Тематическое моделирование (часть 1)

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

ШАД Яндекс • 27 октября 2015

Содержание

Что такое «тема» в коллекции текстовых документов?

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.

Более формально,

- *тема* — условное распределение на множестве терминов, $p(w|t)$ — вероятность термина w в теме t;
- *тематический профиль* документа — условное распределение $p(t|d)$ — вероятность темы t в документе d.

Когда автор писал термин w в документе d, он думал о теме t, и мы хотели бы выявить, о какой именно.

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

Цели и приложения тематического моделирования

- Выявить скрытую тематическую структуру коллекции текстов
- Найти сжатое описание семантики каждого документа

Приложения:

- Категоризация, классификация, аннотирование, суммаризация, сегментация текстовых документов
- Разведочный информационный поиск (exploratory search)
- Аннотирование изображений, видео, музыки
- Анализ и агрегирование новостных потоков
- Поиск трендов, фронта исследований (research front)
- Поиск экспертов, рецензентов, подрядчиков (expert search)
- Рекомендательные системы
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов

Основные предположения

- Порядок слов в документе не важен (bag of words)
- Порядок документов в коллекции не важен (bag of docs)
- Каждое слово в документе связано с некоторой темой $t \in T$
- $D \times W \times T$ — дискретное вероятностное пространство
- 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: $p(w|d, t) = p(w|t)$

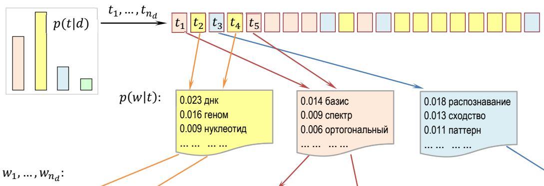
Предварительная обработка текстов:

- Лемматизация (русский) или стемминг (английский)
- Выделение терминов (term extraction)
- Выделение именованных сущностей (named entities)
- Удаление стоп-слов и слишком редких слов

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: W — словарь терминов
D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$
n_{dw} — сколько раз термин w встретился в документе d
n_d — длина документа d

Найти: параметры модели $\frac{n_{dw}}{n_d} \approx p(w|d) = \sum\limits_{t \in T} \phi_{wt}\theta_{td}$:
$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Эта задача стохастического матричного разложения является *некорректно поставленной*, т. к. её решение не единственно:

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

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

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

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

эквивалентна максимизации функционала

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

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

$$\begin{array}{ll} \text{E-шаг:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\[2ex] \phi_{wt} = \operatorname*{norm}_{w\in W}\Bigl(\sum\limits_{d\in D} n_{dw}p_{tdw}\Bigr) \\[2ex] \theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum\limits_{w\in d} n_{dw}p_{tdw}\Bigr) \end{array}\right. \\ \text{M-шаг:} & \end{array}$$

где $\operatorname*{norm}_{t\in T} x_t = \frac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — операция нормировки вектора.

EM-алгоритм. Элементарная интерпретация

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

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

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

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

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

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

LDA — Latent Dirichlet Allocation [Blei 2003]

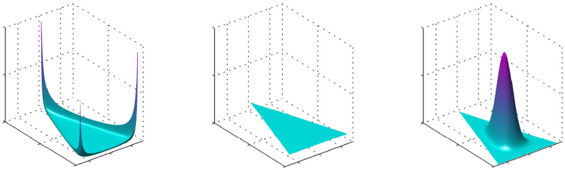
Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})_{w\in W}$ и $\theta_d = (\theta_{td})_{t\in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

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

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

Пример:

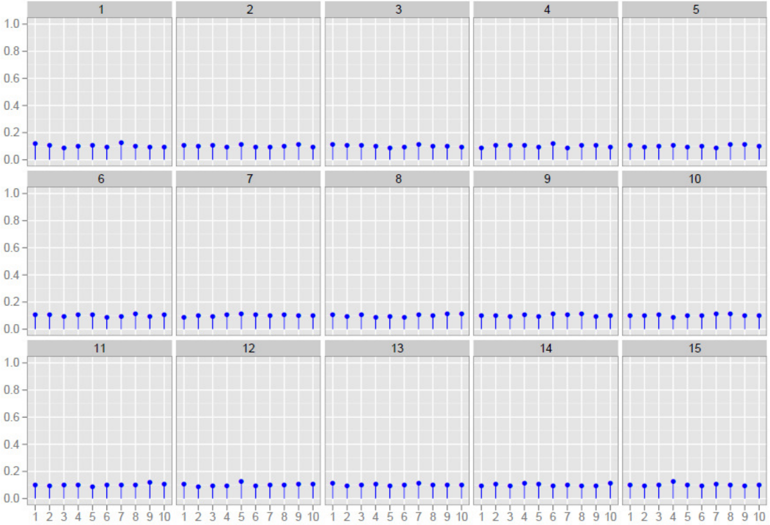
$\mathrm{Dir}(\theta|\alpha)$
$|T| = 3$
$\theta, \alpha \in \mathbb{R}^3$



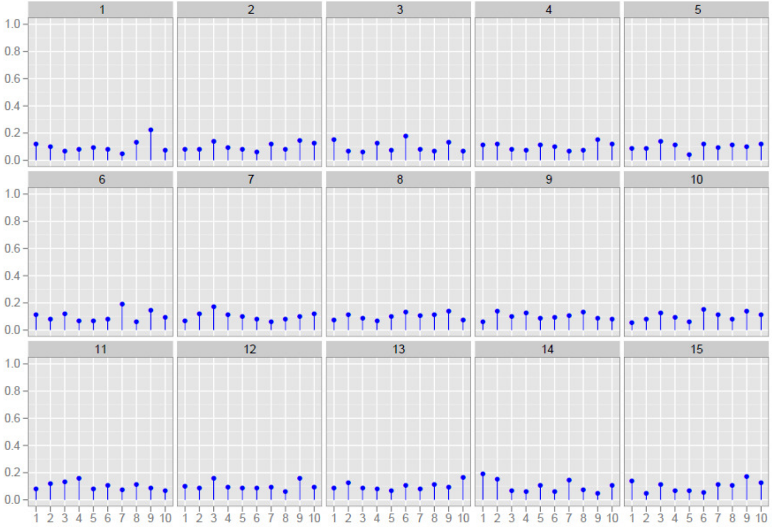
$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$ $\qquad$ $\alpha_1 = \alpha_2 = \alpha_3 = 1$ $\qquad$ $\alpha_1 = \alpha_2 = \alpha_3 = 10$

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

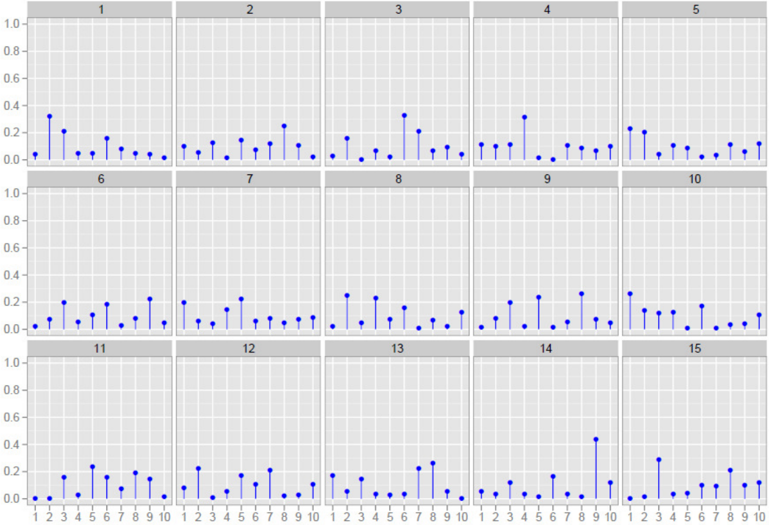
Распределение Дирихле при $\alpha_t \equiv 100$, 10 тем, 15 документов



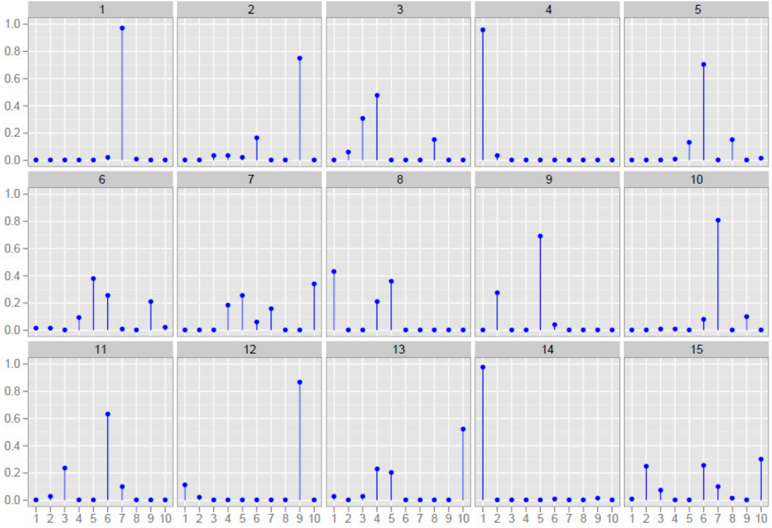
Распределение Дирихле при $\alpha_t \equiv 10$, 10 тем, 15 документов



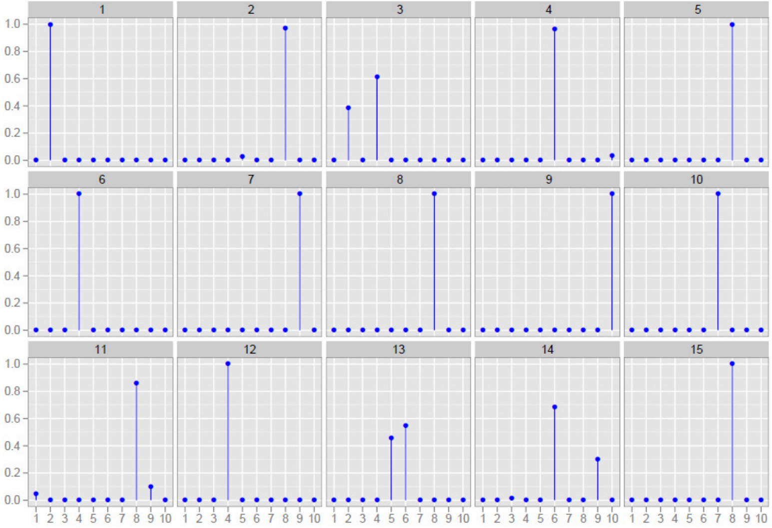
Распределение Дирихле при $\alpha_t \equiv 1$, 10 тем, 15 документов



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



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



Принцип максимума апостериорной вероятности

Совместное правдоподобие данных и модели:

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

Принцип MAP (maximum a posteriori probability)

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

при ограничениях неотрицательности и нормировки

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

Регуляризованный EM-алгоритм

Максимизация апостериорной вероятности ($\tilde\beta_w, \tilde\alpha_t > -1$):

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{rl} \text{E-шаг:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\ \phi_{wt} = \operatorname*{norm}_{w\in W}\Bigl(\sum_{d\in D} n_{dw}p_{tdw} + \tilde\beta_w\Bigr) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum_{w\in d} n_{dw}p_{tdw} + \tilde\alpha_t\Bigr) \end{array} \right. \\ \text{M-шаг:} & \end{array}$$

Неединственность и неустойчивость решения

Эксперимент на модельных данных.

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

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

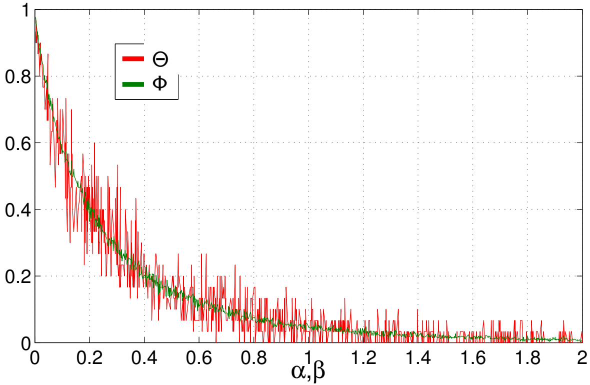
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

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

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

Генерация модельных данных различной разреженности

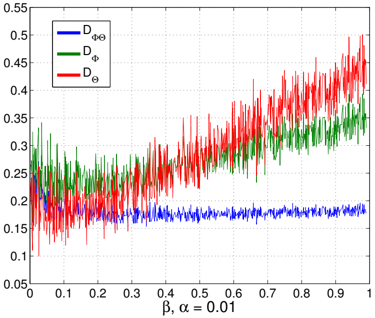
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \text{Dir}(\alpha)$ и $\phi_t^0 \sim \text{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



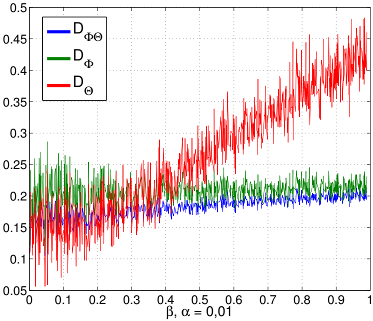
Эксперимент: неустойчивость восстановления Φ, Θ

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

PLSA

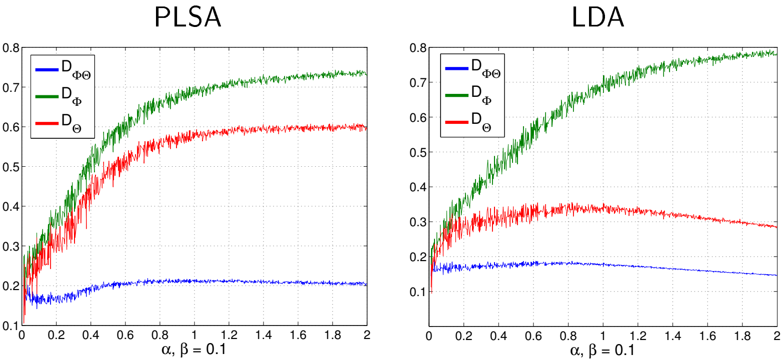


LDA



Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Θ_0 при фиксированном $\beta = 0.1$



Выводы

1. Матрицы Φ, Θ устойчиво восстанавливаются только при сильной разреженности Φ_0, Θ_0 (более 90% нулей)
2. Произведение $\Phi\Theta$ восстанавливается устойчиво, независимо от разреженности исходных Φ_0, Θ_0
3. Задача некорректно поставлена, нет единственности: для любых $S_{T\times T}$ таких, что Φ', Θ' — стохастические,
$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'.$$
4. Распределение Дирихле — слишком слабый регуляризатор

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning. Springer, 2015. Volume 101, Issue 1-3 “Data Analysis and Intelligent Optimization with Applications”, Pp. 303–323.

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация ln правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{rl} \text{E-шаг:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\ \phi_{wt} = \operatorname*{norm}_{w\in W}\Bigl(\sum_{d\in D} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Bigr) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigr) \end{array} \right. \\ \text{M-шаг:} & \end{array}$$

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014. Т. 455., № 3. 268–271.

Комбинирование регуляризованных тематических моделей

Максимизация ln правдоподобия с n регуляризаторами R_i:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi,\Theta},$$

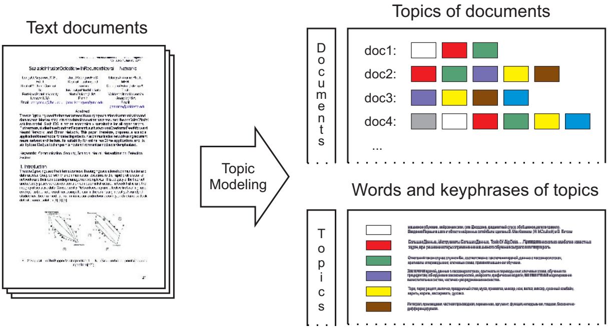
где τ_i — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{aligned}
&\text{E-шаг:} && \left\{\begin{array}{l}
p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\[2ex]
\phi_{wt} = \operatorname*{norm}_{w\in W}\Bigl(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\sum_{i=1}^{n}\tau_i \frac{\partial R_i}{\partial \phi_{wt}}\Bigr) \\[2ex]
\theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\sum_{i=1}^{n}\tau_i \frac{\partial R_i}{\partial \theta_{td}}\Bigr)
\end{array}\right.\\
&\text{M-шаг:} &&
\end{aligned}$$

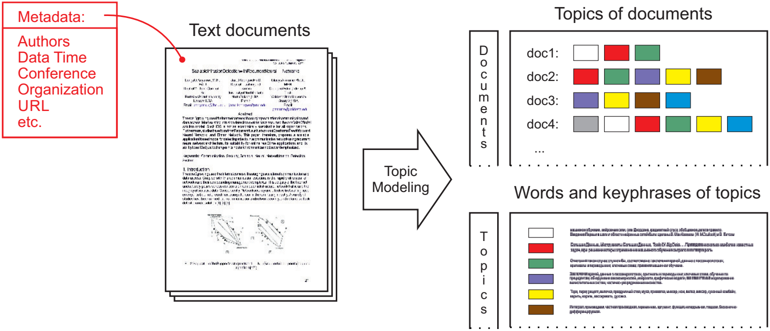
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$,...



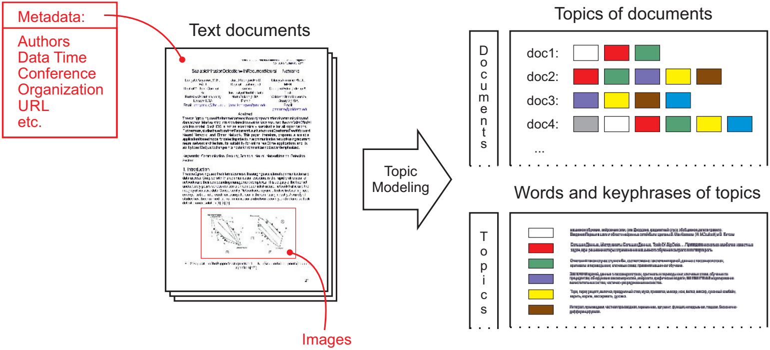
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$,
авторов $p(t|a)$, времени $p(t|a)$,…



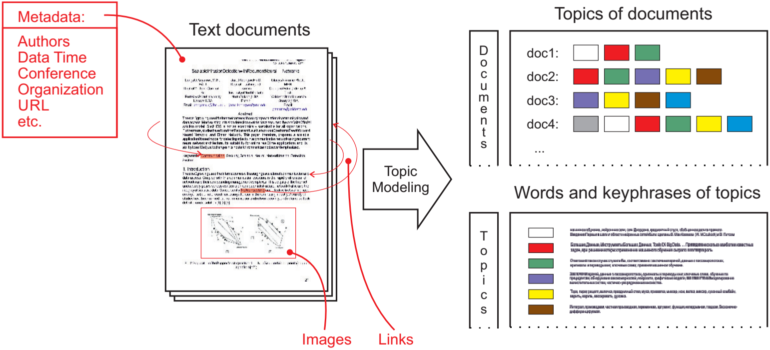
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$,...



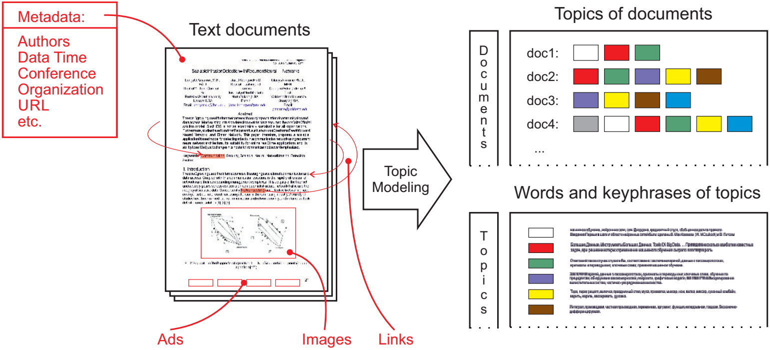
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$, ссылок $p(d'|r)$,...



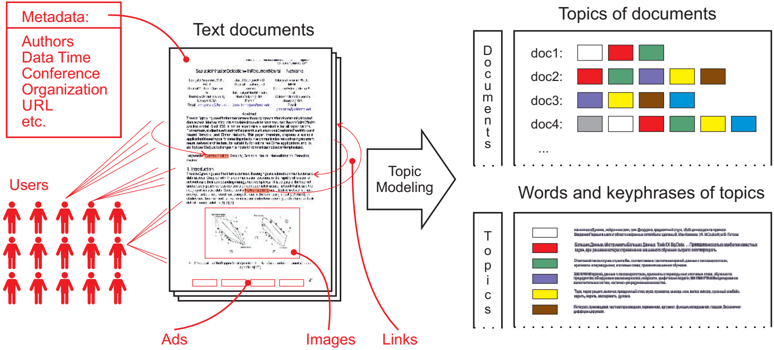
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$, ссылок $p(d'|r)$, баннеров $p(t|b)$,...



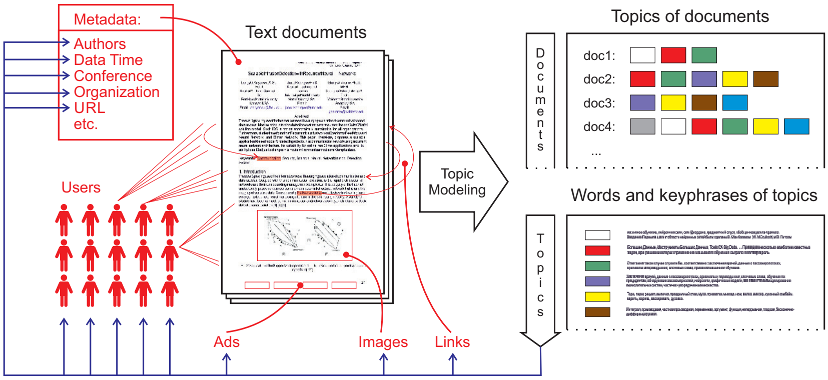
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$, ссылок $p(d'|r)$, баннеров $p(t|b)$, пользователей $p(t|u)$,...



Мультимодальная тематическая модель

Каждая модальность $m \in M$ описывается своим словарём W^m, документы могут содержать *токены* разных модальностей, каждая тема имеет своё распределение $p(w|t)$, $w \in W^m$



Мультимодальная ARTM

W^m — словарь токенов m-й модальности, $m \in M$
$W = W^1 \sqcup \dots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы ln правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{ll} \text{E-шаг:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\ \phi_{wt} = \operatorname*{norm}_{w\in W^m}\Bigl(\sum_{d\in D} \tau_{m(w)} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Bigr) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum_{w\in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigr) \end{array} \right. \\ \text{M-шаг:} & \end{array}$$

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}, $w \in W^m$ (для θ_{td} всё аналогично):
$$\sum_d \tau_m n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \mu_{wt}; \quad \mu_{wt} \geqslant 0; \quad \mu_{wt}\phi_{wt} = 0.$$
2. Умножим обе части равенства на ϕ_{wt} и выделим p_{tdw}:
$$\phi_{wt}\lambda_t = \sum_d \tau_m n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}.$$
3. Альтернатива: либо $\phi_{wt} = 0$ для всех w, либо $\lambda_t > 0$ и
$$\phi_{wt}\lambda_t = \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$
4. Суммируем обе части равенства по $w \in W^m$:
$$\lambda_t = \sum_{w \in W^m} \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$
5. Подставим λ_t из (4) в (3), получим требуемое. ■

Рациональный EM-алгоритм для PLSA

Идея: E-шаг встраивается внутрь M-шага

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

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

Онлайновый параллельный EM-алгоритм для ARTM

Вход: коллекция D, разложенная по пакетам D_b, $b = 1, \dots, B$;
коэффициент дисконтирования $\rho \in (0, 1]$;
Выход: матрица Φ;

инициализировать ϕ_{wt} для всех $w \in W$, $t \in T$;
$n_{wt} := 0$, $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;
для всех пакетов D_b, $b = 1, \dots, B$
- $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \textbf{ProcessBatch}(D_b, \Phi)$;
- **если** пора выполнить синхронизацию, **то**
  - $n_{wt} := \rho n_{wt} + \tilde{n}_{wt}$ для всех $w \in W$, $t \in T$;
  - $\phi_{wt} := \operatorname*{norm}_{w \in W^m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ для всех $w \in W$, $t \in T$;
  - $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;

Hoffman M. D., Blei D. M., Bach F. R. Online learning for latent Dirichlet allocation // NIPS-2010. Pp. 856–864.

Онлайновый параллельный EM-алгоритм для ARTM

ProcessBatch обрабатывает пакет D_b при фиксированной Φ.

Вход: пакет D_b, матрица $\Phi = (\phi_{wt})$;
Выход: матрица $(\tilde{n}_{wt})$;

$\tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;
для всех $d \in D_b$
  инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
  повторять
    $p_{tdw} := \mathop{\mathrm{norm}}\limits_{t \in T} \bigl(\phi_{wt}\theta_{td}\bigr)$ для всех $w \in d,\ t \in T$;
    $\theta_{td} := \mathop{\mathrm{norm}}\limits_{t \in T} \Bigl(\sum\limits_{w \in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \Bigr)$ для всех $t \in T$;
  пока θ_d не сойдётся;
  $\tilde{n}_{wt} := \tilde{n}_{wt} + \tau_{m(w)} n_{dw} p_{tdw}$ для всех $w \in d,\ t \in T$;

ARTM: зоопарк регуляризаторов

- разреживание и декоррелирование предметных тем
- сглаживание фоновых тем общей лексики (LDA)
- энтропийное разреживание для отбора тем
- сглаживание и разреживание тем во времени
- выявление иерархических связей между темами
- многоязычное тематическое моделирование
- выявление внутренней тематической структуры текста
- обучение с учителем для классификации и регрессии
- частичное (semi-supervised) обучение
- и др.

Vorontsov K. V., Potapenko A. A. Tutorial on probabilistic topic modeling: additive regularization for stochastic matrix factorization // Analysis of images, social networks and texts (AIST'2014). Springer CCIS, 2014, Vol. 436, pp. 29-46.

Байесовские тематические модели и ARTM

Методы обучения байесовских тематических моделей

- вариационный вывод (variational inference)
- сэмплирование Гиббса (Gibbs sampling)

Преимущества ARTM:

- байесовские модели представимы в виде регуляризатора
- регуляризаторы не обязаны иметь вероятностный смысл
- регуляризаторы (значит, и модели) легко комбинировать
- стандартизация разработки многофункциональных моделей
- онлайновый параллельный EM-алгоритм
- реализован в проекте с открытым кодом BigARTM

Резюме

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