Тематическое моделирование

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 • 6 октября 2020

Содержание

Задача тематического моделирования

Дано:

- W — конечное множество термов (слов, токенов)
- D — конечное множество текстовых документов
- n_{dw} — частоты слов w в документах d

Найти: вероятностную тематическую модель

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

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

Критерий: максимум логарифма правдоподобия

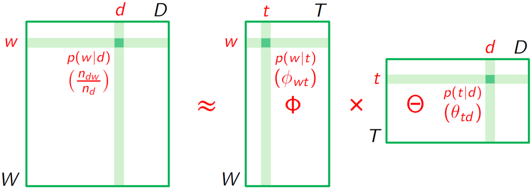
$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

при ограничениях $\phi_{wt} \geqslant 0$, $\sum_w \phi_{wt} = 1$, $\theta_{td} \geqslant 0$, $\sum_t \theta_{td} = 1$

Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

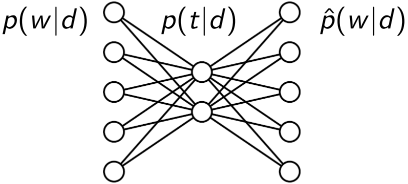
Три интерпретации задачи тематического моделирования

- **Мягкая кластеризация** документов по кластерам-темам
- **Стохастическое матричное разложение:**



- **Автокодировщик** $p_d \xrightarrow{f} \theta_d \xrightarrow{g} \hat{p}_d$

$$p_d \equiv p(w|d)$$
$$\theta_d \equiv p(t|d) = f(p_d, \Phi)$$
$$\hat{p}_d \equiv \hat{p}(w|d) = g(\theta_d, \Phi)$$
$$\sum_d \mathrm{KL}\big(p_d \| \hat{p}_d\big) \to \min_{\Phi}$$

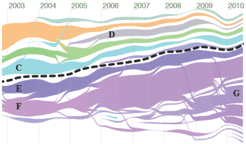


Некоторые приложения тематического моделирования

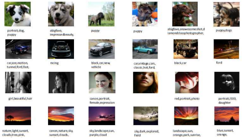
разведочный поиск в электронных библиотеках

поиск тематического контента в соцсетях

детектирование и трекинг новостных сюжетов

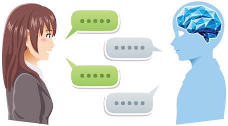


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

управлением диалогом в разговорном интеллекте



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

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

приводит к задаче математического программирования:

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

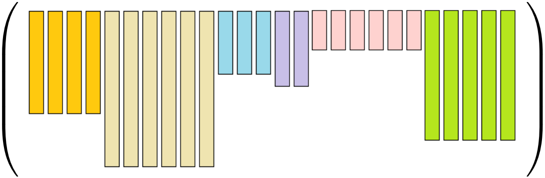
при ограничениях неотрицательности и нормировки

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

Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$\Omega =$



Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Лемма о максимизации функции на единичных симплексах

Операция нормировки вектора: $p_i = \underset{i\in I}{\mathrm{norm}}(x_i) = \dfrac{\max\{x_i, 0\}}{\sum\limits_{k\in I} \max\{x_k, 0\}}$

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω. Тогда векторы ω_j локального экстремума задачи $f(\Omega) \to \max$ удовлетворяют системе уравнений

$$\begin{cases} \omega_{ij} = \underset{i\in I_j}{\mathrm{norm}}\left(\omega_{ij}\dfrac{\partial f}{\partial \omega_{ij}}\right), & \text{если } \exists i\colon \omega_{ij}\dfrac{\partial f}{\partial \omega_{ij}} > 0 \\ \omega_{ij} = \underset{i\in I_j}{\mathrm{norm}}\left(-\omega_{ij}\dfrac{\partial f}{\partial \omega_{ij}}\right), & \text{иначе, если } \exists i\colon \omega_{ij}\dfrac{\partial f}{\partial \omega_{ij}} < 0 \\ \omega_{ij}\dfrac{\partial f}{\partial \omega_{ij}} = 0, & \text{иначе} \end{cases}$$

Замечания к Лемме о максимизации на единичных симплексах

- Лемма применима для построения широкого класса моделей, параметрами которых являются дискретные распределения вероятности (нормированные неотрицательные векторы)
- Численное решение системы — методом простых итераций
- Существование стационарной точки Ω гарантировано
- Первый из трёх случаев является основным:

$$\omega_{ij} := \operatorname*{norm}_{i \in I_j} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right)$$

- Остальные случаи будем считать вырожденными, отбрасывая $\omega_j \equiv 0$ и сокращая размерность модели
- Итерации похожи на градиентную оптимизацию, но учитывают ограничения и не требуют подбора шага η:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}}$$

Напоминания. Условия Каруша–Куна–Таккера

Задача математического программирования:

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

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

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

Доказательство Леммы

Запишем условия Каруша–Куна–Таккера для ω_{ij}:

$$\frac{\partial f}{\partial \omega_{ij}} = \lambda_j - \mu_{ij}; \quad \mu_{ij}\omega_{ij} = 0.$$

Предполагая $\omega_{ij} > 0$, умножим обе части равенства на ω_{ij}:

$$A_{ij} \equiv \omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = \omega_{ij}\lambda_j.$$

Возможны три случая:

1. Если $\lambda_j > 0$, то либо $A_{ij} > 0$, либо $\omega_{ij} = 0$. Тогда $\omega_{ij}\lambda_j = (A_{ij})_+$; $\lambda_j = \sum\limits_i (A_{ij})_+ \ \Rightarrow\ \omega_{ij} = \operatorname*{norm}\limits_i(A_{ij})$.
2. Если $\lambda_j < 0$ и $(\exists i)\ A_{ij} < 0$, то $(\forall i)\ A_{ij} \leqslant 0$. Тогда $\omega_{ij}\lambda_j = -(-A_{ij})_+$; $\lambda_j = -\sum\limits_i (-A_{ij})_+ \ \Rightarrow\ \omega_{ij} = \operatorname*{norm}\limits_i(-A_{ij})$.
3. Иначе $\lambda_j = 0$ и ω_j находится из уравнений $\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = 0$.

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

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

Регуляризация — стандартный приём доопределения решения с помощью дополнительных критериев.

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{rl} \text{Е-шаг:} & \left\{ \begin{array}{l} p_{tdw} \equiv p(t \mid d,w) = \underset{t\in T}{\operatorname{norm}} \bigl(\phi_{wt}\theta_{td}\bigr) \\ \phi_{wt} = \underset{w\in W}{\operatorname{norm}} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t\in T}{\operatorname{norm}} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_{w\in d} n_{dw} p_{tdw} \end{array} \right. \\ \text{М-шаг:} & \end{array}$$

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

Доказательство (по лемме о максимизации на симплексах)

Применим лемму к log-правдоподобию с регуляризатором:

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

$$\phi_{wt} = \operatorname*{norm}_{w\in W}\left(\phi_{wt}\frac{\partial f}{\partial \phi_{wt}}\right) = \operatorname*{norm}_{w\in W}\left(\phi_{wt}\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) =$$

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

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\theta_{td}\frac{\partial f}{\partial \theta_{td}}\right) = \operatorname*{norm}_{t\in T}\left(\theta_{td}\sum_{w\in W} n_{dw}\frac{\phi_{wt}}{p(w|d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) =$$

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

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 термов $w \in W$

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

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.

Два частных случая: модели PLSA и LDA

PLSA: probabilistic latent semantic analysis [Hofmann, 1999] (вероятностный латентный семантический анализ):

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

M-шаг — частотные оценки условных вероятностей:

$$\phi_{wt} = \operatorname*{norm}_{w}\bigl(n_{wt}\bigr), \qquad \theta_{td} = \operatorname*{norm}_{t}\bigl(n_{td}\bigr).$$

LDA: latent Dirichlet allocation (латентное размещение Дирихле):

$$R(\Phi, \Theta) = \sum_{t,w}(\beta_w - 1)\ln\phi_{wt} + \sum_{d,t}(\alpha_t - 1)\ln\theta_{td}.$$

M-шаг — сглаженные частотные оценки с параметрами β_w, α_t:

$$\phi_{wt} = \operatorname*{norm}_{w}\bigl(n_{wt} + \beta_w - 1\bigr), \qquad \theta_{td} = \operatorname*{norm}_{t}\bigl(n_{td} + \alpha_t - 1\bigr).$$

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

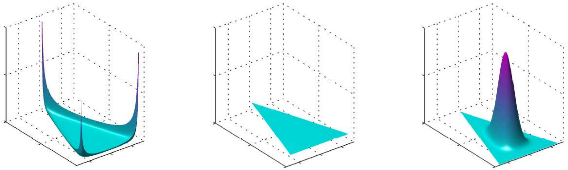
Распределение Дирихле

Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})$ и $\theta_d = (\theta_{td})$ 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. Распределение $\mathrm{Dir}(\theta|\alpha)$ при $|T| = 3$, $\theta, \alpha \in \mathbb{R}^3$



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

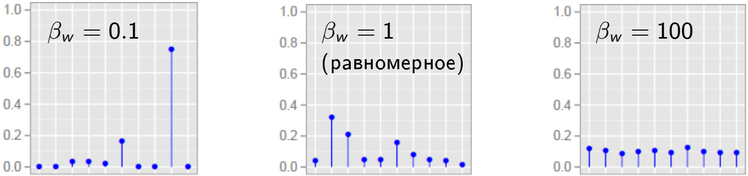
Распределение Дирихле

Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})$ и $\theta_d = (\theta_{td})$ 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. Распределение $\mathrm{Dir}(\phi|\beta)$ при $|T| = 10$, $\phi, \beta \in \mathbb{R}^{10}$:



Максимизация апостериорной вероятности для модели LDA

Совместное правдоподобие данных и модели:

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

Регуляризатор — логарифм априорного распределения:

$$R(\Phi,\Theta) = \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}$$

M-шаг — сглаженные или разреженные частотные оценки:

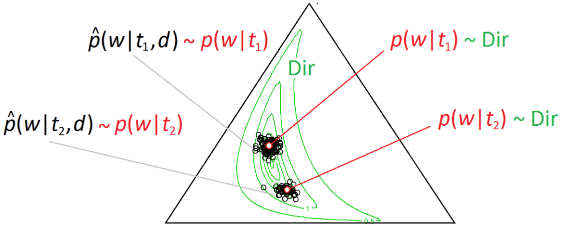
$$\phi_{wt} = \mathop{\mathrm{norm}}_{w} \bigl(n_{wt} + \beta_w - 1\bigr), \qquad \theta_{td} = \mathop{\mathrm{norm}}_{t} \bigl(n_{td} + \alpha_t - 1\bigr).$$

при $\beta_w > 1$, $\alpha_t > 1$ — сглаживание,
при $0 < \beta_w < 1$, $0 < \alpha_t < 1$ — слабое разреживание,
при $\beta_w = 1$, $\alpha_t = 1$ априорное распределение равномерно, PLSA.

Почему именно распределение Дирихле?

- оно способно порождать разреженные векторы,
- имеет параметры, управляющие степенью разреженности,
- описывает кластерные структуры на симплексе (см. рис.),
- математически удобно для байесовского вывода

Распределение $\mathrm{Dir}(\phi|\alpha)$ порождает векторы тем $\phi_t = p(w|t)$, которые порождают мультиномиальные распределения $\hat{p}(w|t,d)$.

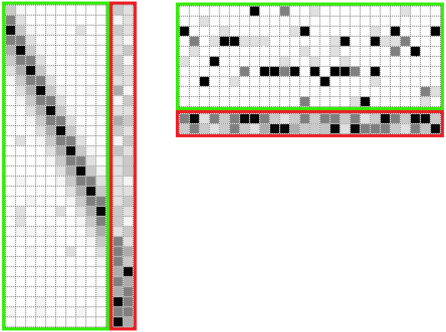


Разделение тем на предметные и фоновые

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

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



Обобщение LDA: 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:

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

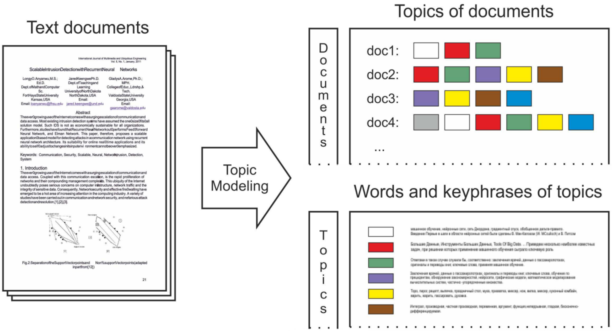
Подставляем, получаем ещё один вариант разреживания — постепенное контрастирование строк матрицы Φ:

$$\phi_{wt} = \operatorname*{norm}_{w}\Bigl(n_{wt} - \tau\phi_{wt} \sum_{s \in T \setminus t} \phi_{ws}\Bigr).$$

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

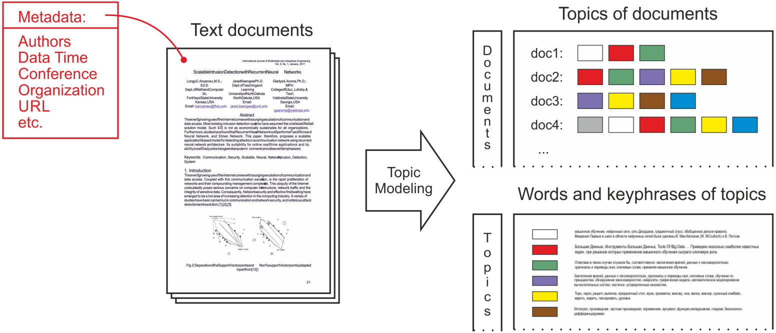
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$,



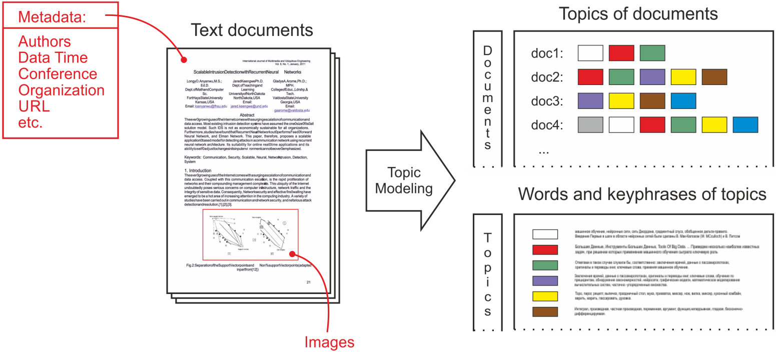
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$, …



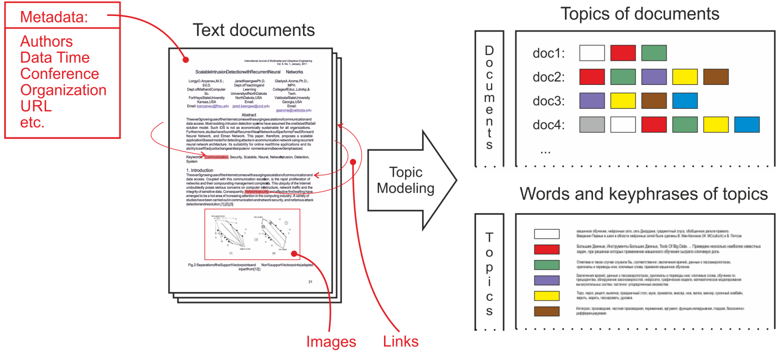
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$, $p(\text{объект}|t)$,



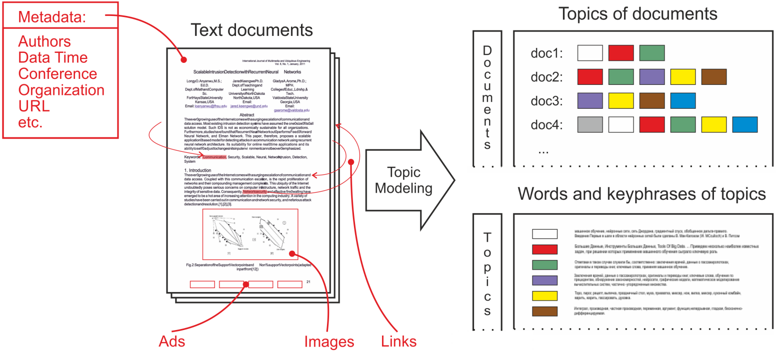
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$, $p(\text{объект}|t)$, $p(\text{ссылка}|t)$,



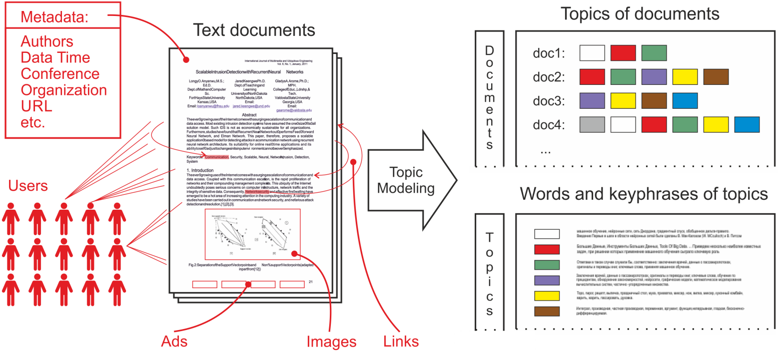
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$, $p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$,



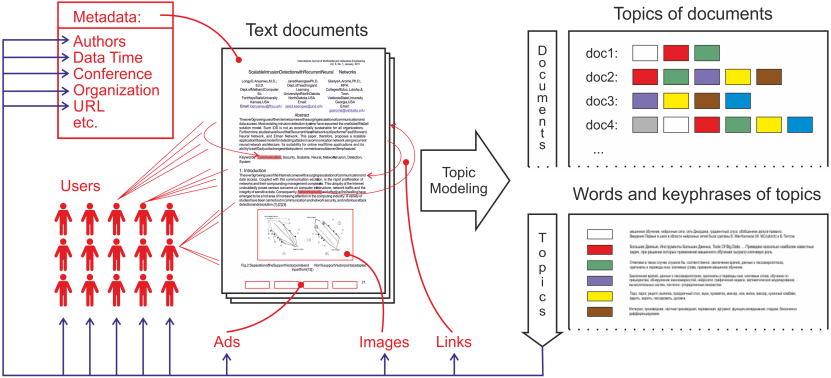
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$, $p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{пользователь}|t)$



Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$, $p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{пользователь}|t)$



Мультимодальная ARTM

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

Максимизация суммы log правдоподобий с регуляризацией:

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

Е-шаг:
М-шаг:

$$\begin{cases} p_{tdw} = \operatorname{norm}_{t \in T} \bigl(\phi_{wt} \theta_{td}\bigr) \\ \phi_{wt} = \operatorname{norm}_{w \in W^m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw} \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw} \end{cases}$$

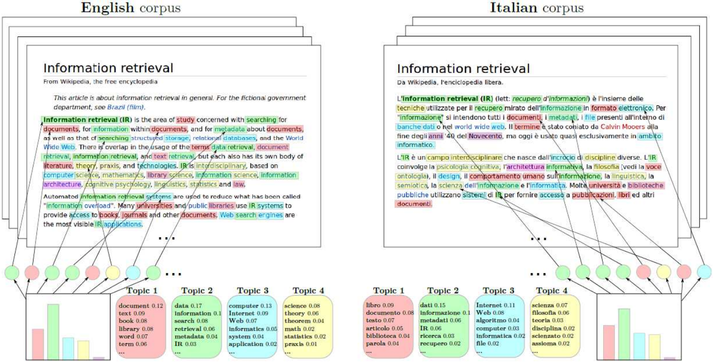
Модальность биграмм улучшает интерпретируемость тем

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.

Многоязычные модели параллельных коллекций



Для построения мультиязычных тем достаточно иметь парные документы, без выравнивания, без двуязычных словарей!

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

Пример. 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.

Тематическая модель классификации (категоризации)

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

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

Правдоподобие вероятностной модели бинарных данных:

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

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

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

Регуляризатор для задач регрессии

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

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

Регуляризатор — среднеквадратичная ошибка (МНК):

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

Подставляем, получаем формулы М-шага:

$$\theta_{td} = \operatorname*{norm}_{t \in T} \Bigl(n_{td} + \tau v_t \theta_{td} \Bigl(y_d - \sum_{t \in T} v_t \theta_{td} \Bigr) \Bigr);$$

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

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.

Проблема коротких текстов

Короткие тексты (short text):

- Twitter и другие микроблоги
- социальные медиа
- заголовки статей и новостных сообщений

Тривиальные подходы:

- считать каждое сообщение отдельным документом
- разреживать $p(t|d)$ вплоть до единственной темы
- объединить сообщения по автору/времени/региону/и т. п.
- объединить посты с комментариями
- дополнить коллекцию длинными текстами (Википедия и др.)

Более интересная идея:

- использовать сочетаемость пар слов в сообщениях

Битермы: модель сочетаемости слов в коротких текстах

Битерм — пара слов, встречающихся рядом:
в одном коротком сообщении / предложении / окне $\pm h$ слов.

Тематическая модель битермов (Biterm Topic Model):

$$p(u, v) = \sum_{t \in T} p(u|t)p(v|t)p(t) = \sum_{t \in T} \phi_{ut}\phi_{vt}\pi_t,$$

где $\phi_{wt} = p(w|t)$, $\pi_t = p(t)$ — параметры модели.

Критерий максимума логарифма правдоподобия:

$$\sum_{u,v} n_{uv} \ln \sum_t \phi_{ut}\phi_{vt}\pi_t \;\to\; \max_{\Phi,\pi},$$

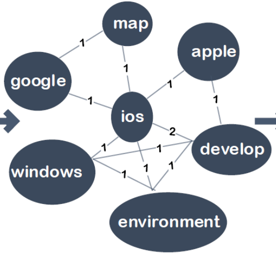
$$\phi_{vt} \geqslant 0; \quad \sum\nolimits_v \phi_{vt} = 1; \qquad \pi_t \geqslant 0; \quad \sum\nolimits_t \pi_t = 1$$

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng. A Biterm Topic Model for Short Texts. WWW 2013.

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 (Word Network) 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} — частота сочетания пары слов (w, u).

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.

Регуляризатор Θ для учёта связей между документами

Цель: улучшить темы, используя ссылки или цитирования (если документы ссылаются друг на друга, то их темы близки):

n_{dc} — число ссылок из d на c.

Максимизируем ковариации тематических векторных представлений связанных документов θ_d, θ_c:

$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \sum_{t \in T} \theta_{td} \theta_{tc} \to \max .$$

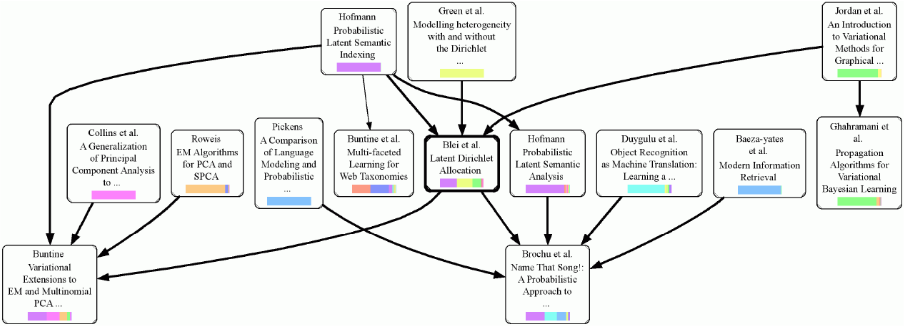
Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} = \operatorname*{norm}_{t} \Bigl(n_{td} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc} \Bigr).$$

Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences. ICML-2007.

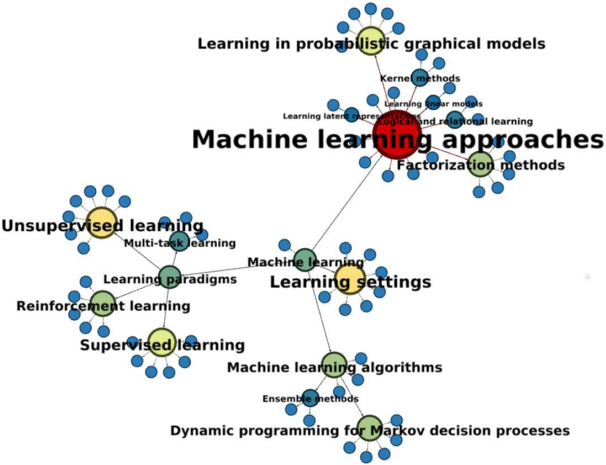
Модели, учитывающие цитирования или гиперссылки

- Учёт ссылок уточняет тематическую модель
- Тематическая модель выявляет влиятельные ссылки



Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences. ICML-2007.

Пример древовидной тематической иерархии



Georgeta Bordea. Domain adaptive extraction of topical hierarchies for Expertise Mining. 2013.

Иерархические тематические модели

- структура иерархии: дерево / многодольный граф
- направление: снизу вверх / сверху вниз / одновременно
- наращивание: повершинное / послойное

Открытые проблемы:

- "Despite recent activity in the field of HPTMs, determining the hierarchical model that best fits a given data set, in terms of the structure and size of the learned hierarchy, still remains a challenging task and an open issue."
- "The evaluation of hierarchical PTMs is also an open issue."

Zavitsanos E., Paliouras G., Vouros G. A. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes. 2011.

Иерархическая тематическая модель: послойное построение

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:

$$\sum_{t\in T} n_{wt} \ln p(w|t) = \sum_{t\in T} n_{wt} \ln \sum_{s\in S} p(w|s)p(s|t) \;\to\; \max_{\Phi,\Psi},$$

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

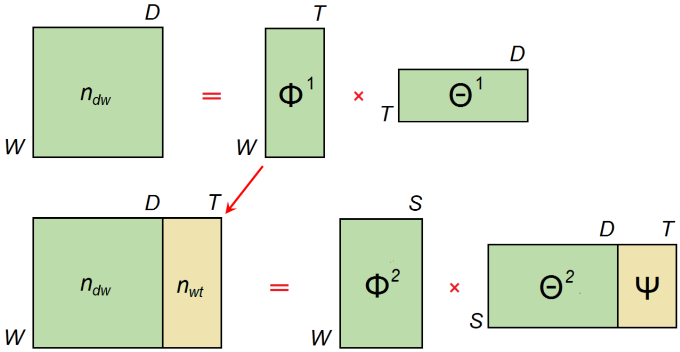
Родительская $\Phi^p \approx \Phi\Psi$, отсюда регуляризатор матрицы Φ:

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

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

Построение второго уровня иерархии с подтемами S

В коллекцию добавляются $|T|$ псевдодокументов родительских тем с частотами термов $\tau n_{wt} = \tau n_t \phi_{wt}$, $t \in T$



Матрица связей тем с подтемами $\Psi = \bigl(p(s|t)\bigr)$ образуется в столбцах матрицы Θ, соответствующих псевдодокументам.

Пример тематической иерархии

Тексты научно-просветительского ресурса Postnauka.ru:
2976 документов, 43196 слов, 1799 тэгов

Chirkova N.A., Vorontsov K.V. Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.

Belyy A.V., Seleznova M.S., Sholokhov A.K., Vorontsov K.V. Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue 2018.

Резюме

- Тематическое моделирование — «мягкая кластеризация», автокодировщик или стохастическое матричное разложение
- Стандартные методы — PLSA и LDA
- Нестандартные — огромное разнообразие регуляризаторов
- Аддитивная регуляризация позволяет комбинировать модели
- Обычно в ТМ используется байесовское обучение. **Почему оно не нужно:** на практике используются не апостериорные распределения, а лишь их точечные оценки
- В ARTM те же модели выводятся намного проще, с помощью общей Леммы о максимизации на симплексах