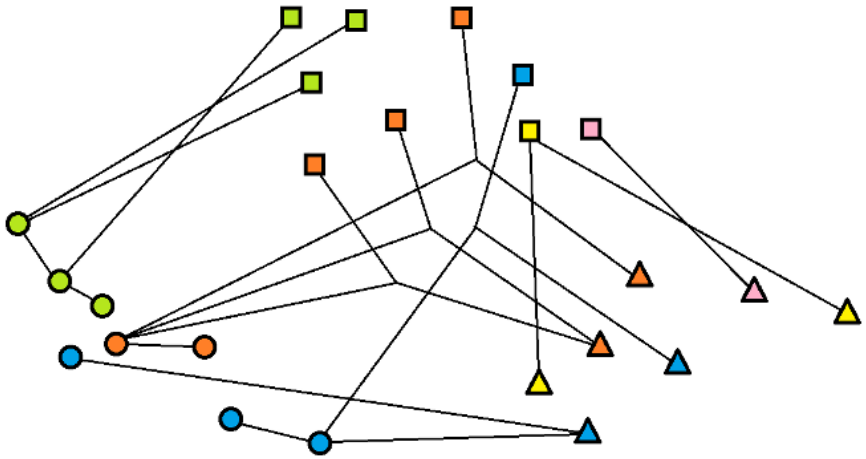
$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:



T — множество тем:



X^k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x): вершина-контейнер $d \in V$ и вершины $x \subset V$,
n_{dx} — число вхождений ребра (d, x) в выборку X^k
$p(d, x)$ — неизвестное распределение на рёбрах типа k

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k:

$$p(x|d) = \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{vt},$$

$\theta_{td} = p(t|d)$ — тематика контейнера не зависит от типа ребра k
$\phi_{vt} = p(v|t)$ — распределение термов модальности v в теме t

Задача максимизации log правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in X^k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{vt} \to \max_{\Phi, \Theta},$$

$$\phi_{vt} \geqslant 0, \quad \sum_{v \in V^m} \phi_{vt} = 1; \qquad \theta_{td} \geqslant 0, \quad \sum_{t \in T} \theta_{td} = 1;$$

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k\in K}\tau_k \sum_{(d,x)\in X^k} n_{dx} \ln \sum_{t\in T} \theta_{td} \prod_{v\in x} \phi_{vt} \;+\; R(\Phi,\Theta) \to \max_{\Phi,\Theta}.$$

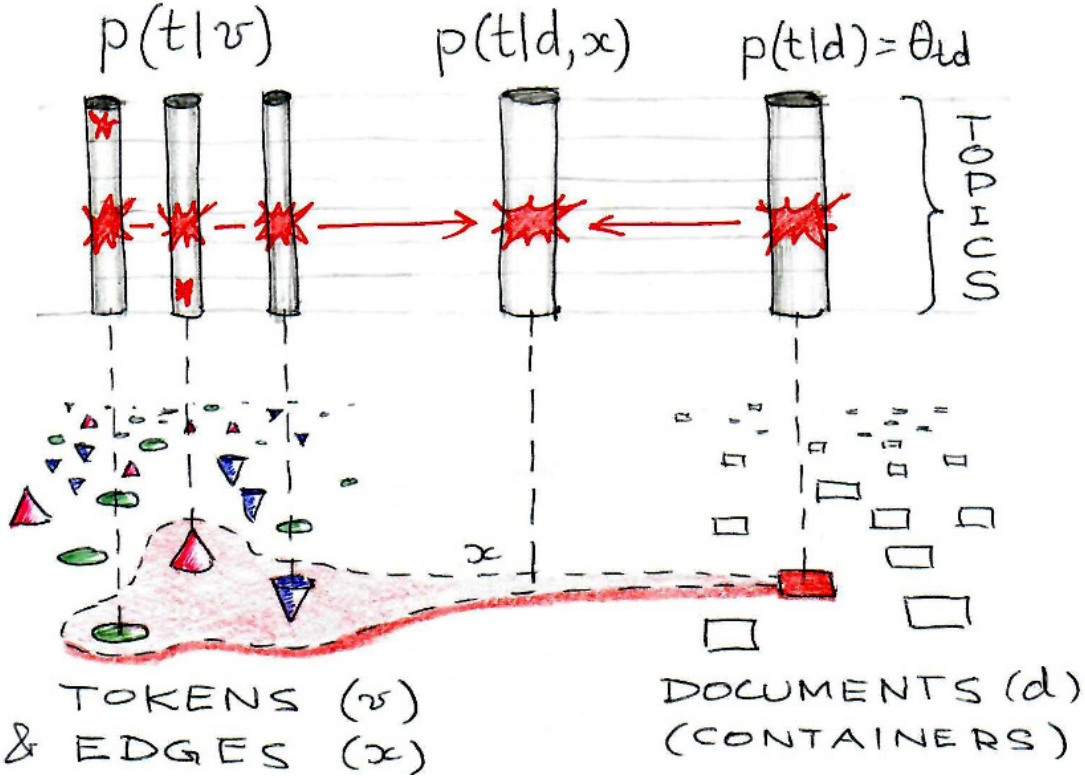
EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdx} = p(t|d,x)$:

$$\text{E-шаг:}\quad p_{tdx} = \underset{t\in T}{\text{norm}}\left(\theta_{td}\prod_{v\in x}\phi_{vt}\right)$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{vt} = \underset{v\in V^m}{\text{norm}}\left(\sum_{k\in K}\tau_k \sum_{(d,x)} \big[v\in x\big] n_{dx} p_{tdx} + \phi_{vt}\frac{\partial R}{\partial \phi_{vt}}\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{k\in K}\tau_k \sum_{(d,x)} n_{dx} p_{tdx} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

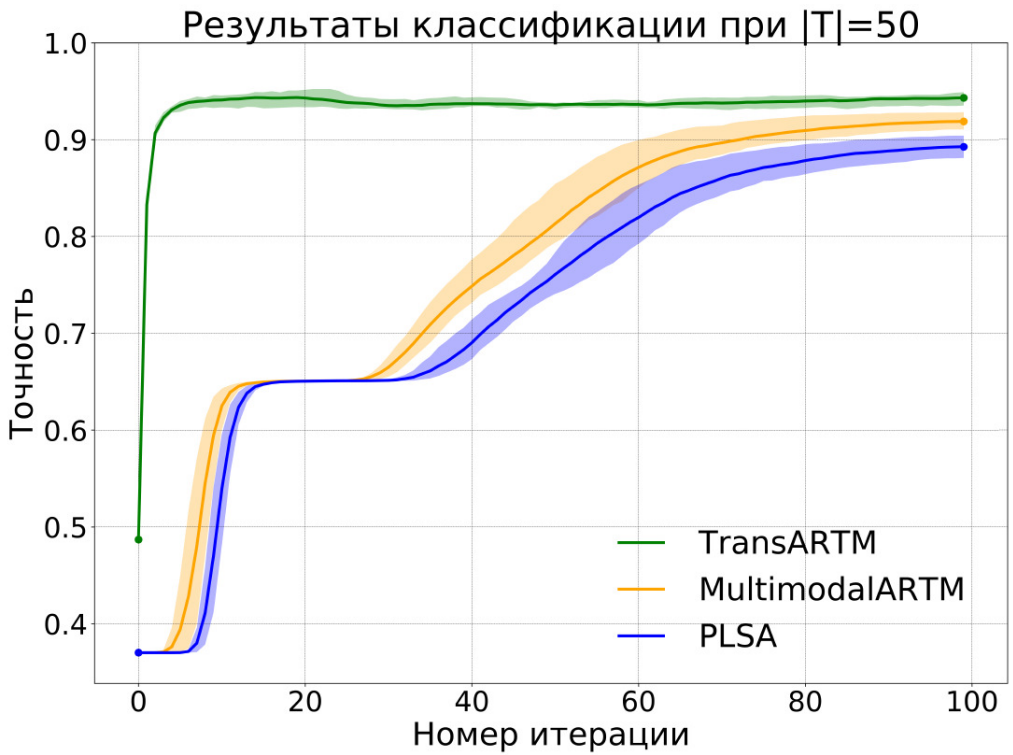
Интерпретируемые эмбединги транзакционных данных

- *Гиперграф* — множество подмножеств вершин-токенов
- Транзакция = подмножество токенов = ребро гиперграфа
- Транзакция происходит, когда токены имеют общие темы



Эксперименты на модельных данных

13M транзакций, 3 модальности, 5 классов, 9 типов рёбер



Вывод: обычные модели не могут восстановить гиперграф.

Илья Кариков. Гиперграфовые тематические модели транзакционных данных. Магистерская диссертация, МФТИ, 2018.

Модели предложений и коротких текстов TwitterLDA, senLDA

S_d — множество предложений документа d

n_{sw} — сколько раз терм w встречается в предложении s

Тематическая модель предложения s:

$$p(s|d) = \sum_{t\in T} p(t|d) \prod_{w\in s} p(w|t)^{n_{sw}} = \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}}$$

Максимизация регуляризованного log-правдоподобия

$$\sum_{d\in D} \sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

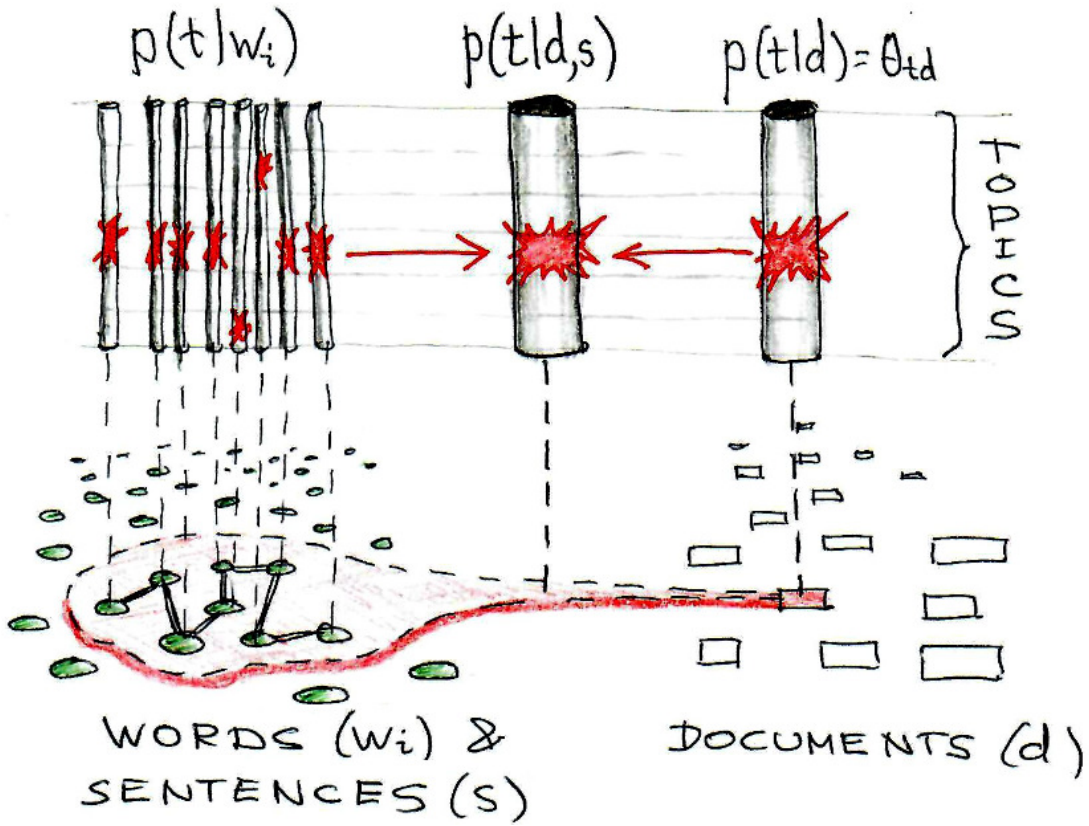
это частный случай гиперграфовой модели, в которой предложения являются «транзакциями» или гипер-рёбрами.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim et al. Comparing Twitter and traditional media using topic models. ECIR 2011.

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Интерпретируемые эмбединги предложений

- Предложение — семантически однородная единица языка
- Предложение образуется из слов, имеющих общие темы
- Предложение = подмножество слов = ребро гиперграфа



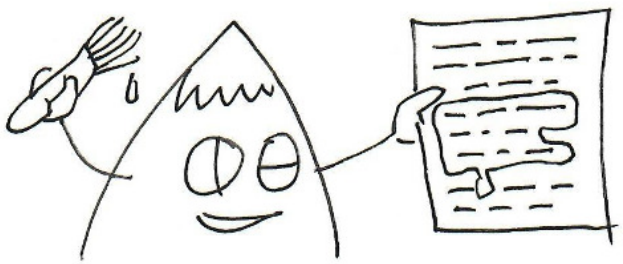
Обобщение №7: модели с регуляризацией E-шага

Проблема

Гипотеза «мешка слов» — самое часто критикуемое допущение тематического моделирования.
Как строить модели, учитывающие порядок слов?

Решение

Пост-обработка $p(t|d, w_i)$ как пучка временных рядов, например, сглаживание или сегментирование, с учётом предложений, секционирования, синтаксических связей, лексических цепочек, и т. д.



Сегментная структура текста и пост-обработка E-шага

Документ $d = \{w_1, \ldots, w_{n_d}\}$, n_d — длина документа d

Матрица *тематики слов в документах* $p(t|d, w_i)$ размера $T \times n_d$:

Регуляризация E-шага

Трёхмерная матрица $\Pi = \left(p_{tdw} = p(t|d,w)\right)_{T\times D\times W}$

Максимизация log правдоподобия с регуляризаторами R и $\tilde{R}$:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ & \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}}\right) \end{cases}$$

Набросок доказательства: три леммы

Лемма 1. Для функции $p_{tdw}(\Phi,\Theta) = \frac{\phi_{wt}\theta_{td}}{\sum_z \phi_{wz}\theta_{zd}}$ и любого $z \in T$

$$\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}} = \theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

Введём функцию от вспомогательных переменных Π:

$$Q_{tdw}(\Pi) = \frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}.$$

Лемма 2. Если $R(\Pi)$ не зависит от p_{tdw} при $w \notin d$, то

$$\phi_{wt}\frac{\partial R(\Pi)}{\partial \phi_{wt}} = \sum_{d\in D} p_{tdw}Q_{tdw}(\Pi); \qquad \theta_{td}\frac{\partial R(\Pi)}{\partial \theta_{td}} = \sum_{w\in d} p_{tdw}Q_{tdw}(\Pi).$$

Лемма 3. Формулы M-шага:

$$\phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \sum_{d\in D} Q_{tdw}p_{tdw} + \phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}}\Big);$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \sum_{w\in d} Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}}\Big).$$

Гипотеза: пост-обработка E-шага — это неявная регуляризация

Между E- и M-шагом добавляется обработка матрицы $p_{tdw} = p(t|d,w)$ тематики слов документа:

$$\tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right) \qquad (1)$$

Пост-обработка E-шага позволяет учитывать порядок слов в каждом документе в обход гипотезы «мешка слов».

Гипотеза

Любое «разумное» преобразование $p_{tdw} \to \tilde{p}_{tdw}$ эквивалентно некоторому регуляризатору $R(\Pi(\Phi,\Theta))$.

Открытый вопрос: при каких условиях по заданным p_{tdw} и $\tilde{p}_{tdw}$ возможно подобрать функцию $R(\Pi)$ так, чтобы выполнялось уравнение пост-обработки (1)?

Пример 1. Кросс-энтропийное разреживание $p(t|d,w)$

Путь каждый термин относится к небольшому числу тем:

$$\mathrm{KL}\Big(\tfrac{1}{|T|} \,\Big\|\, p(t|d,w)\Big) \to \max.$$

Суммируем по всем терминам всех документов:

$$R(\Pi) = -\frac{\tau}{|T|}\sum_{d\in D}\sum_{w\in d} n_{dw}\sum_{t\in T}\ln p_{tdw} \to \max.$$

Формула регуляризованного E-шага:

$$\tilde{p}_{tdw} = p_{tdw} - \tau\Big(\tfrac{1}{|T|} - p_{tdw}\Big).$$

Интерпретация: Если $p_{tdw} < \frac{1}{|T|}$, то p_{tdw} станет ещё меньше.
Тематика термина концентрируется в небольшом числе тем.

Недостаток: Тематика соседних слов разреживается независимо.

Пример 2. Тематическая модель сегментированного текста

S_d — множество микро-сегментов документа d

n_{sw} — число вхождений слова w в сегмент s длины n_s

Тематика сегмента $s \in S_d$ — средняя тематика его слов:

$$p_{tds} \equiv p(t|d,s) = \frac{1}{n_s} \sum_{w \in s} n_{sw} p_{tdw}.$$

Кросс-энтропийный регуляризатор разреживания $p(t|d,s)$:

$$R(\Pi) = -\sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T} \ln \sum_{w \in s} n_{sw} p_{tdw} \to \max.$$

Формула регуляризованного E-шага:

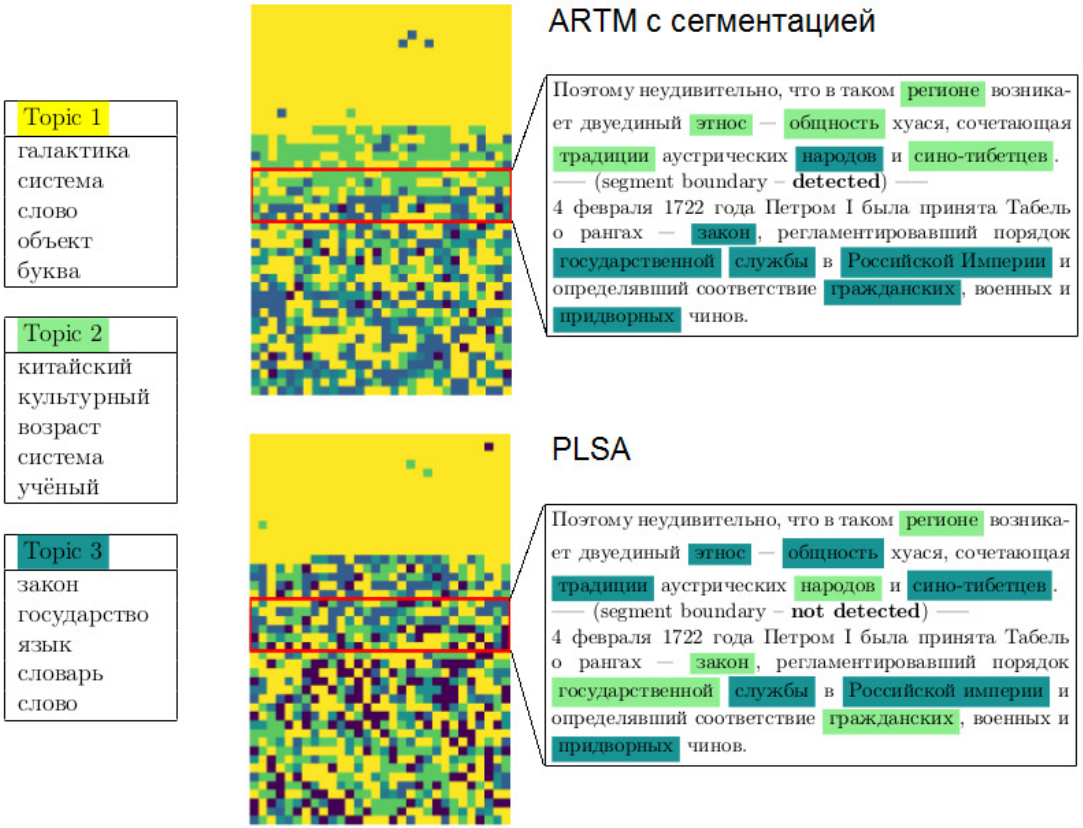
$$\tilde{p}_{tdw} = p_{tdw} \left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s} \left(\frac{1}{p_{tds}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zds}}\right)\right).$$

Интерпретация: если $p_{tds} < \frac{1}{|T|}$, то p_{tdw} уменьшатся $\forall\, w \in s$.

Тематика сегмента концентрируется в небольшом числе тем.

Пример. Регуляризатор E-шага для сегментации текста

Полусинтетическая коллекция из фрагментов postnauka.ru



N.Skachkov, K.Vorontsov. Improving topic models with segmental structure of texts. Dialogue, 2018.

Правдоподобие и перплексия (perplexity)

Правдоподобие языковой модели $p(w|d)$ (чем выше, тем лучше):

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

Перплексия языковой модели $p(w|d)$ (чем меньше, тем лучше):

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

Интерпретация перплексии:

- если распределение $p(w|d) = \frac{1}{|W|}$ равномерное, то $\mathcal{P} = |W|$
- мера различности или неопределённости слов в тексте
- коэффициент ветвления (branching factor) текста

Перплексия тестовой (отложенной) коллекции

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

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

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

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

Интерпретируемости и когерентность

Тема интерпретируемая, если по топовым словам темы эксперт может определить, о чём эта тема, и дать ей название.

- Экспертные оценки:
  — интерпретируемость темы по балльной шкале;
  — каждую тему оценивают несколько экспертов.
- Метод интрузий (intrusion):
  — в список топовых слов внедряется лишнее слово;
  — измеряется доля ошибок экспертов его при определении

Нужна автоматически вычисляемая мера интерпретируемости, коррелирующая с экспертными оценками.

Ею оказалась *когерентность* (согласованность, coherence).

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Эксперимент. Связь когерентности и интерпретируемости

Измерялась ранговая корреляция Спирмена между 15 метрикам и экспертными оценками интерпретируемости.

PMI — лучшая метрика.

Gold-standard — средняя корреляция Спирмена между оценками разных экспертов.

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	−0.20	0.19
	LCH	−0.31	−0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	−0.08	0.27
	WUP	0.41	0.26
Wikipedia	RACO	0.62	0.69
	MIW	0.68	0.70
	DOCSIM	0.59	0.60
	PMI	**0.74**	**0.77**
Google	TITLES	0.51	
	LOGHITS	−0.19	
Gold-standard	IAA	0.82	0.78

Вывод: когерентность близка к «золотому стандарту».

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Когерентность как внутренняя мера интерпретируемости

Когерентность (согласованность) темы t по k топовым словам:

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

где w_i — i-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),

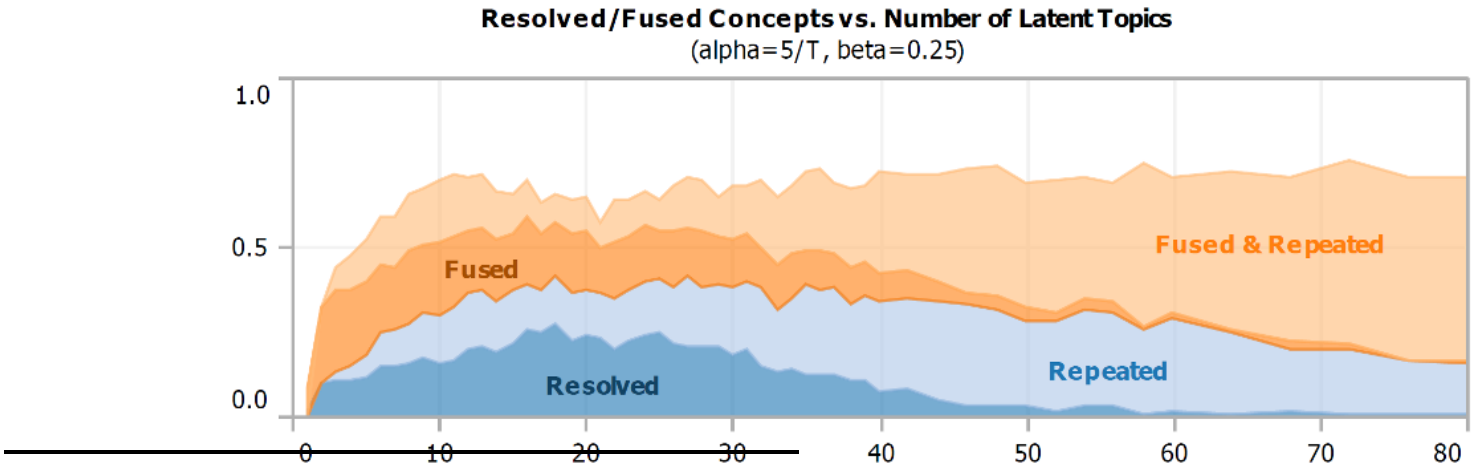
N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов),

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.

Внешние критерии качества

- Полнота и точность тематического поиска
- Качество классификации документов
- Качество сегментации или суммаризации
- Экспертное оценивание тем *методом интрузий*
- Точность соответствия тем заданным *концептам*
  (число ненайденных и расщеплённых тем и концептов)



Chuang J., Gupta S., Manning C., Heer J. Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment. ICML-2013.

Методика оценивания качества разведочного поиска

Поисковый запрос

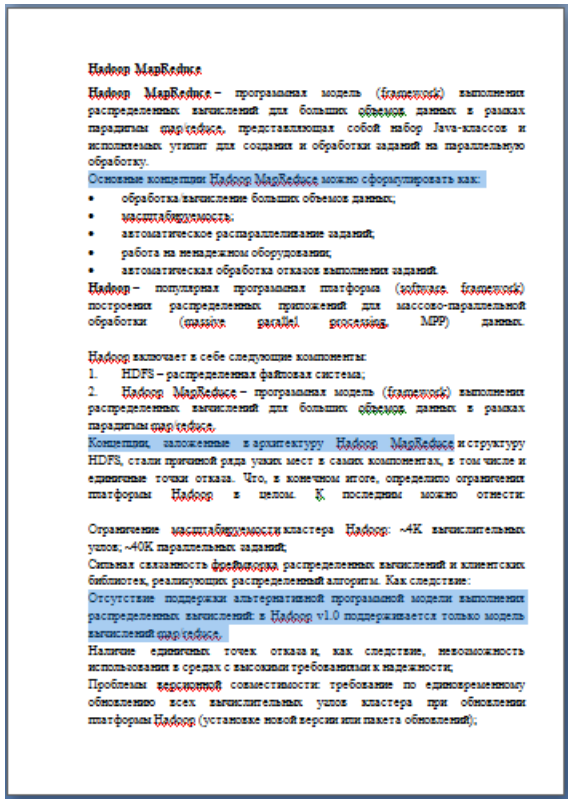
набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

документы d с распределением $p(t|d)$, близким к распределению $p(t|q)$ запроса

Два задания асессорам

- найти как можно больше статей, пользуясь любыми средствами поиска (и засечь время)
- оценить релевантность поисковой выдачи на том же запросе



Hadoop MapReduce

Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce, представляющая собой набор Java-классов и исполняемых утилит для создания и обработки заданий на параллельную обработку.

Основные концепции Hadoop MapReduce можно сформулировать как:

- обработка/вычисление больших объемов данных;
- масштабируемость;
- автоматическое распараллеливание заданий;
- работа на ненадежном оборудовании;
- автоматическая обработка отказов выполнения заданий.

Hadoop – популярная программная платформа (software framework) построения распределенных приложений для массово-параллельной обработки (massive parallel processing, MPP) данных.

Hadoop включает в себя следующие компоненты:

1. HDFS – распределенная файловая система;
2. Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce.

Концепции, заложенные в архитектуру Hadoop MapReduce и структуру HDFS, стали причиной ряда узких мест в самих компонентах, в том числе и единичные точки отказа. Что, в конечном итоге, определило ограничения платформы Hadoop в целом. К последним можно отнести:

Ограничение масштабируемости кластера Hadoop: ~4K вычислительных узлов; ~40K параллельных заданий;

Сильная связанность фреймворка распределенных вычислений и клиентских библиотек, реализующих распределенный алгоритм. Как следствие:

Отсутствие поддержки альтернативной программной модели выполнения распределенных вычислений: в Hadoop v1.0 поддерживается только модель вычислений map/reduce.

Наличие единичных точек отказа и, как следствие, невозможность использования в средах с высокими требованиями к надежности;

Проблемы версионной совместимости: требование по единовременному обновлению всех вычислительных узлов кластера при обновлении платформы Hadoop (установке новой версии или пакета обновлений);

Пример запроса для разведочного поиска

Две коллекции новостей про технологии

Habrahabr.ru

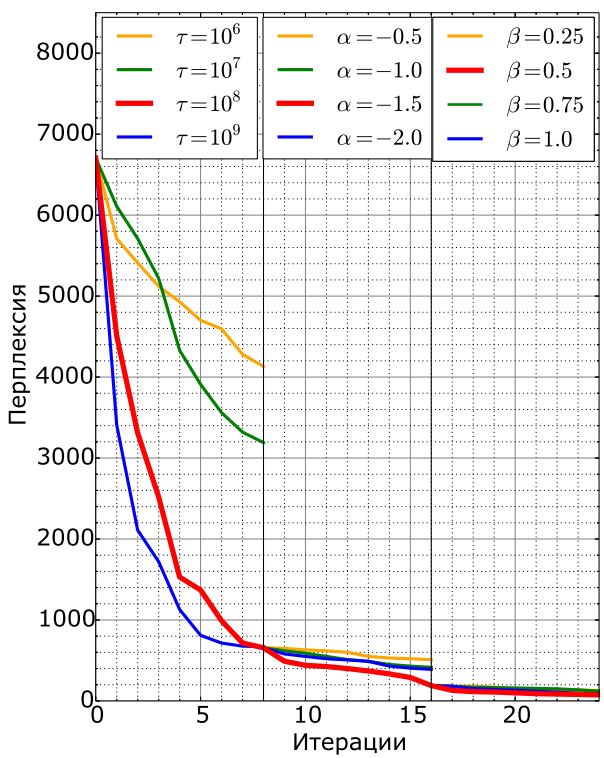
175 143 статей на русском
10 552 слов (униграмм)
742 000 биграмм
524 авторов статей
10 000 авторов комментариев
2546 тегов
123 хаба (категории)

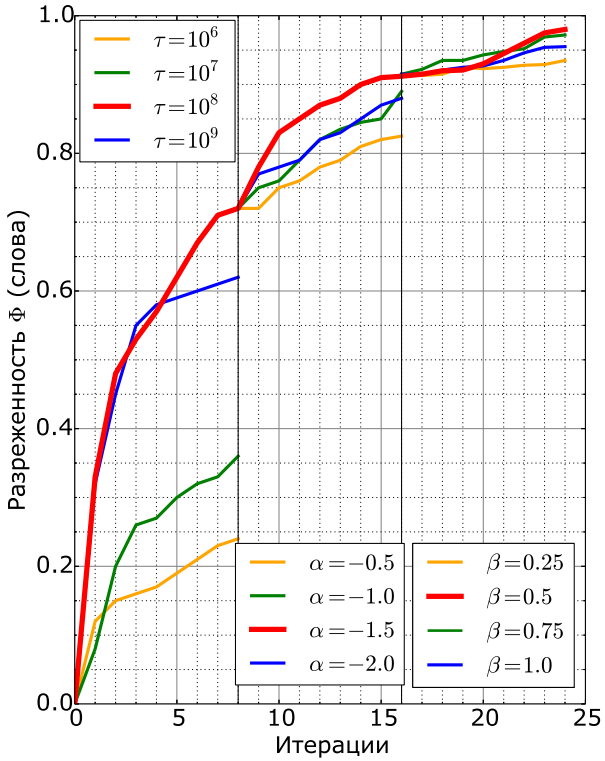
TechCrunch.com

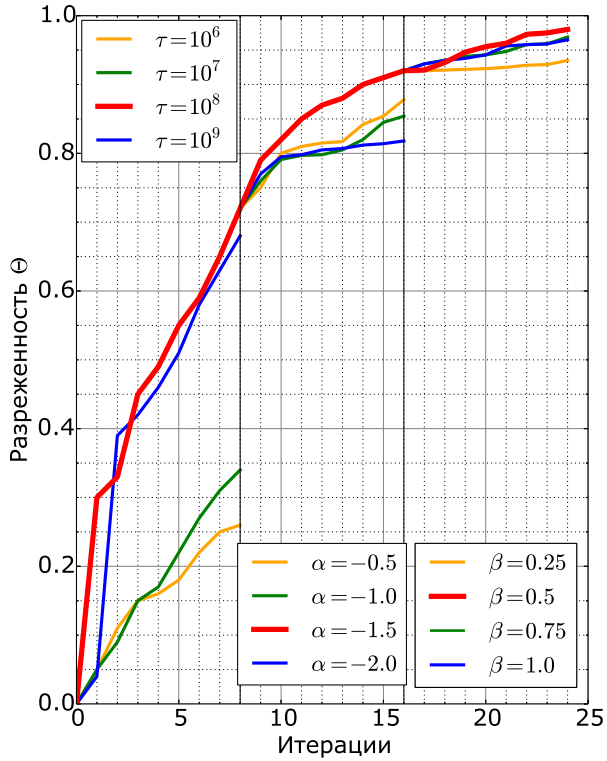
759 324 статей на английском
11 523 слов (униграмм)
1.2 млн. биграмм
605 авторов
184 категорий

Последовательный подбор коэффициентов регуляризации

- декоррелирование распределений терминов в темах (τ),
- разреживание распределений тем в документах (α),
- сглаживание распределений терминов в темах (β).

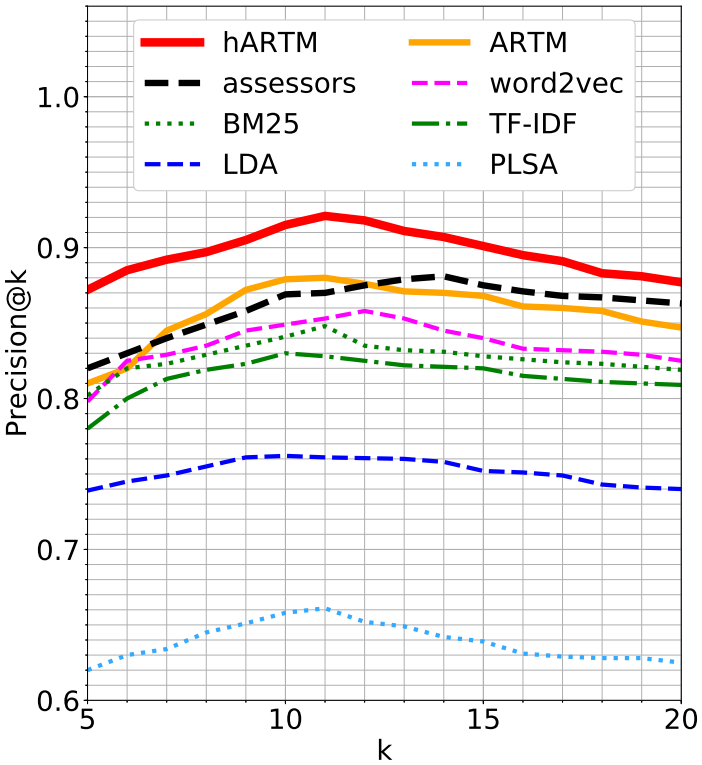


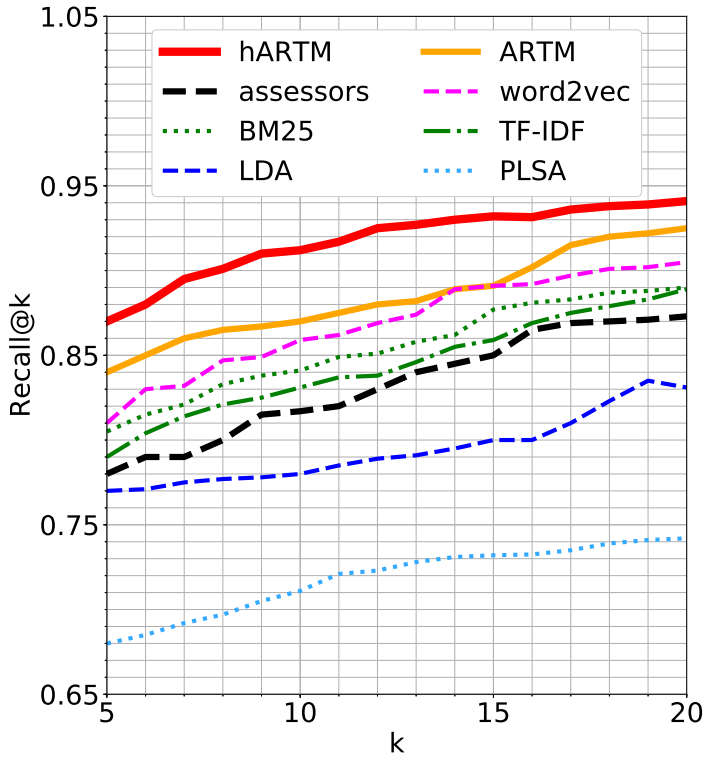




Сравнение качества поиска с асессорами и простыми моделями

Точность и полнота по первым k позициям поисковой выдачи (коллекция Habrahabr.ru)

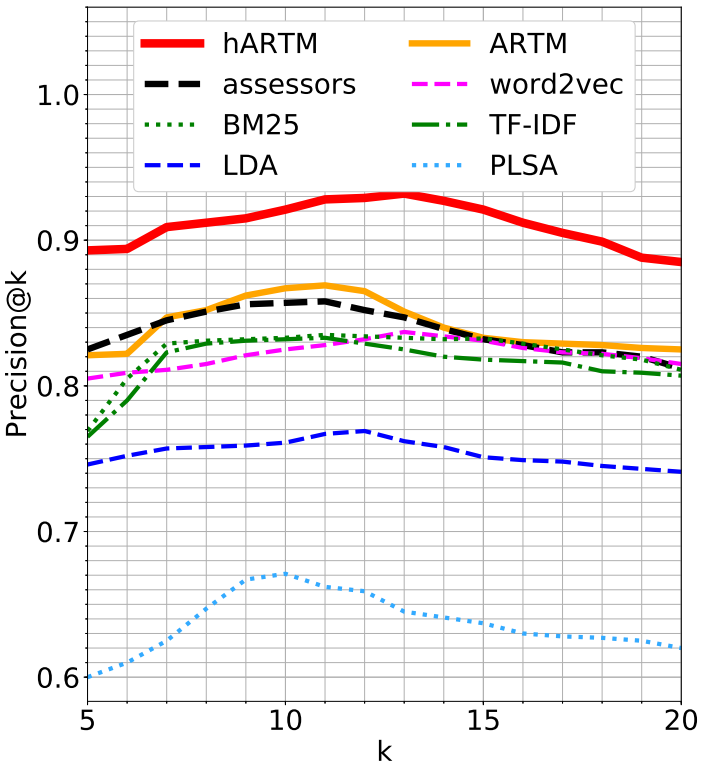


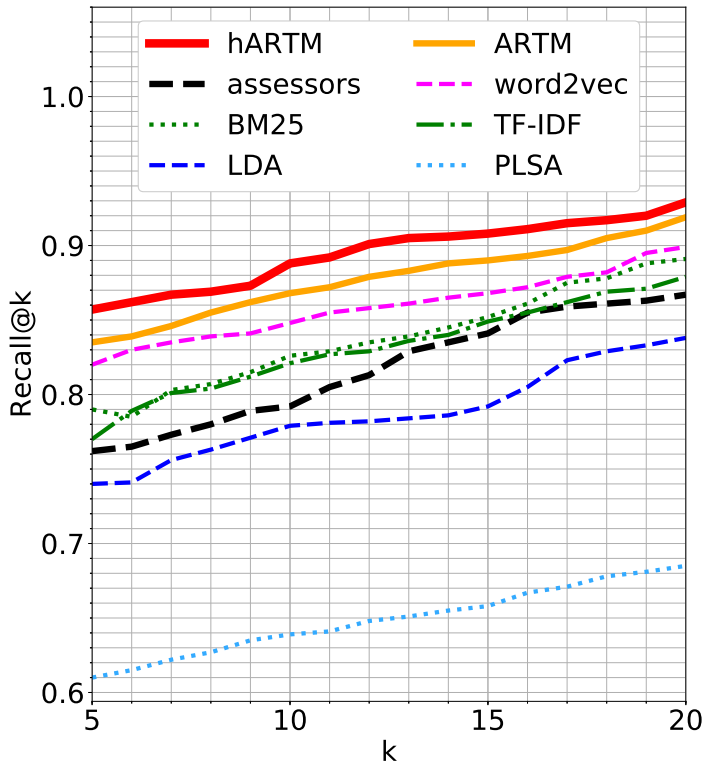


A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Сравнение качества поиска с асессорами и простыми моделями

Точность и полнота по первым k позициям поисковой выдачи (коллекция TechCrunch.com)





A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Влияние числа тем на качество поиска

Коллекция Habrahabr.ru

Используем 3 регуляризатора, 5 модальностей, меняем $|T|$

	acессоры	100	150	**200**	250	400
Prec@5	0.821	0.662	0.721	**0.810**	0.761	0.693
Prec@10	0.869	0.761	0.812	**0.879**	0.825	0.673
Prec@15	0.875	0.733	0.795	**0.868**	0.791	0.651
Prec@20	0.863	0.724	0.795	**0.847**	0.792	0.642
Recall@5	0.780	0.732	0.807	**0.840**	0.821	0.721
Recall@10	0.817	0.771	0.843	**0.870**	0.851	0.751
Recall@15	0.850	0.824	**0.895**	0.891	0.871	0.773
Recall@20	0.873	0.857	0.905	**0.925**	0.892	0.771

- Наилучшее качество поиска — при 200 темах

Влияние числа тем на качество поиска

Коллекция TechCrunch.com

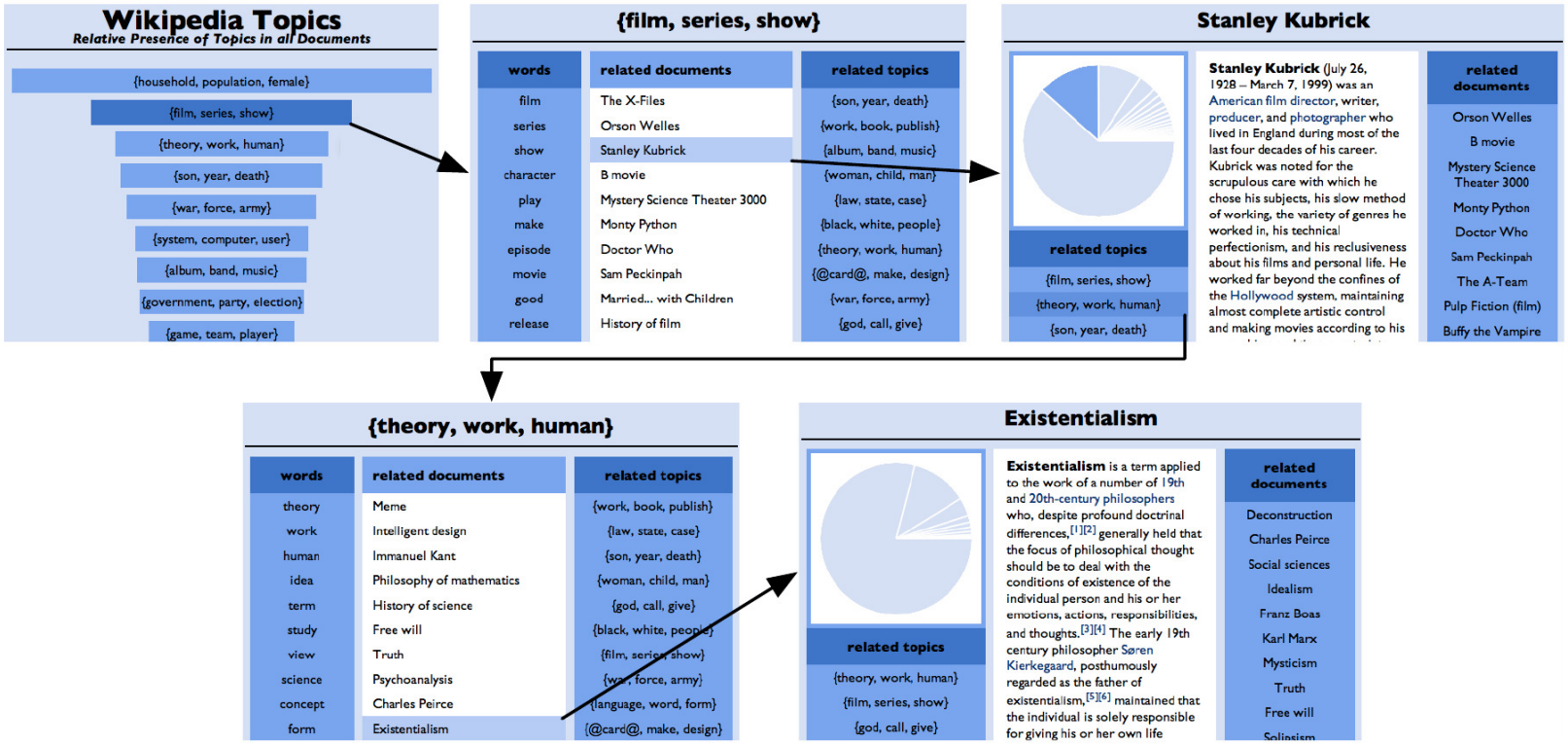
Используем 3 регуляризатора, 4 модальности, меняем $|T|$

	асессоры	350	400	450	**475**	500
Prec@5	0.822	0.653	0.725	0.752	**0.819**	0.777
Prec@10	0.851	0.663	0.732	0.762	**0.867**	0.811
Prec@15	0.835	0.682	0.743	0.787	**0.833**	0.793
Prec@20	0.813	0.650	0.743	0.773	**0.825**	0.793
Recall@5	0.762	0.731	0.762	0.793	**0.835**	0.817
Recall@10	0.792	0.763	0.793	0.812	**0.868**	0.855
Recall@15	0.835	0.782	0.807	0.855	**0.890**	0.882
Recall@20	0.867	0.792	0.823	0.862	**0.919**	0.903

- Наилучшее качество поиска — при 475 темах

Система TMVE — Topic Model Visualization Engine

Тематический навигатор с веб-интерфейсом:

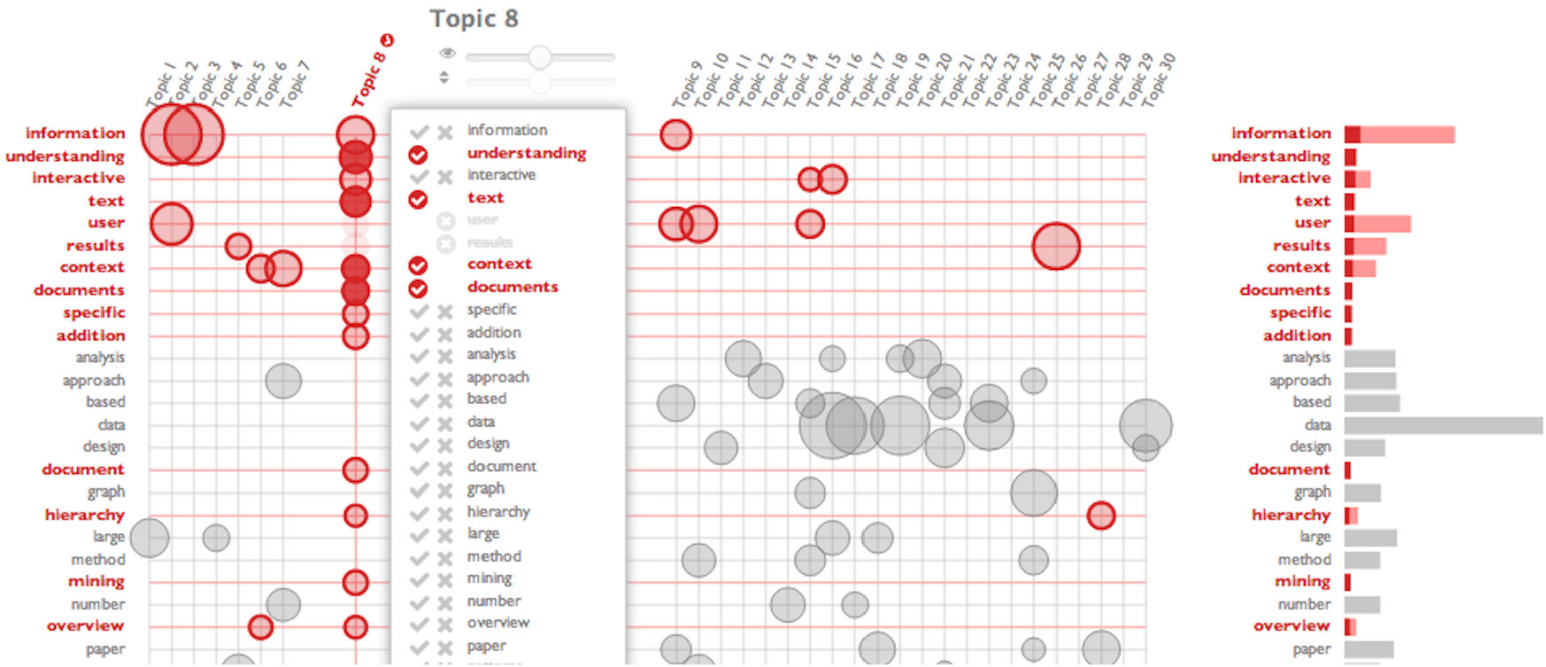


https://github.com/ajbc/tmv

Chaney A., Blei D. Visualizing Topic Models // Frontiers of computer science in China, 2012. — 55(4), pp. 77–84.

Система Termite

Интерактивная визуализация матрицы Φ и сравнение тем:

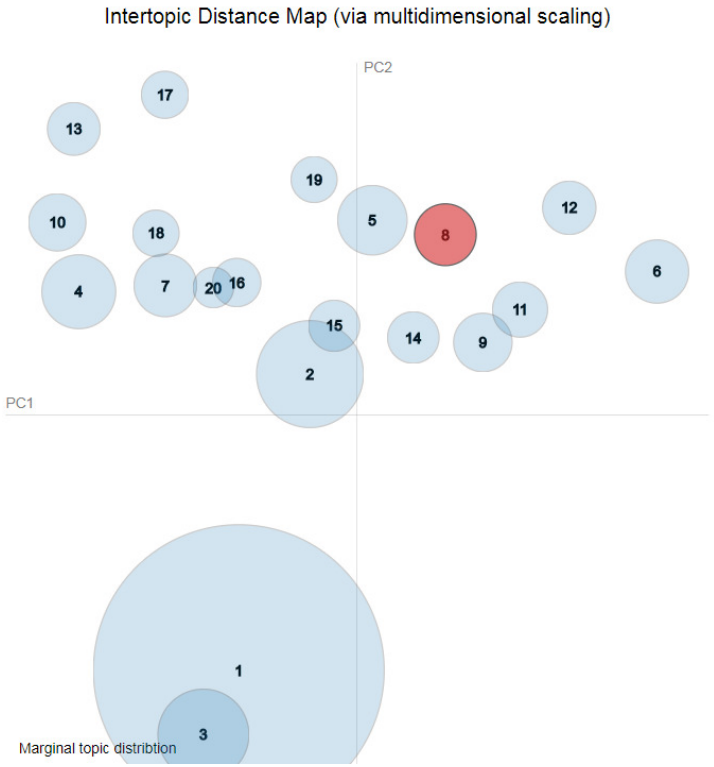


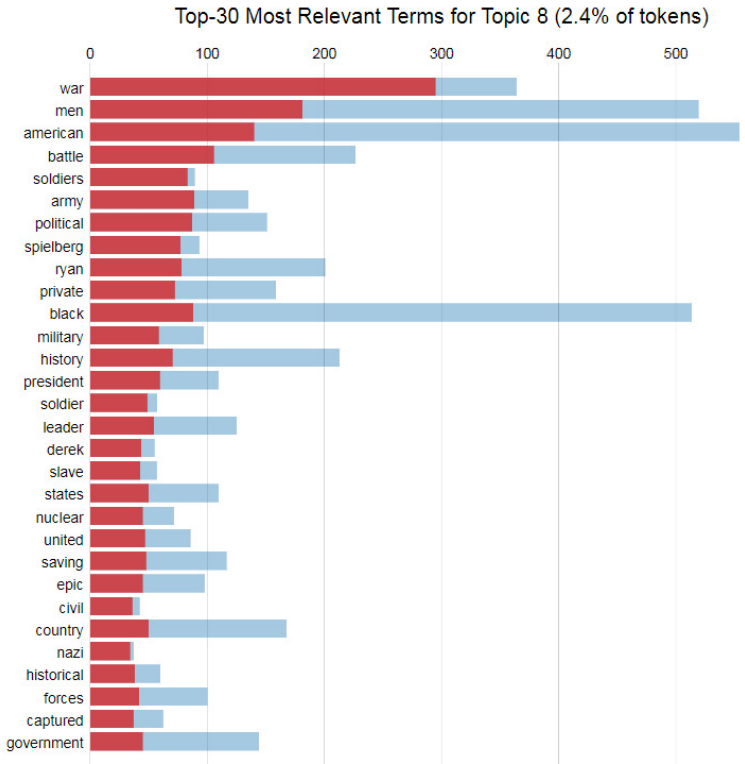
https://github.com/uwdata/termite-visualizations

Chuang J., Manning C., Heer J. Termite: Visualization Techniques for Assessing Textual Topic Models. IWCAVI 2012.

Система LDAvis

Карта сходства тем и сравнение $p(w|t)$ с $p(w)$:

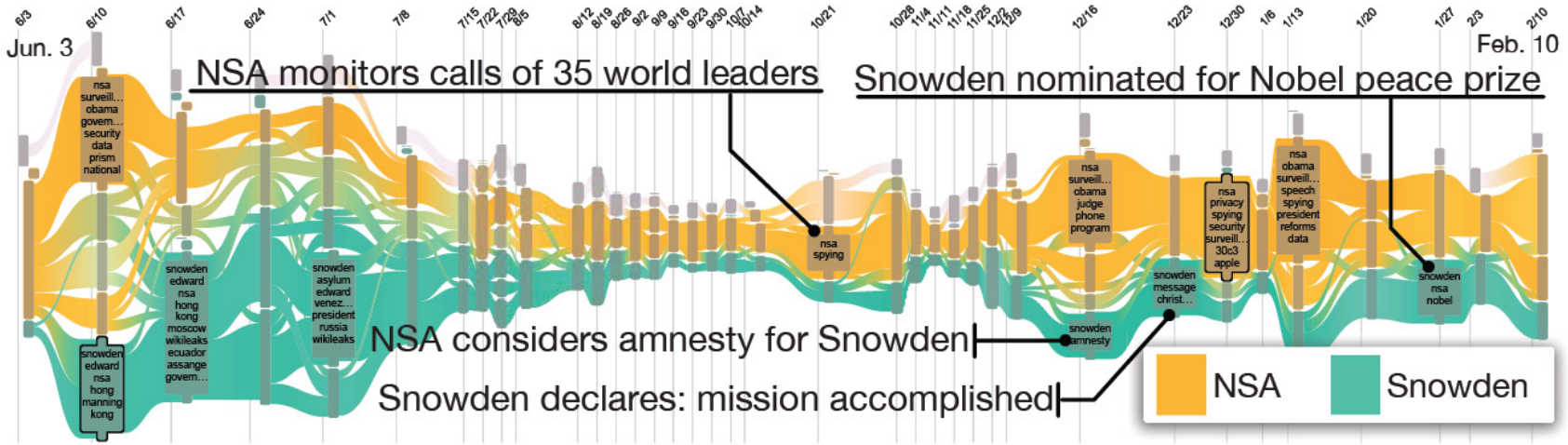




https://github.com/cpsievert/LDAvis

C.Sievert, K.Shirley. LDAvis: A method for visualizing and interpreting topics. 2014.

Динамика тем: эволюция предметной области



Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- генерирует отчёт.

Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. 2014.

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
- Аддитивная регуляризация позволяет комбинировать модели и строить модели с заданными свойствами
- В отличие от классических задач машинного обучения, регуляризаторы весьма разнообразны
- На практике важны внешние критерии качества моделей

Литература

- *К.В.Воронцов.* Обзор вероятностных тематических моделей. 2018. – NEW!
  http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf
- *К.В.Воронцов.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
- *K.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.
- *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.
- *N.Chirkova, K.Vorontsov.* Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.
- *A.Ianina, K.Vorontsov.* Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.
- *A.Potapenko, A.Popov, K.Vorontsov.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.
- *V.Alekseev, V.Bulatov, K.Vorontsov.* Intra-Text Coherence as a Measure of Topic Models Interpretability. Dialogue, 2018.
- *A.Belyy, M.Seleznova, A.Sholokhov, K.Vorontsov.* Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue, 2018.
- *N.Skachkov, K.Vorontsov.* Improving topic models with segmental structure of texts. Dialogue, 2018.