Вероятностные тематические модели
Лекция 1. Постановка задачи, оптимизация и регуляризация

К. В. Воронцов
vokov@forecsys.ru

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

МФТИ – ФИЦ ИУ РАН • 9 сентября 2021

Вероятностные тематические модели (BTM): о чём этот курс

Probabilistic Topic Modeling (PTM) — область автоматической обработки текстов (Natural Language Processing, NLP)

Курс о том, как

- выявлять тематическую структуру текстовых коллекций
- искать тексты по смыслу, а не по ключевым словам
- применять разнообразные методы NLP совместно с PTM
- ставить задачи анализа текстов в терминах оптимизации
- упрощать математическую теорию, отказываясь от избыточных вероятностных допущений

Пререквизиты (какие знания потребуются)

- теория вероятности (в основном на конечных множествах)
- машинное обучение (базовые понятия и методология)
- линейная алгебра, методы оптимизации (самые азы)
- язык Python

Содержание

Что такое «тема» в коллекции текстовых документов?

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

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

Имея коллекцию текстовых документов, хотим узнать:

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

Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их частоты $p(w|t)$ 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 1. Мультиязычная модель Википедии

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

Пример 2. Биграммная модель научных конференций

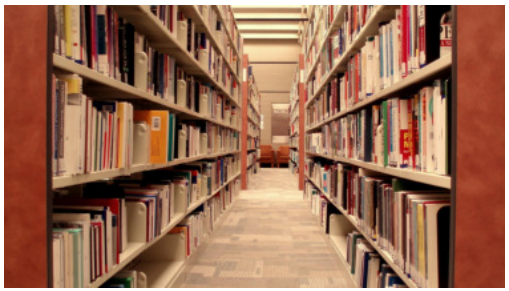
Коллекция 1000 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
униграммы	биграммы	униграммы	биграммы
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

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

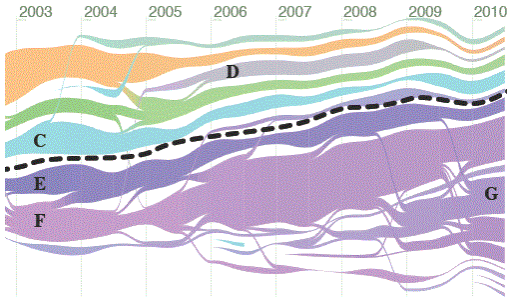
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

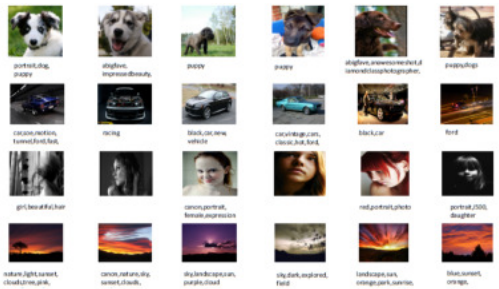


поиск тематического контента в соцсетях

выявление и отслеживание цепочек новостей

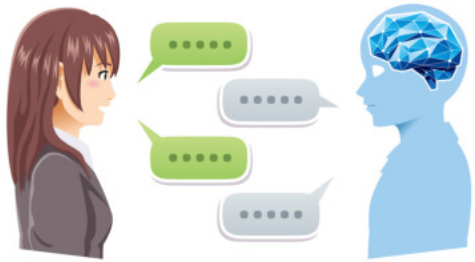


мультимодальный поиск текстов и изображений



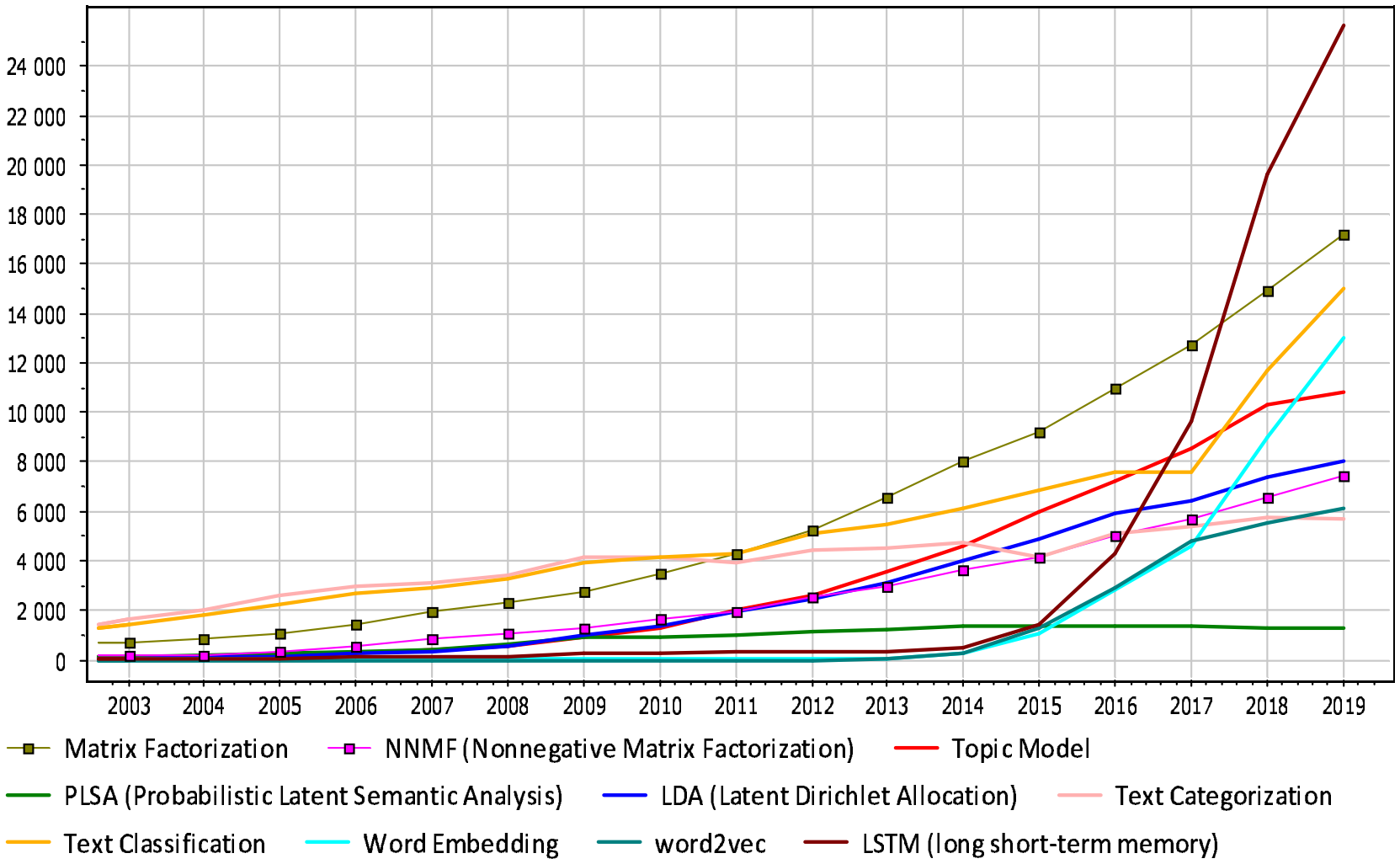
анализ банковских транзакционных данных

управлением диалогом в разговорном интеллекте



Тематическое моделирование и смежные области исследований

Динамика цитирования, по данным Google Scholar:



Пусть

- W — конечное множество *термов* (слов, терминов)
- D — конечное множество текстовых документов
- T — конечное множество тем
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- каждый терм w в документе d связан с некоторой темой 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: $p(w|d,t) = p(w|t)$

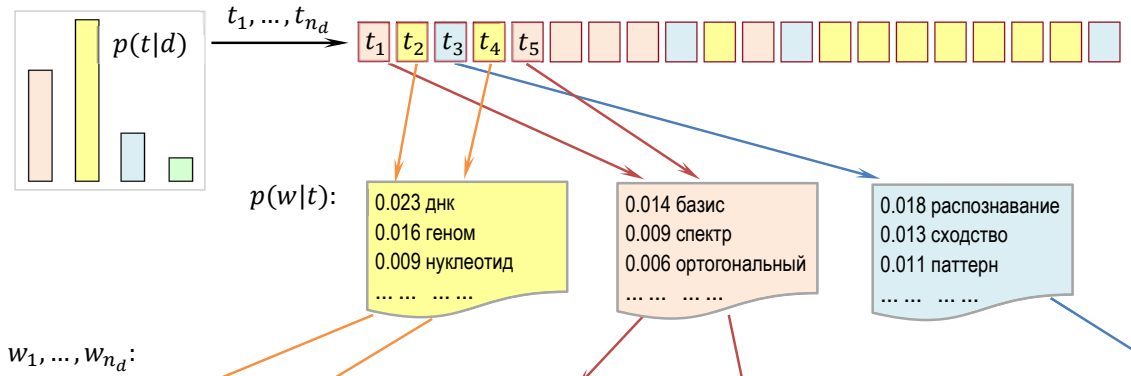
Тематическая модель, по формуле полной вероятности:

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

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

Вероятностная тематическая модель коллекции документов D описывает появление термов 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).

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

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

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

Вход: распределение $p(w|t)$ для каждой темы $t \in T$;
распределение $p(t|d)$ для каждого документа $d \in D$;
Выход: коллекция документов;

для всех $d \in D$
  для всех позиций $i = 1, \ldots, n_d$ в документе d
    сгенерировать тему t_i из $p(t|d)$;
    сгенерировать терм w_i из $p(w|t_i)$;

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

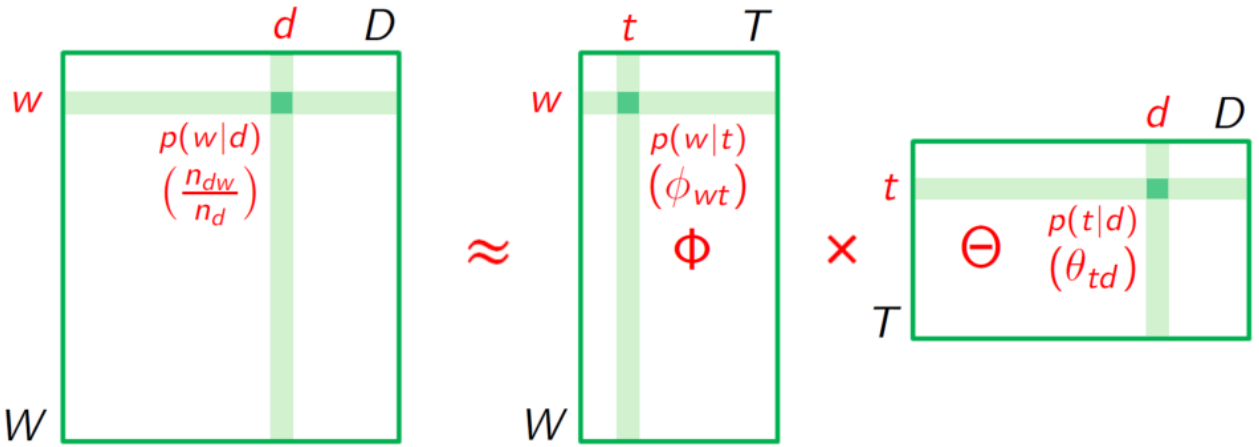
Дано: коллекция текстовых документов

- n_{dw} — частоты термов 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 термов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:



Система обозначений для частот — счётчиков числа термов

Ненаблюдаемые частоты, зависящие от t:

$n_{dwt} = \sum_{i=1}^{n} [d_i = d]\,[w_i = w]\,[t_i = t]$ — частота (d, w, t) в коллекции

$n_{wt} = \sum_d n_{dwt}$ — частота терма w в теме t

$n_{td} = \sum_w n_{dwt}$ — частота термов темы t в документе d

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

Наблюдаемые частоты, не зависящие от t:

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

$n_w = \sum_d n_{dw}$ — частота терма w в коллекции

$n_d = \sum_w n_{dw}$ — длина документа d

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

Элементарное решение обратной задачи

Выразим $p(t|d,w)$ через ϕ_{wt}, θ_{td} по формуле Байеса:

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

Частотные оценки условных вероятностей $\phi_{wt} = \frac{n_{wt}}{n_t}$, $\theta_{td} = \frac{n_{td}}{n_d}$, $p(t|d,w) = \frac{n_{dwt}}{n_{dw}}$ приводят к системе уравнений для ϕ_{wt} и θ_{td}:

$$\begin{cases} n_{dwt} = n_{dw}\dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}, & d \in D,\ w \in W,\ t \in T \\ \phi_{wt} = \dfrac{\sum_d n_{dwt}}{\sum_{d,w} n_{dwt}}, & w \in W,\ t \in T \\ \theta_{td} = \dfrac{\sum_w n_{dwt}}{\sum_{t,w} n_{dwt}}, & d \in D,\ t \in T \end{cases}$$

Численное решение — методом простых итераций

И это работает? Почему? Какую задачу решает эта система?

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

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

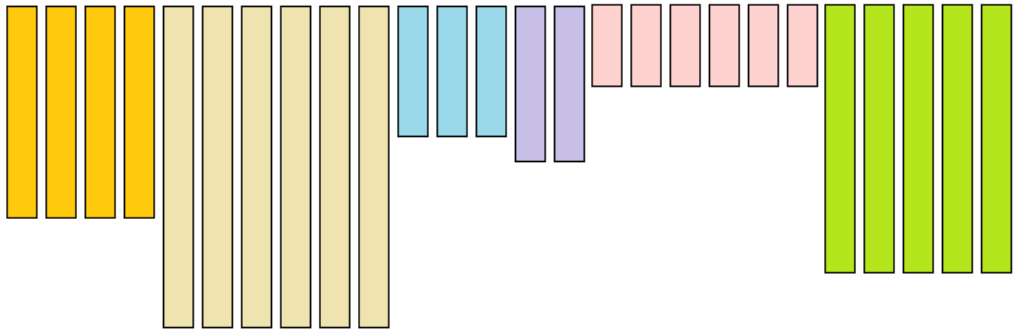
при ограничениях неотрицательности и нормировки

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

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 $|I_j|$:

$$\Omega = \left(\quad \right)$$



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

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

$$\omega_{ij} = \underset{i \in I_j}{\text{norm}}\left(\omega_{ij}\frac{\partial f}{\partial \omega_{ij}}\right), \quad \text{если } \exists i: \omega_{ij}\frac{\partial f}{\partial \omega_{ij}} > 0$$

$$\omega_{ij} = \underset{i \in I_j}{\text{norm}}\left(-\omega_{ij}\frac{\partial f}{\partial \omega_{ij}}\right), \quad \text{иначе, если } \exists i: \omega_{ij}\frac{\partial f}{\partial \omega_{ij}} < 0$$

$$\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = 0, \quad \text{иначе}$$

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:

$$\omega_{ij} := \underset{i \in I_j}{\text{norm}} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right)$$

- В остальных случаях нормирующий знаменатель нулевой; такие векторы будем удалять из модели как вырожденные
- 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:

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

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

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

Доказательство Леммы

Запишем условия Каруша–Куна–Таккера для $\omega_j = (\omega_{ij} : i \in I_j)$:

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

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

■

Теорема о сходимости итерационного процесса

$$\omega_{ij}^{t+1} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^{t}} \right)$$

Теорема. Пусть $f(\Omega)$ — ограниченная сверху, непрерывно дифференцируемая функция, и все Ω^t, начиная с некоторой итерации t^0 обладают свойствами:

- $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t = 0 \to \omega_{ij}^{t+1} = 0$ (сохранение нулей)
- $\exists \varepsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t \notin (0, \varepsilon)$ (отделимость от нуля)
- $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ (невырожденность)

Тогда $f(\Omega^{t+1}) > f(\Omega^t)$ и $\left|\omega_{ij}^{t+1} - \omega_{ij}^t\right| \to 0$ при $t \to \infty$.

Ирхин И. А., Воронцов К. В. Сходимость алгоритма аддитивной регуляризации тематических моделей // Труды Института математики и механики УрО РАН. 2020.

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена по Адамару*, 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

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \operatorname*{norm}_{t \in T}(\phi_{wt}\theta_{td})$$

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

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

Доказательство (по лемме о максимизации на симплексах)

Применим лемму к log-правдоподобию с регуляризатором:

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

Дифференцируя, выделим вспомогательную переменную p_{tdw}:

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

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

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

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

■

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.

Модель вероятностного латентного семантического анализа

PLSA — Probabilistic Latent Semantic Analysis:

- $R(\Phi, \Theta) = 0$ — нет никакой регуляризации.

Получаем то самое «элементарное решение обратной задачи».

EM-алгоритм: метод простой итерации для системы уравнений

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

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Модель латентного размещения Дирихле

LDA — latent Dirichlet allocation:

- распределения ϕ_t близки к заданному распределению β;
- распределения θ_d близки к заданному распределению α.

Логарифм правдоподобия простой вероятностной модели:

$$R(\Phi, \Theta) = \sum_t \sum_w \beta_w \ln \phi_{wt} + \sum_d \sum_t \alpha_t \ln \theta_{td}.$$

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

$$\phi_{wt} = \underset{w}{\text{norm}}\big(n_{wt} + \beta_w\big), \qquad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td} + \alpha_t\big).$$

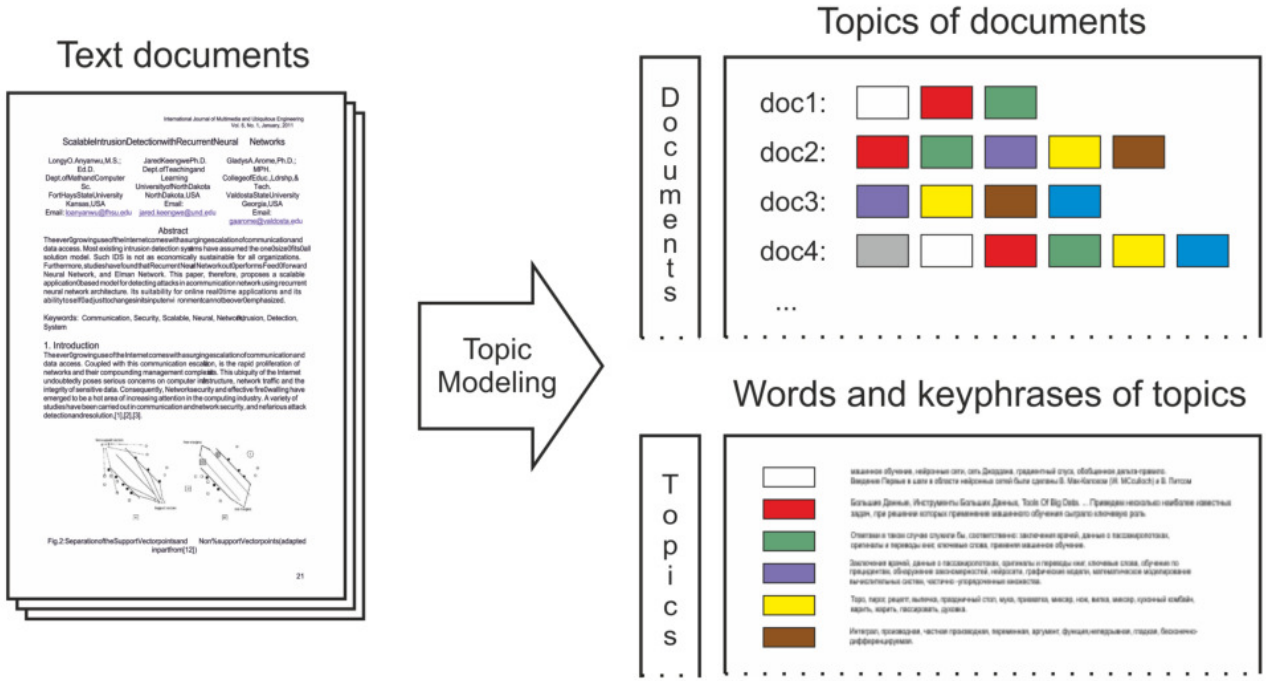
При $\beta_w < 0$ или $\alpha_t < 0$ — *разреживание* вместо сглаживания.

Причём здесь распределение Дирихле? — в следующих лекциях...

Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

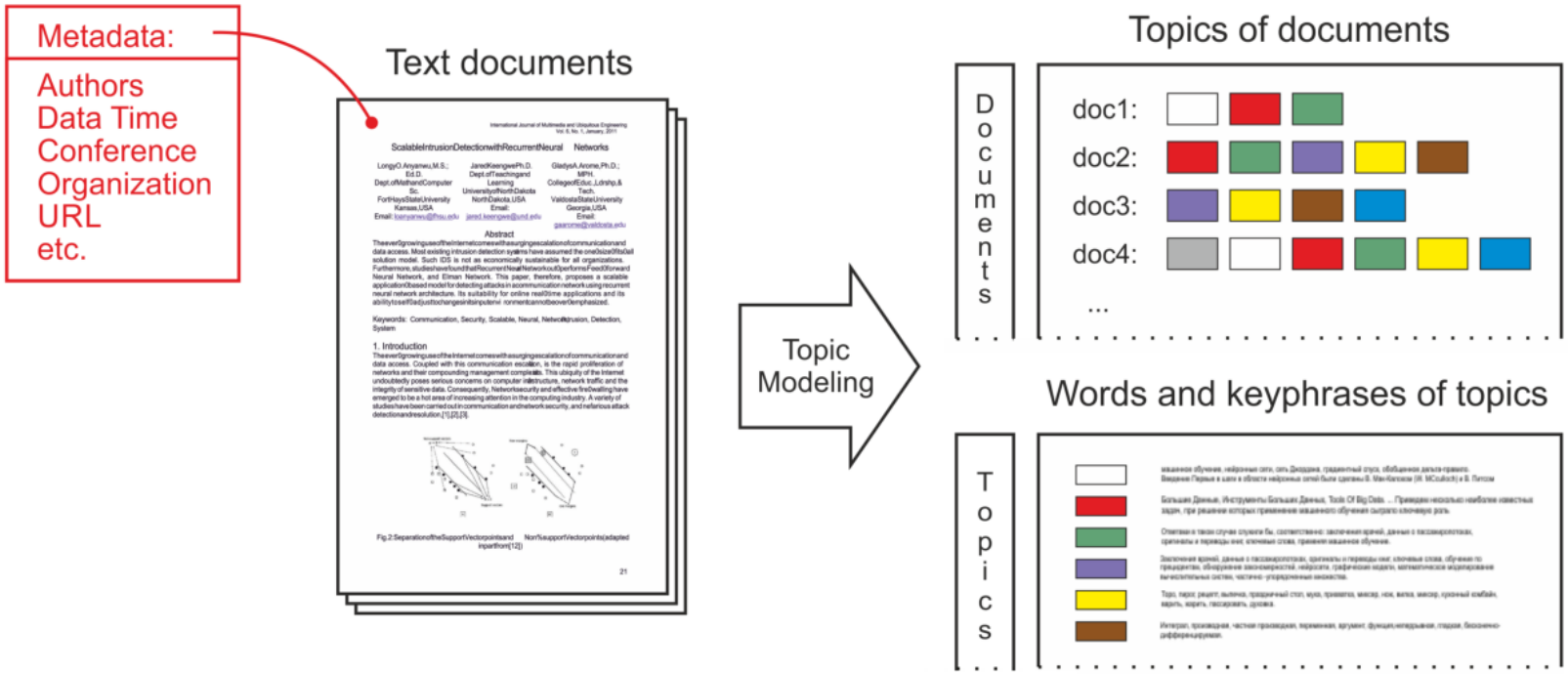
Мультимодальная тематическая модель

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



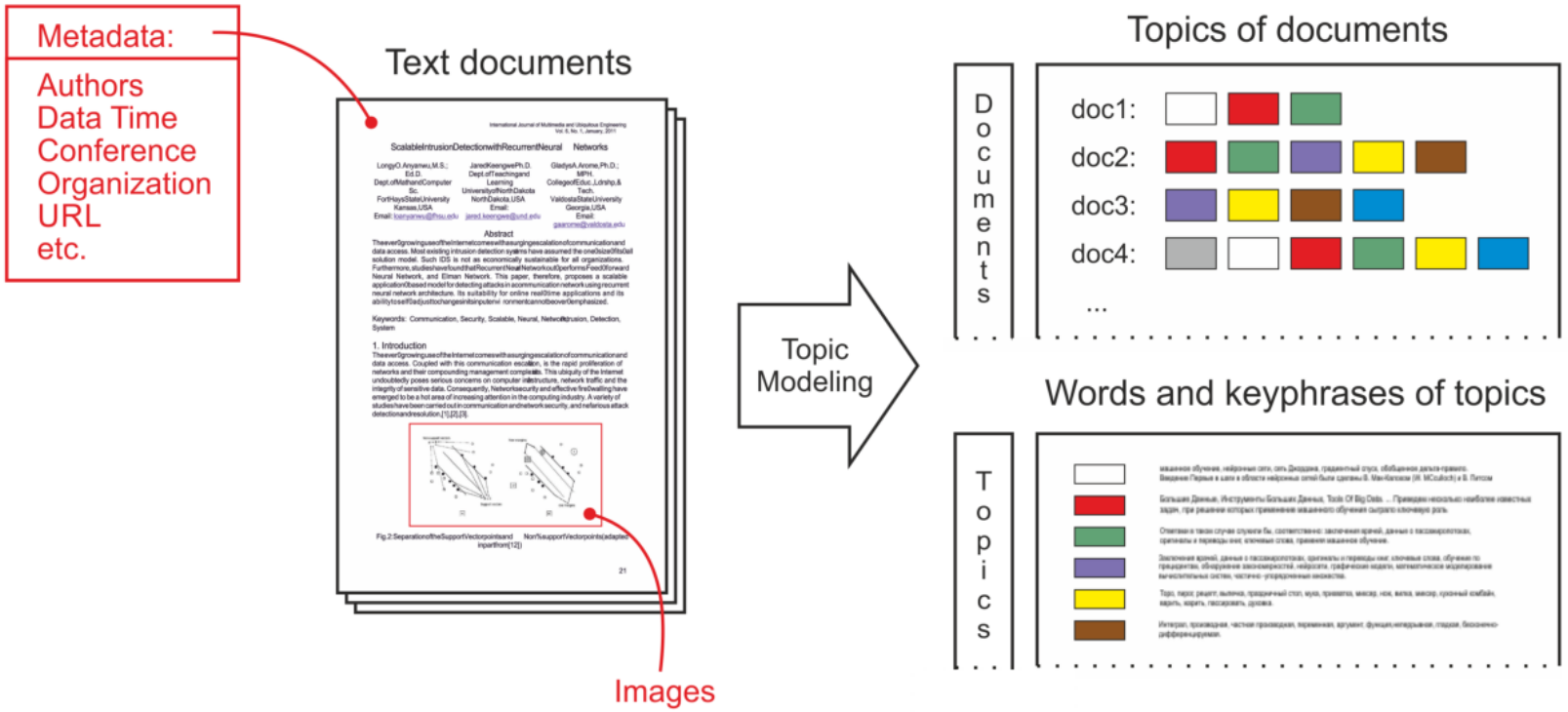
Мультимодальная тематическая модель

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



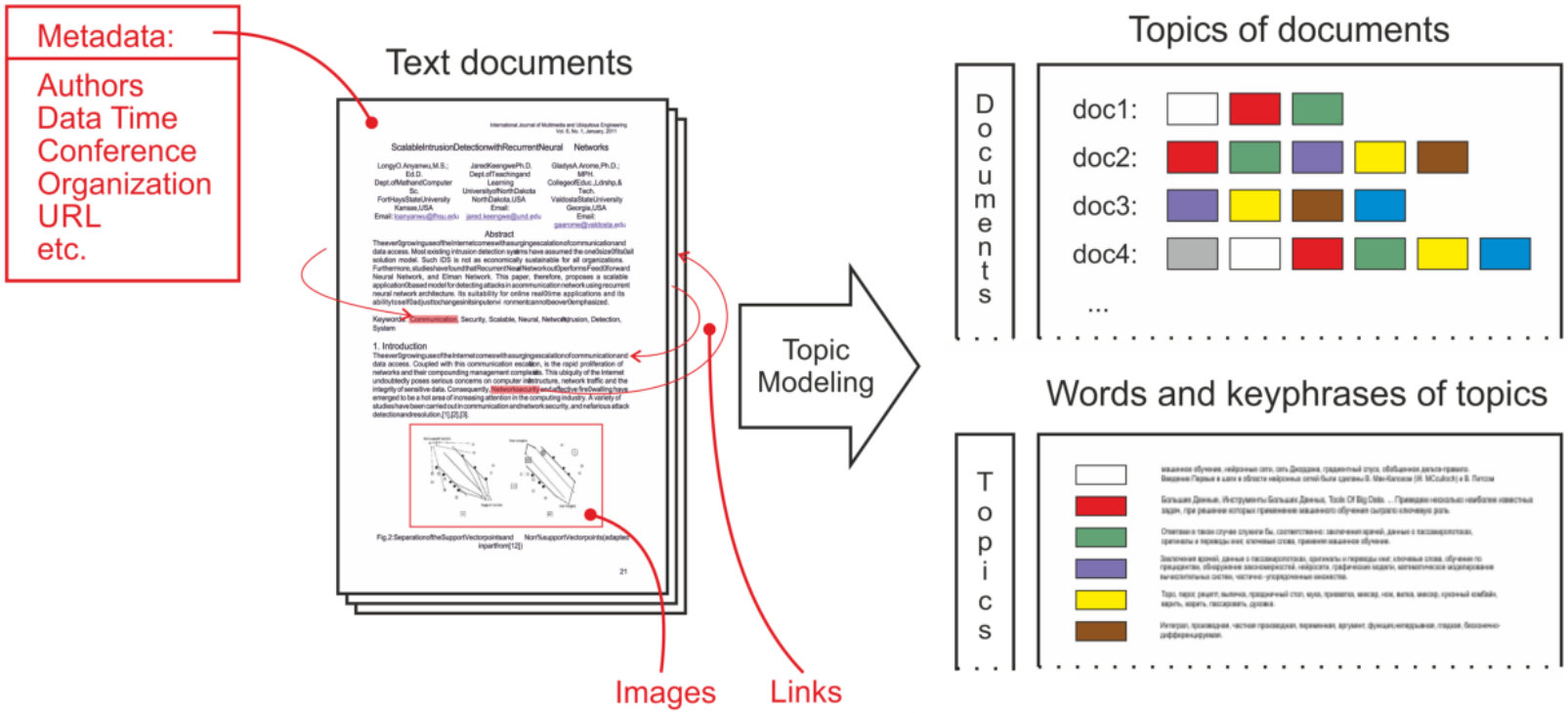
Мультимодальная тематическая модель

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



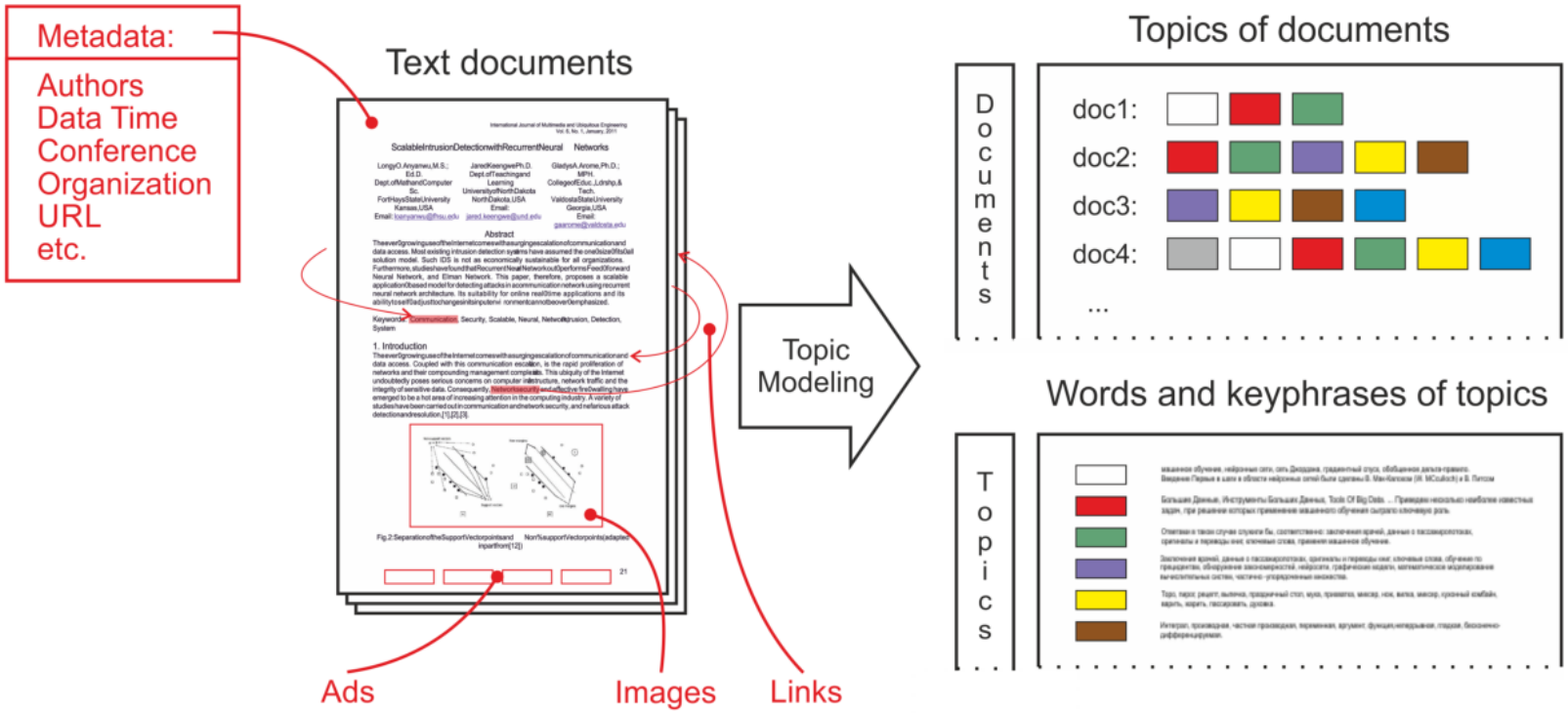
Мультимодальная тематическая модель

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



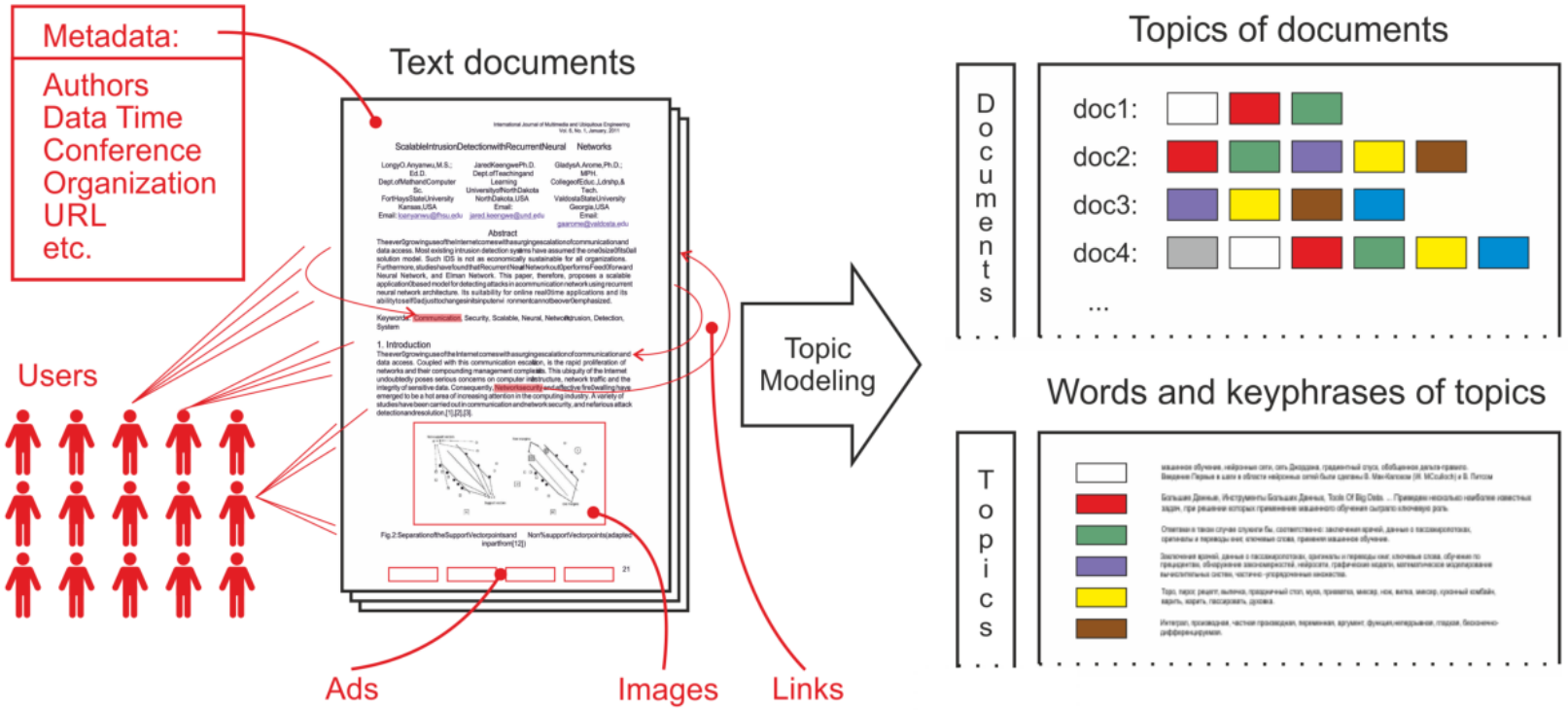
Мультимодальная тематическая модель

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



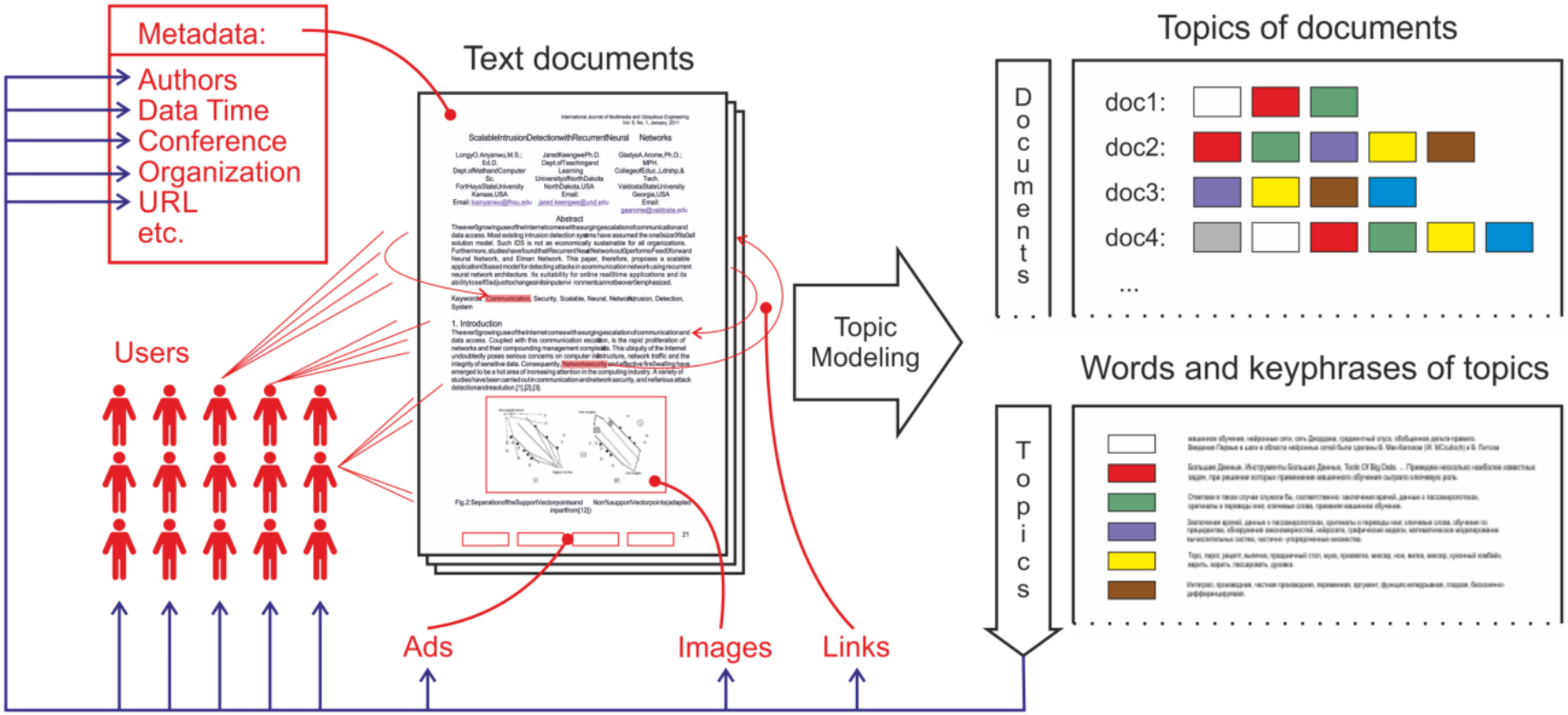
Мультимодальная тематическая модель

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 термов 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.

BigARTM: библиотека тематического моделирования

Ключевые возможности:

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

Сообщество:

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

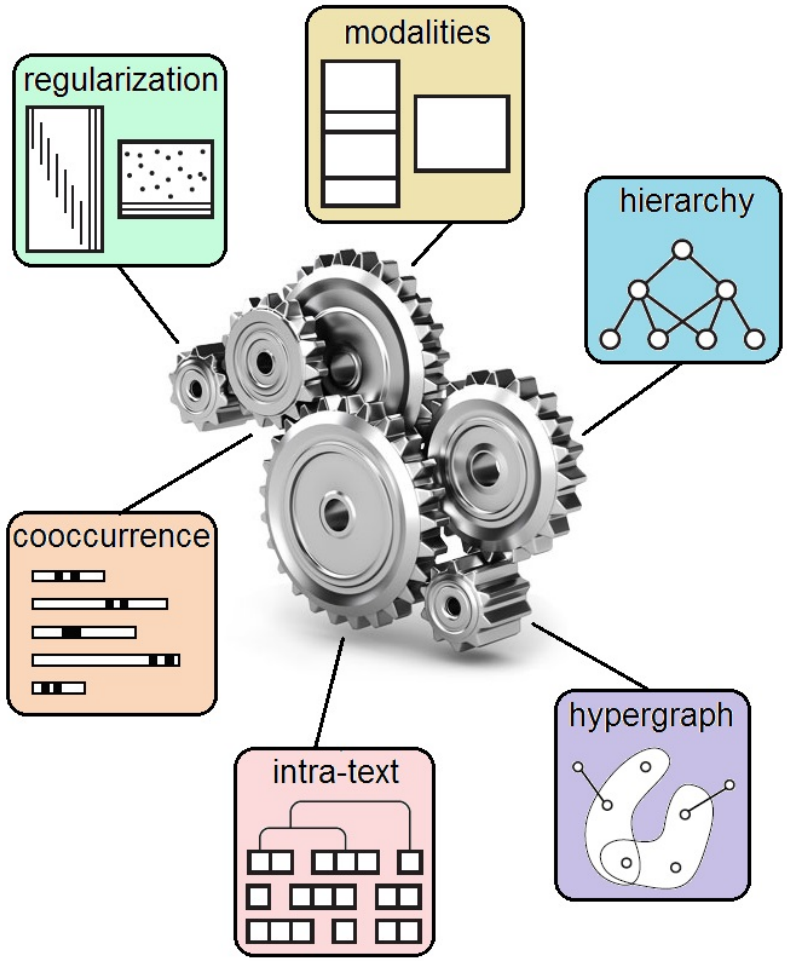
Лицензия и среда разработки:

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

Шесть ключевых механизмов BigARTM

1. ✓ регуляризация
2. ✓ модальности
3. иерархия тем
4. сочетаемость термов
5. внутритекстовые связи
6. гиперграфовые данные

(далее в курсе они будут изучаться подробно)



Этапы решения практических задач

- Установка BigARTM, запуск примеров
- Получение коллекции, перевод в удобный формат
- Предварительная обработка текстов
- Реализация базовой модели (обычно PLSA)
- Прикладное использование тематической модели
- При необходимости интеграция с другими методами NLP
- Измерение качества тематической модели
- Добавление регуляризаторов и модальностей
- Оптимизация коэффициентов регуляризации
- Оптимизация весов / оценка полезности модальностей
- Оптимизация числа тем
- Интерпретация и визуализация тем

Методы предварительной обработки текста

- Удаление чисел, не-слов и «прочей грязи»
- Устранение переносов (когда текст был в pdf)
- Исправление опечаток (для пользовательских данных)
- Лемматизация (для русского языка)
- Стемминг (для английского языка)
- Удаление слишком редких слов (если «мешок слов»)
- Удаление стоп-слов (если не строить фоновые темы)
- Автоматическое выделение терминов (ATE)
- Выделение именованных сущностей (NER)
- Сокращение словаря (Vocabulary Reduction)

Извлечение объектов и фактов из текстов в Яндексе. Лекция для Малого ШАДа, 2013. `https://habr.com/ru/company/yandex/blog/205198`

`https://nlpub.ru/Обработка_текста`

Какими будут практические задания по курсу

Задача-минимум: научиться решать задачи NLP с использованием тематического моделирования в BigARTM

Задача-максимум: исследовать одну из открытых проблем:

1. несбалансированность тем: слишком крупные темы дробятся, слишком мелкие объединяются
2. динамическое создание новых тем в текстовых потоках (first story detection)
3. автоматическое определения ролей слов и разделение их на тематическую и общую лексику
4. автоматическое именование и суммаризация тем (пусть каждая тема сама рассказывает о себе)
5. автоматизация подбора гиперпараметров
6. графическая визуализация «карт знаний» (D3.js)

Датасеты для заданий по спецкурсу

- Научные статьи: arXiv, PubMed, Semantic Scholar
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
- Википедия
- Акты арбитражных судов РФ
- Данные социальных сетей: Twitter, VK, LJ,...
- TechCrunch (английский)
- Открытые датасеты (английский): 20 newsgroups, NIPS, KOS
- Транзакции клиентов Sberbank DSD 2016

---

http://bigartm.org
http://drive.google.com/drive/folders/1PPnw6aZOJAJoLRYuwdGm437RssV-XQx0

Десять открытых проблем тематического моделирования

1. Гарантирование качества (интерпретируемости) всех тем
2. Надёжное разделение лексики на тематическую и общую
3. Моделирование тематики в локальном контексте
4. Динамическое создание новых тем в текстовых потоках
5. Обеспечение устойчивости тематических моделей
6. Оптимизация гиперпараметров в потоковом режиме
7. Бережное слияние моделей нескольких коллекций
8. Автоматическое именование и реферирование тем
9. Создание больших предобученных тематических моделей
10. Применение гиперграфовых тематических моделей

Эти задачи должны решаться на любых данных «из коробки», т.е. без экспериментов по подбору гиперпараметров.

Теоретическое задание сегодня

Два упражнения на принцип максимума правдоподобия:

- Униграммная модель документов: $p(w|d) = \xi_{dw}$
  Найти параметры модели ξ_{dw}.
- Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d
  Найти параметры модели ξ_w.

Творческое задание:

- Предложить модель, которая определяет роли слов в текстах и разделяет их на три группы:
  — слова общей лексики (фон)
  — тематические слова
  — специфичные слова документа (шум)

Подсказка 1: регуляризаторы сглаживания фоновых и шумовых тем.
Подсказка 2: можно использовать документную частоту слов.

Резюме

- Тематическая модель — это:
  — мягкая кластеризация документов по кластерам-темам
  — стохастическое матричное разложение
  — вероятностное векторное представление текстов и слов
- Задача некорректно поставлена, её решение не единственно
- ARTM позволяет накладывать много ограничений на модель
- Лемма о максимизации на единичных симплексах позволяет выводить EM-алгоритм для широкого класса моделей
- Онлайновый EM-алгоритм имеет сложность $O(n|T|)$
- BigARTM — эффективная открытая реализация
- Что дальше:
  — придумываем регуляризаторы, обобщаем модели
  — применяем для решения практических задач
  — измеряем и оптимизируем качество моделей