Постановки задач оптимизации в машинном обучении
Часть 3. Тематическое моделирование текстовых и транзакционных данных

Воронцов Константин Вячеславович

(Лаборатория машинного интеллекта МФТИ)

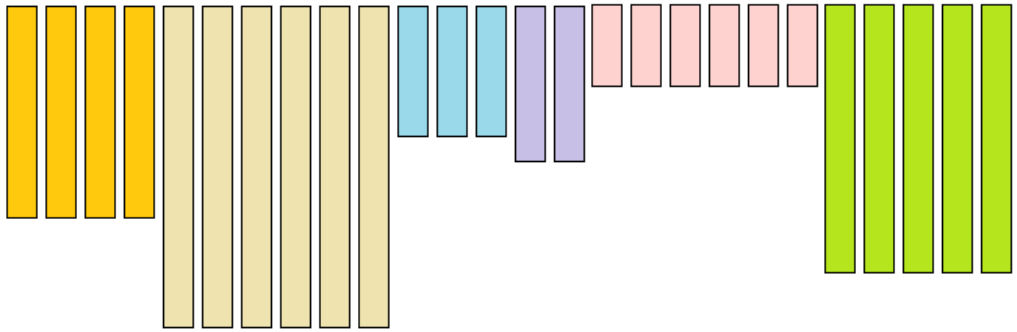
• Управление, информация, оптимизация • «Сириус», Сочи • 23–29 августа 2020

Содержание

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, можно отбросить $\omega_j \equiv 0$, сократив 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:

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

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

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

Доказательство Леммы

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

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

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 $\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = 0$.

Вероятностная тематическая модель текстовой коллекции

- W — конечное множество (словарь) *термов* (слов)
- D — конечное множество текстовых документов
- T — конечное множество тем
- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция — i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- переменные d_i, w_i — наблюдаемые, темы t_i — скрытые
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- гипотеза условной независимости: $p(w|d,t) = p(w|t)$

Тематическая модель языка, по формуле полной вероятности:

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

Постановка задачи тематического моделирования

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

Критерий: максимум правдоподобия

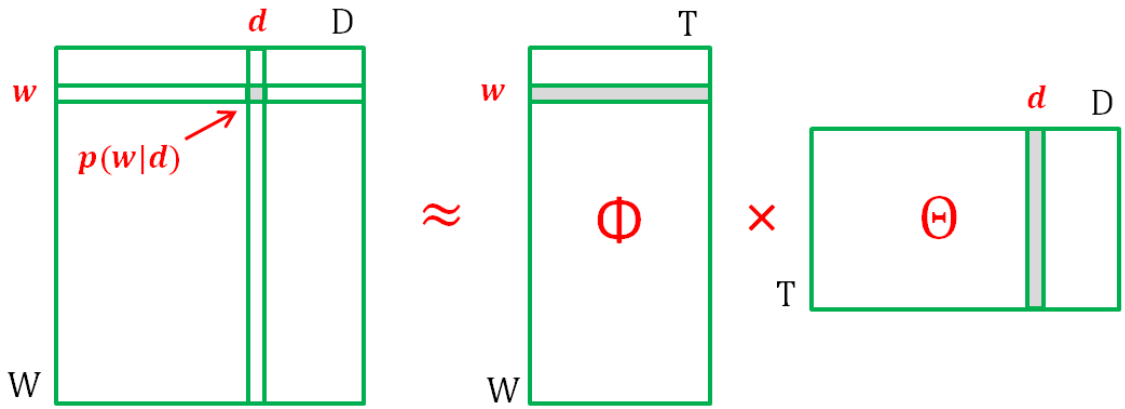
$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi, \Theta}$$

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 тематического моделирования некорректно поставлена

Тематическое моделирование — это задача низкорангового стохастического матричного разложения ($|T| \ll |D|, |W|$)



Задача *некорректно поставлена* по Адамару, так как множество её решений в общем случае бесконечно:
если $\Phi\Theta$ — решение, то $(\Phi S)(S^{-1}\Theta)$ — тоже решение

Регуляризация — доопределение решения с помощью дополнительного критерия: $L(\Phi, \Theta) + \tau R(\Phi, \Theta) \to \max$

PLSA и LDA — наиболее известные тематические модели

PLSA, Probabilistic Latent Semantic Analysis:

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

LDA, Latent Dirichlet Allocation:

$$R(\Phi, \Theta) = \beta_0 \sum_t \sum_w \beta_w \ln \phi_{wt} + \alpha_0 \sum_d \sum_t \alpha_t \ln \theta_{td}$$

— распределения ϕ_t близки к заданному распределению β
— распределения θ_d близки к заданному распределению α

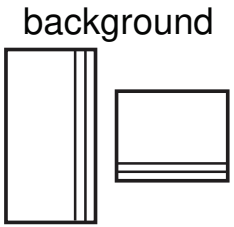
Можно также говорить об априорных распределениях Дирихле, но регуляризация проще и понятнее, а эффект тот же

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

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

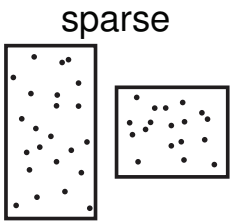
Воронцов К.В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM // www.MachineLearning.ru, 2020.

Регуляризаторы для улучшения интерпретируемости тем



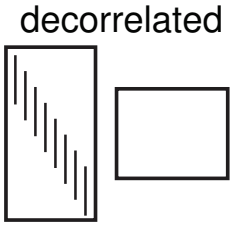
Сглаживание фоновых тем $B \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_w \beta_w \ln \phi_{wt} + \alpha_0 \sum_d \sum_{t \in B} \alpha_t \ln \theta_{td}$$



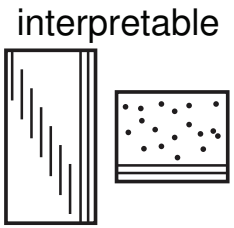
Разреживание предметных тем $S = T \setminus B$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_w \beta_w \ln \phi_{wt} - \alpha_0 \sum_d \sum_{t \in S} \alpha_t \ln \theta_{td}$$



Декоррелирование для повышения различности тем:

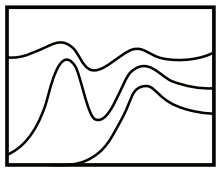
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем

Регуляризаторы для учёта дополнительной информации

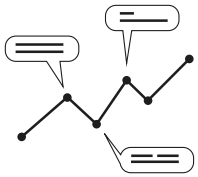
temporal



Темпоральные модели с модальностью времени i:

$$R(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} \left| \phi_{it} - \phi_{i-1,t} \right|$$

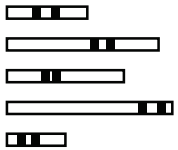
regression



Линейная модель регрессии $\hat{y}_d = \langle v, \theta_d \rangle$ документов:

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

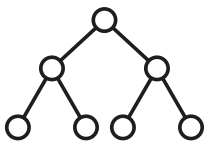
coherence



Модели сочетаемости слов (n_{uv} — частота битерма):

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt}$$

hierarchy



Связь родительских тем t с дочерними подтемами s:

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

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

ЕМ-алгоритм — метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$:

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

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

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

Доказательство (по Лемме о максимизации на симплексах)

Применим Лемму к log-правдоподобию с регуляризатором R:

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

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

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

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

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

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).

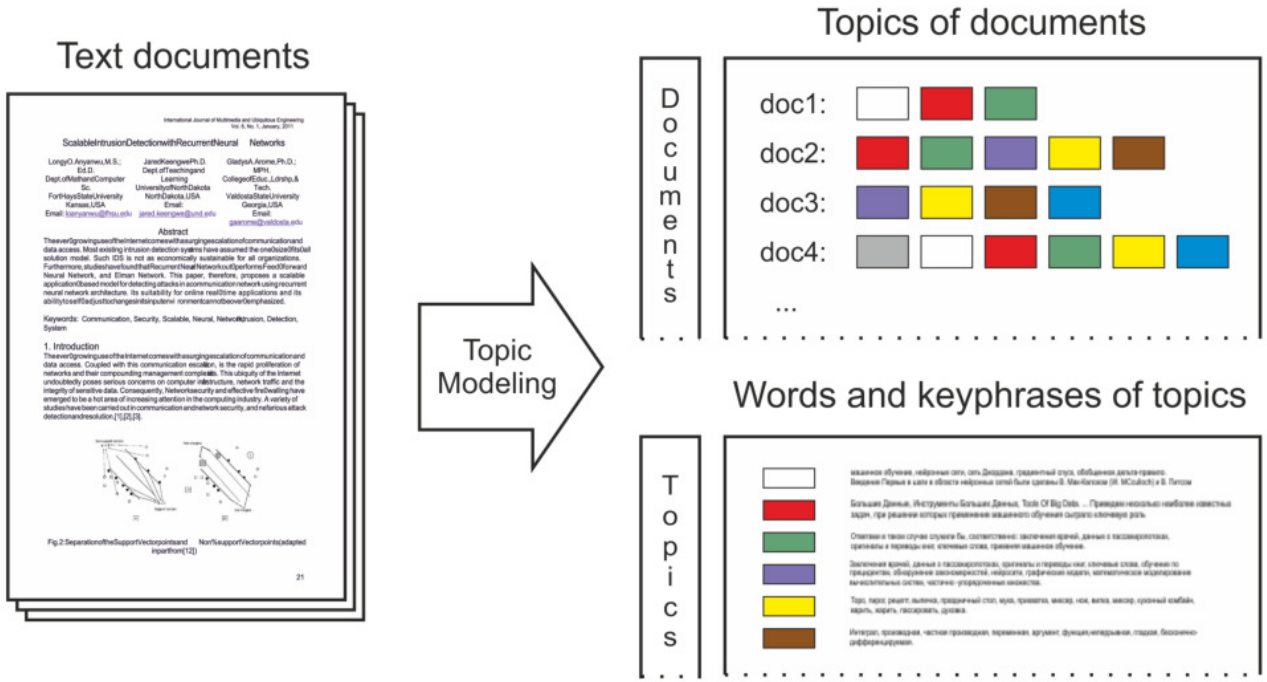
Документ d вырожден, если для всех тем $t \in T$

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

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

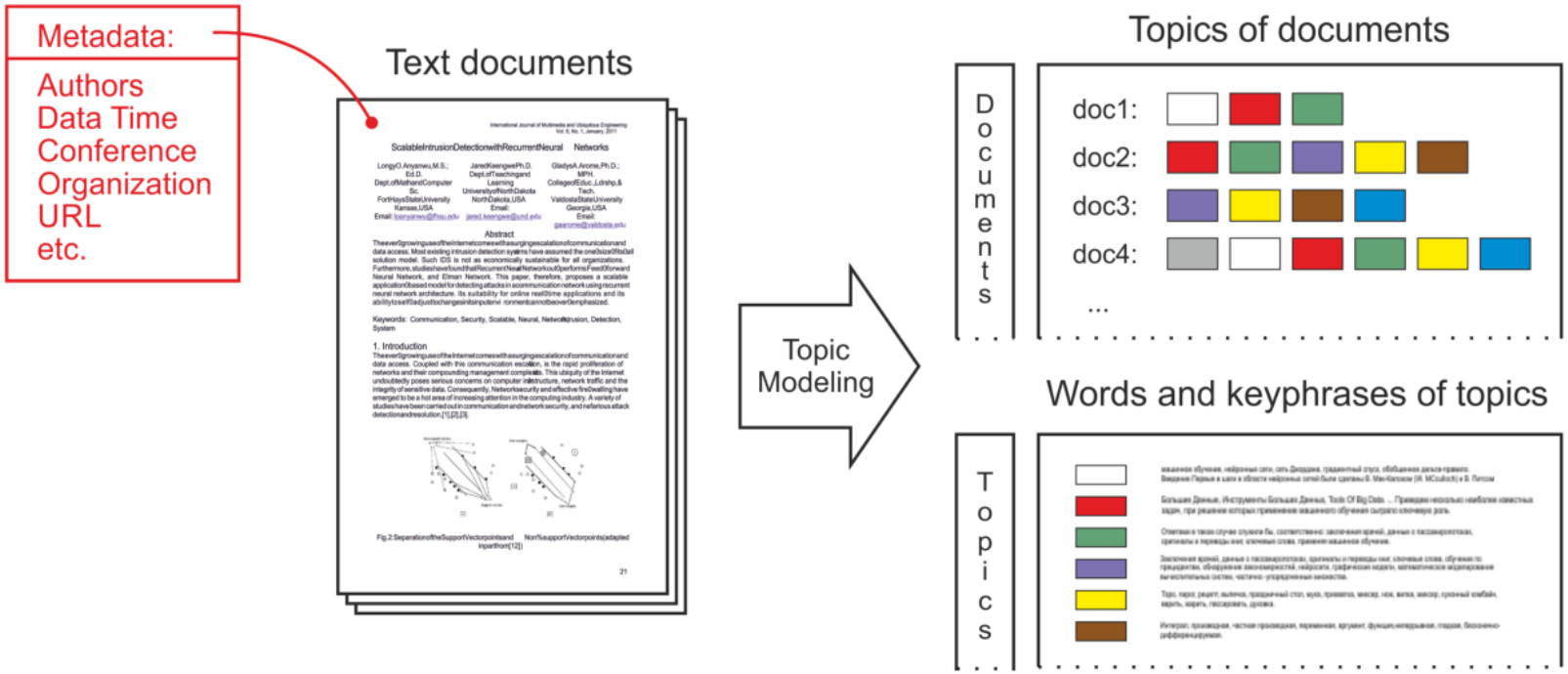
Мультимодальная тематическая модель

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



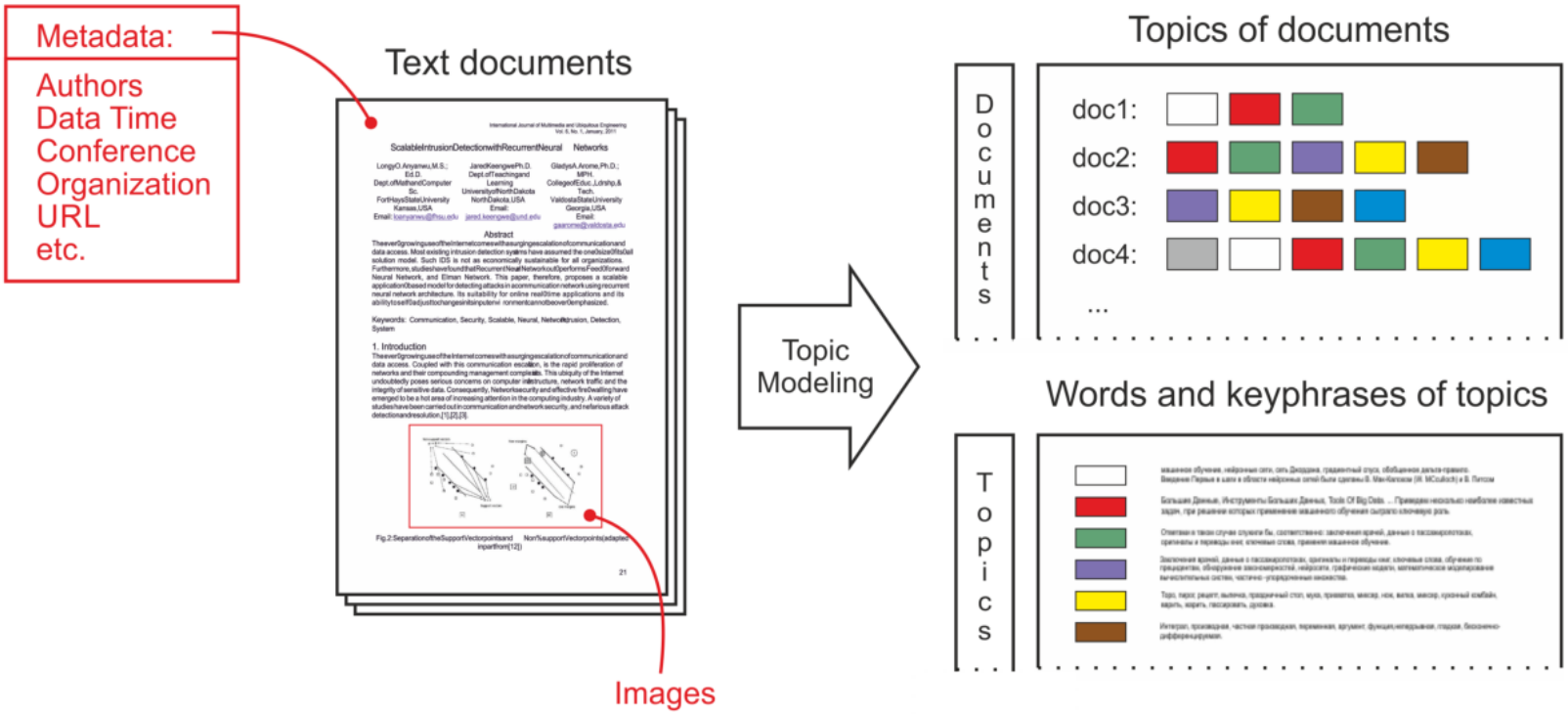
Мультимодальная тематическая модель

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



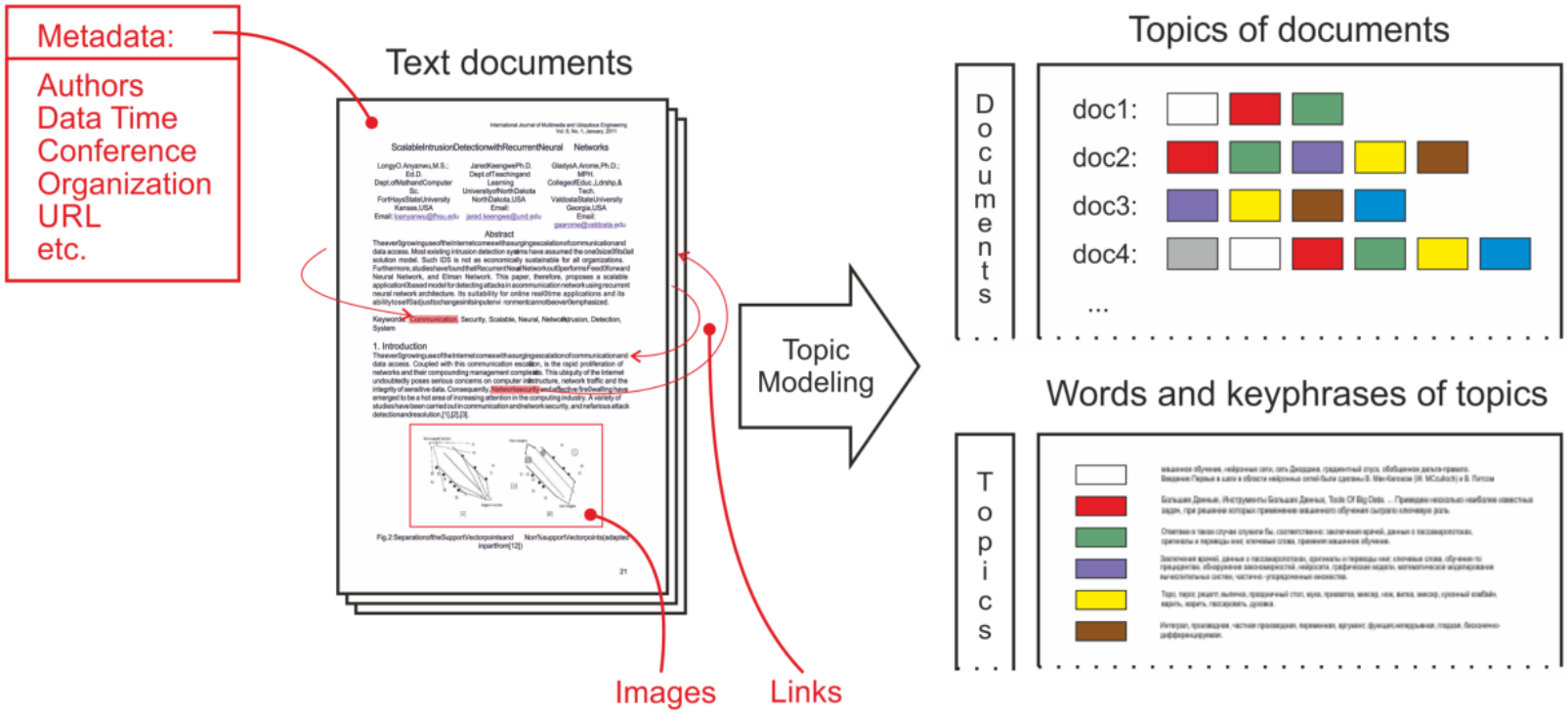
Мультимодальная тематическая модель

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



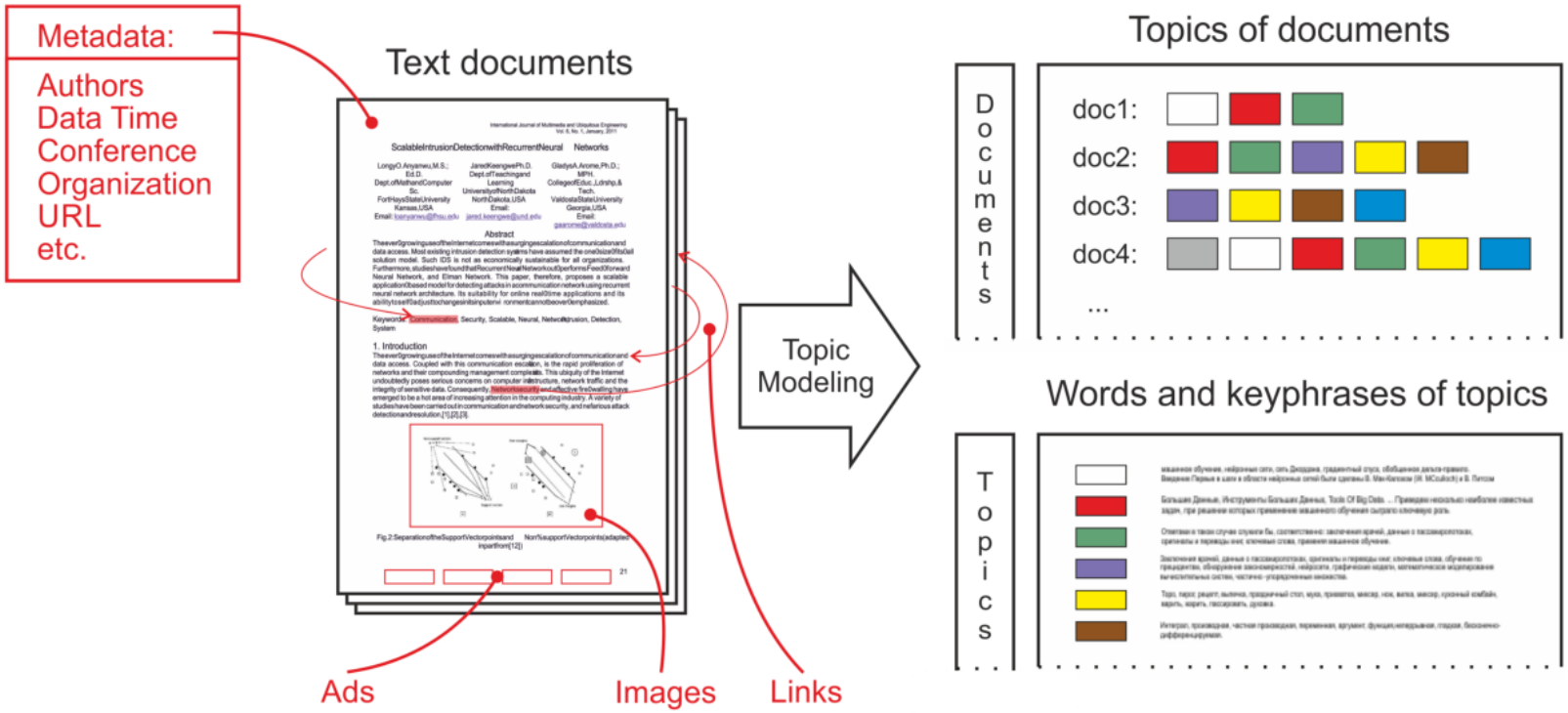
Мультимодальная тематическая модель

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



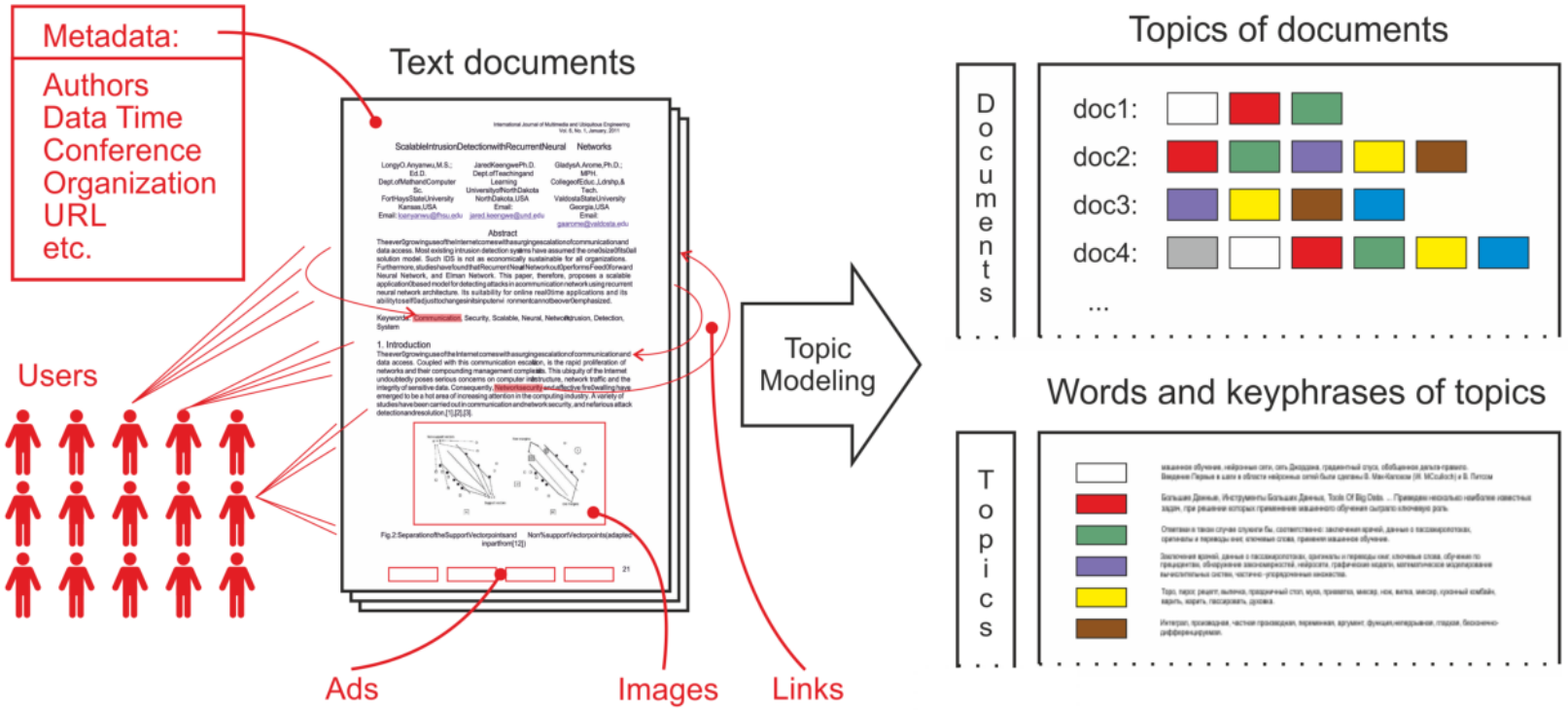
Мультимодальная тематическая модель

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



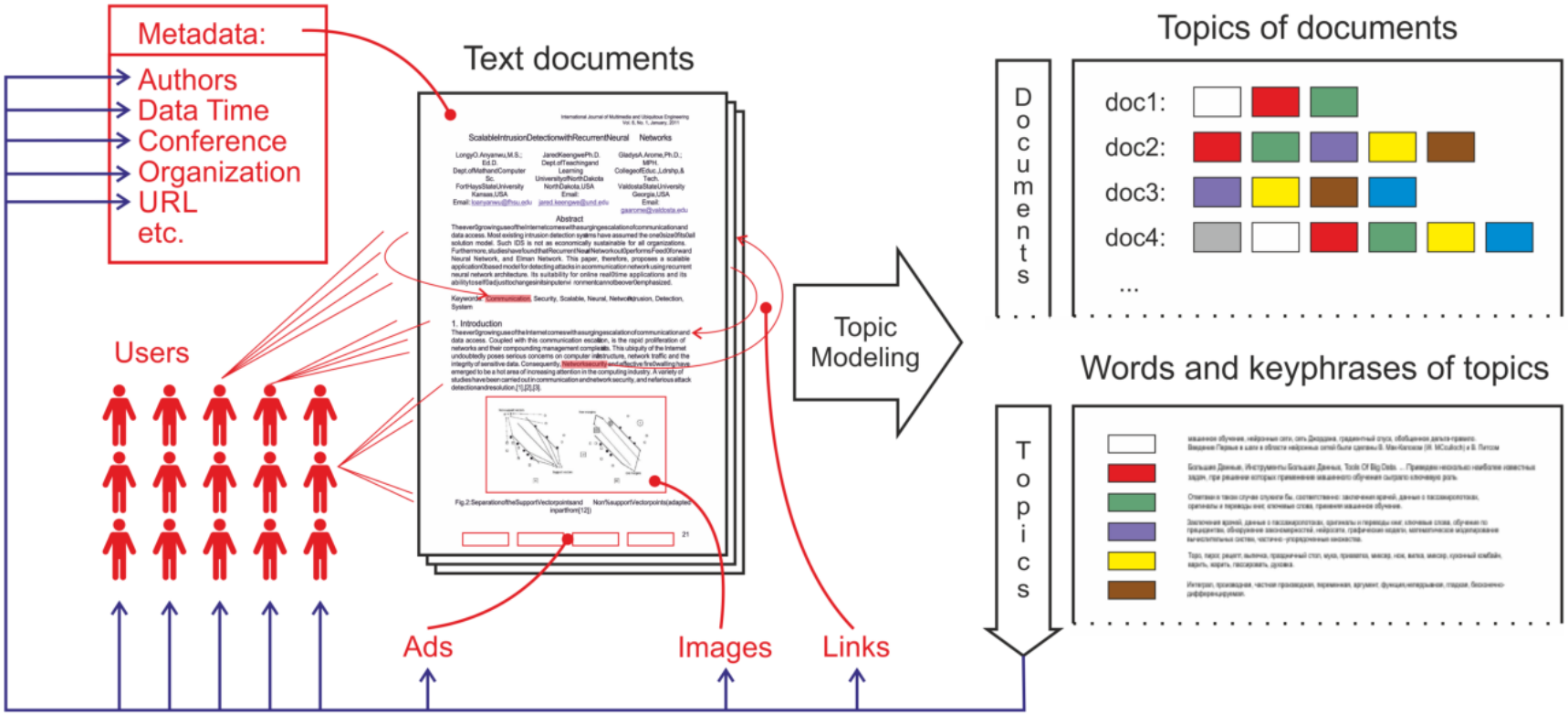
Мультимодальная тематическая модель

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:

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

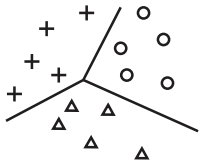
$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}} \left(\phi_{wt} \theta_{td} \right)$$

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

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

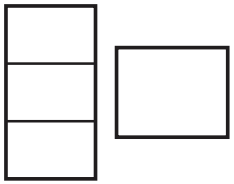
Регуляризаторы для мультимодальных тематических моделей

supervised



Модальности меток классов, категорий или тегов для классификации/категоризации/тегирования текстов

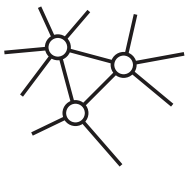
multilanguage



Модальность языков и регуляризация со словарём $\pi_{uwt} = p(u|w,t)$ переводов с языка k на ℓ:

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

graph



Модальность вершин графа v, содержащих D_v:

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v)\in E} S_{uv} \sum_{t\in T} n_t^2 \left(\frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|}\right)^2.$$

geospatial

Модальность геолокаций g с близостью $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g'\in G} S_{gg'} \sum_{t\in T} n_t^2 \left(\frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}}\right)^2$$

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.

Мультимодальные транзакционные данные (примеры)

Выборка может содержать не только пары (d, w), но также n-ки термов разных модальностей, связанных общей темой.

- **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 финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g
- **Данные о пассажирских авиаперелётах:**
  (u, a, b, c) — перелёт клиента u из a в b авиакомпанией c
- **Данные о связности текста:**
  $(w_1, \ldots, w_{n_s})$ — слова образуют фразу или предложение s

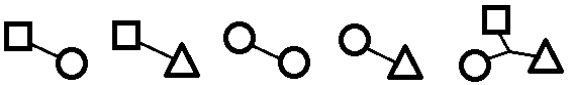
Задача: по наблюдаемой выборке рёбер гиперграфа найти латентные тематические векторные представления его вершин.

Тематическая модель гиперграфа: определения и обозначения

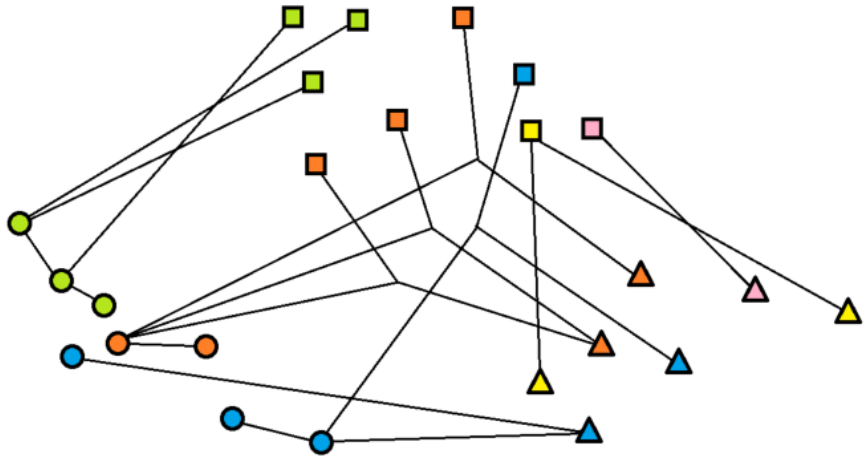
$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:



T — множество тем:



Исходные данные:

E_k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x): вершина-контейнер $d \in V$ и вершины $x \subset V$,
n_{kdx} — число вхождений ребра (d, x) в выборку E_k

Тематическая модель гиперграфа: основные предположения

- в ребре (d, x) подмножество $x \subset V$ может быть любым, независимо от типа ребра k
- первая *гипотеза условной независимости*: тематика контейнера $p(t|d)$ не зависит от типа ребра k
- вторая *гипотеза условной независимости*: распределение $p(v|t)$ термов v модальности V^m в теме t не зависит ни от контейнера d, ни от типа ребра k
- третья *гипотеза условной независимости*: термы $v \in x$ в ребре (d, x) не зависят друг от друга
- *гипотеза «мешка транзакций»*: выборка транзакций типа k порождается случайно и независимо из

$$p_k(d, x) = p(d) \sum_{t \in T} p(t|d) \prod_{v \in x} p(v|t)$$

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k

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

Задача максимизации взвешенной суммы log-правдоподобий по всем типам рёбер:

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

при ограничениях неотрицательности и нормировки:

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

где $\tau_k > 0$ — веса типов рёбер.

ЕМ-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

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

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

$$\text{E-шаг:}\quad p_{tdx} = \operatorname*{norm}_{t\in T}\Big(\theta_{td}\prod_{v\in x}\phi_{vt}\Big)$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{vt} = \operatorname*{norm}_{v\in V^m}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k} \big[v\in x\big] n_{kdx} p_{tdx} + \phi_{vt}\dfrac{\partial R}{\partial \phi_{vt}}\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k} n_{kdx} p_{tdx} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

Доказательство (по лемме о максимизации на симплексах)

Применим Лемму к log-правдоподобию с регуляризатором R:

$$\phi_{vt} = \operatorname*{norm}_{v \in V_m} \left(\phi_{vt} \sum_{k \in K} \tau_k \sum_{dx \in E_k} n_{kdx} \frac{\theta_{td}}{p(x|d)} \frac{\partial}{\partial \phi_{vt}} \prod_{u \in x} \phi_{ut} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right) =$$

$$= \operatorname*{norm}_{v \in V_m} \left(\sum_{k \in K} \sum_{dx \in E_k} \tau_k n_{kdx} [v \in x] p_{tdx} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right)$$

$$\theta_{td} = \operatorname*{norm}_{t \in T} \left(\theta_{td} \sum_{k \in K} \tau_k \sum_{x \in d} n_{kdx} \frac{1}{p(x|d)} \prod_{v \in x} \phi_{vt} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) =$$

$$= \operatorname*{norm}_{t \in T} \left(\sum_{k \in K} \sum_{x \in d} \tau_k n_{kdx} p_{tdx} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$$

Доводы в пользу исключения матрицы Θ из модели

- Вычисление $\theta_{td}(\Phi)$ за один линейный проход по документу
- Ограничение-равенство $\theta_{td} = \theta_{td}(\Phi)$ играет роль регуляризатора и повышает устойчивость модели
- Сокращение размерности модели, уменьшение переобучения
- Размерность Φ растёт сублинейно с ростом коллекции

Первая итерация ЕМ-алгоритма без регуляризации при равномерном начальном приближении $\theta^0_{td} = \frac{1}{|T|}$:

$$\theta_{td}(\Phi) = \underset{t\in T}{\operatorname{norm}}\Big(\sum_w n_{dw} p_{tdw}\Big) = \sum_w \frac{n_{dw}}{n_d} \frac{\phi_{wt}\theta^0_{td}}{\sum_s \phi_{ws}\theta^0_{sd}} = \sum_w \frac{p_{dw}\phi_{wt}}{\sum_s \phi_{ws}},$$

где $p_{dw} = \frac{n_{dw}}{n_d}$ — частотная оценка условной вероятности $p(w|d)$

И.А.Ирхин, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста. ЖВМиМФ. 2020.

ЕМ-алгоритм для ARTM с исключённой матрицей Θ

Максимизация логарифма правдоподобия:

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\big(\phi_{wt}\theta_{td}\big);$$

$$n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw};$$

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} + \phi_{wt} \sum_{d \in D} \sum_{s \in T} \left(\frac{n_{sd}}{\theta_{sd}} + \frac{\partial R}{\partial \theta_{sd}} \right) \frac{\partial \theta_{sd}}{\partial \phi_{wt}} \right)$$

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста, 2020.

Доказательство (по Лемме о максимизации на симплексах)

Оптимизационная задача М-шага относительно Φ и $\Theta(\Phi)$:

$$Q(\Phi) = \sum_{d\in D}\sum_{u\in W}\sum_{s\in T} n_{du}p_{sdu}\big(\ln\phi_{us} + \ln\theta_{sd}(\Phi)\big) + R\big(\Phi, \Theta(\Phi)\big) \to \max_{\Phi}$$

Применим Лемму к регуляризованному log-правдоподобию Q:

$$\phi_{wt}\frac{\partial Q}{\partial \phi_{wt}} = \sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} +$$

$$+ \sum_{d,s,u} n_{du}p_{sdu}\frac{\phi_{wt}}{\theta_{sd}}\frac{\partial \theta_{sd}}{\partial \phi_{wt}} + \phi_{wt}\sum_{d,s}\frac{\partial R}{\partial \theta_{sd}}\frac{\partial \theta_{sd}}{\partial \phi_{wt}} =$$

$$= n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \phi_{wt}\sum_{d,s}\left(\frac{n_{sd}}{\theta_{sd}} + \frac{\partial R}{\partial \theta_{sd}}\right)\frac{\partial \theta_{sd}}{\partial \phi_{wt}}$$

Частный случай $\theta_{td}(\Phi) = \sum_w p_{dw} \operatorname{norm}_t(\phi_{wt})$

Частные производные: $\frac{\partial \theta_{sd}}{\partial \phi_{wt}} = p_{wd} h_w \left(\delta_{st} - \phi_{ws} h_w\right)$

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\theta_{td} = \sum_{w \in d} p_{dw} \phi_{wt} h_w; \qquad h_w = \left(\sum_t \phi_{wt}\right)^{-1};$$

$$p_{tdw} = \operatorname*{norm}_{t \in T}\left(\phi_{wt}\theta_{td}\right); \qquad c_{td} = \frac{n_{td}}{\theta_{td}} + \frac{\partial R}{\partial \theta_{td}};$$

$$n_{td} = \sum_{w \in d} n_{dw} p_{tdw}; \qquad \gamma_{dw} = \sum_{t \in T} \phi_{wt} c_{td};$$

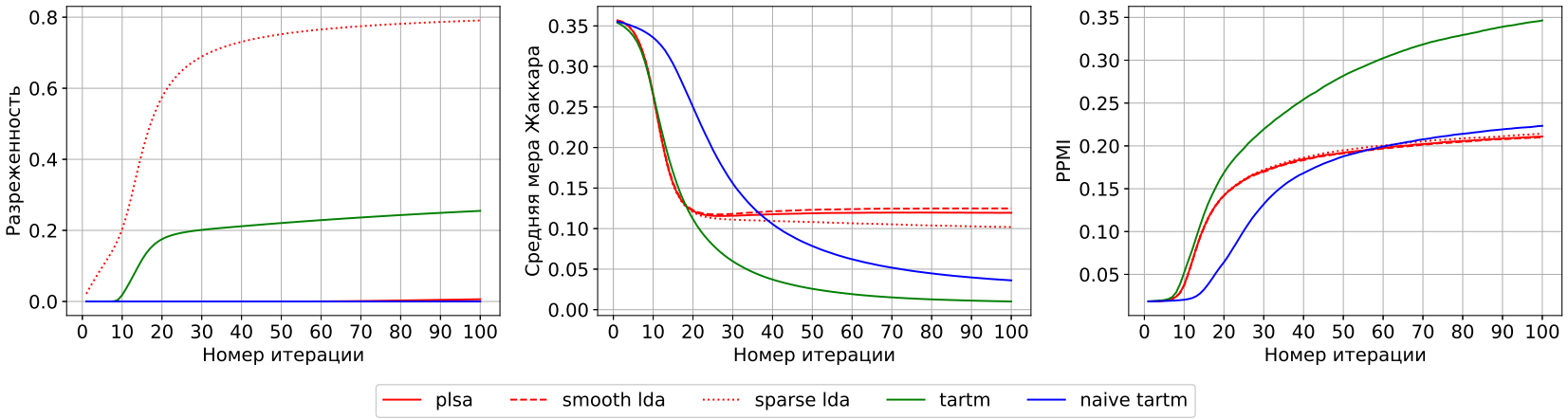
$$p'_{tdw} = p_{tdw} + n_d^{-1} \phi_{wt} h_w (c_{td} - h_w \gamma_{dw});$$

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

Эксперимент. Проверка модифицированного ЕМ-алгоритма

Коллекция NIPS, $|T| = 50$, модели:

- TARTM (Θless ARTM) — модифицированный ЕМ-алгоритм
- naive TARTM — одна итерация обычного ЕМ-алгоритма



- TARTM очищает темы от общеупотребительных слов,
- улучшает разреженность, различность, когерентность тем,
- обрабатывает документ за $O(n_d|T|)$ операций

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста, 2020.

Модульный подход ARTM: сравнение с байесовским подходом

Для построения композитных моделей в ARTM 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*

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Самый быстрый онлайновый параллельный ARTM
- Встроенная библиотека регуляризаторов и мер качества

Сообщество:

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

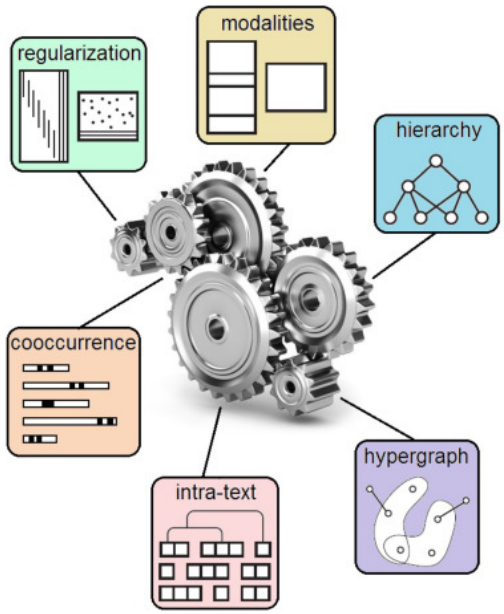
Лицензия и среда разработки:

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

Ключевые возможности библиотек BigARTM и TopicNet

BigARTM

- библиотека регуляризаторов
- мультимодальные модели
- иерархические модели
- гиперграфовые модели
- модели связности текста



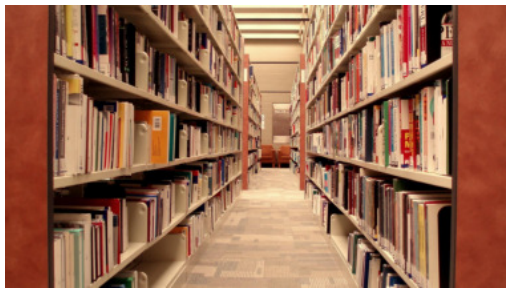
TopicNet

- Перебор сценариев регуляризации для выбора моделей
- Автоматическое протоколирование экспериментов
- Построение «банка тем» из множества моделей
- Визуализация тематических моделей

V.Bulatov, E.Egorov, E.Veselova, D.Polyudova, V.Alekseev, A.Goncharov, K.Vorontsov.
TopicNet: making additive regularisation for topic modelling accessible. LREC-2020

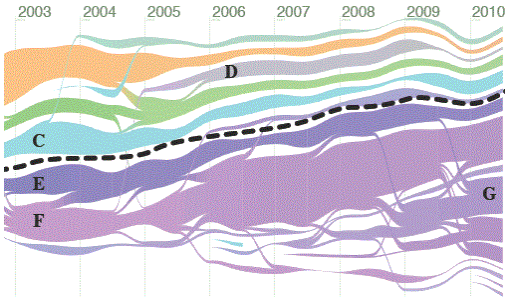
Разнообразие приложений тематического моделирования

разведочный поиск в электронных библиотеках



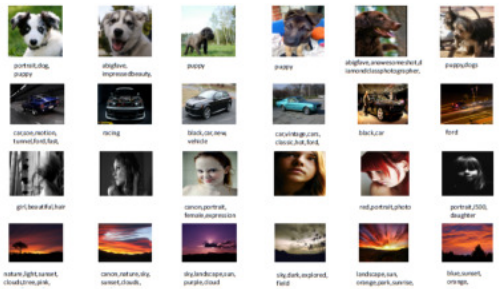
поиск тематического контента в соцсетях

детектирование и трекинг новостных сюжетов

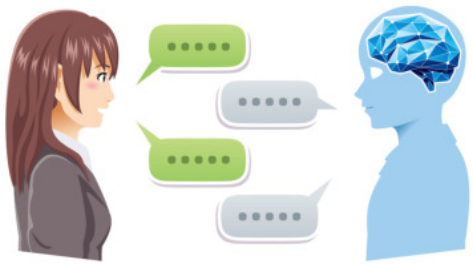


анализ банковских транзакционных данных

мультимодальный поиск текстов и изображений



управлением диалогом в разговорном интеллекте



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

Предобработка текстов

- отброшены 5% наиболее частотных слов (общая лексика)
- удаление пунктуации
- нижний регистр, ё→е
- лемматизация pymorphy2

Методика оценивания качества разведочного поиска

Поисковый запрос

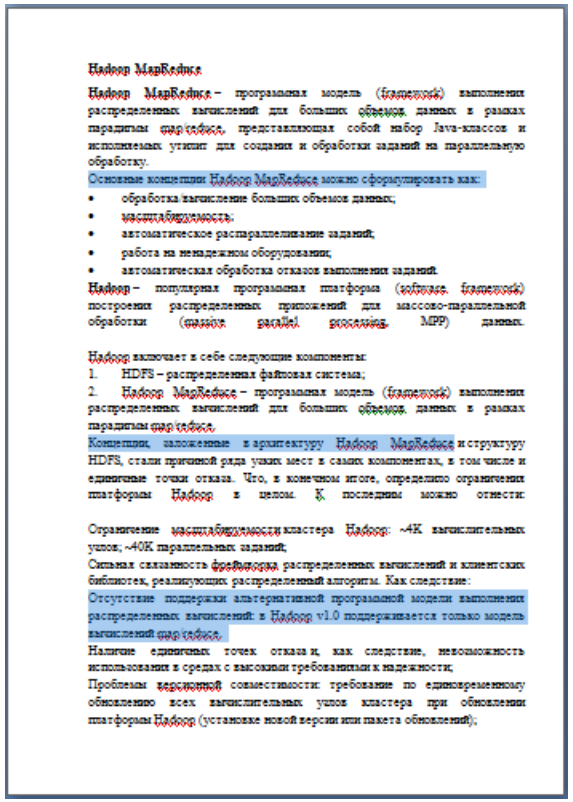
набор ключевых слов или фрагментов текста, около одной страницы А4

Поисковая выдача

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

Два задания асессорам

1. найти как можно больше статей, пользуясь любыми средствами поиска (и засечь время)
2. 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 себе 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

Векторный поиск тематически близких документов

$\theta_{tq} = p(t|q)$ — тематический вектор запроса q
$\theta_{td} = p(t|d)$ — тематические векторы документов $d \in D$

Косинусная мера близости документа d и запроса q:

$$\text{sim}(q,d) = \frac{\sum_t \theta_{tq}\theta_{td}}{\left(\sum_t \theta_{tq}^2\right)^{1/2}\left(\sum_t \theta_{td}^2\right)^{1/2}}.$$

Ранжируем документы коллекции $d \in D$ по убыванию $\text{sim}(q,d)$
Выдача тематического поиска — k первых документов.

Реализация: *векторный индекс* для быстрого поиска документов d по каждой из тем t запроса

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

A.Ianina, K.Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. FRUCT–ISMW, 2019.

Какие модели поиска сравнивались

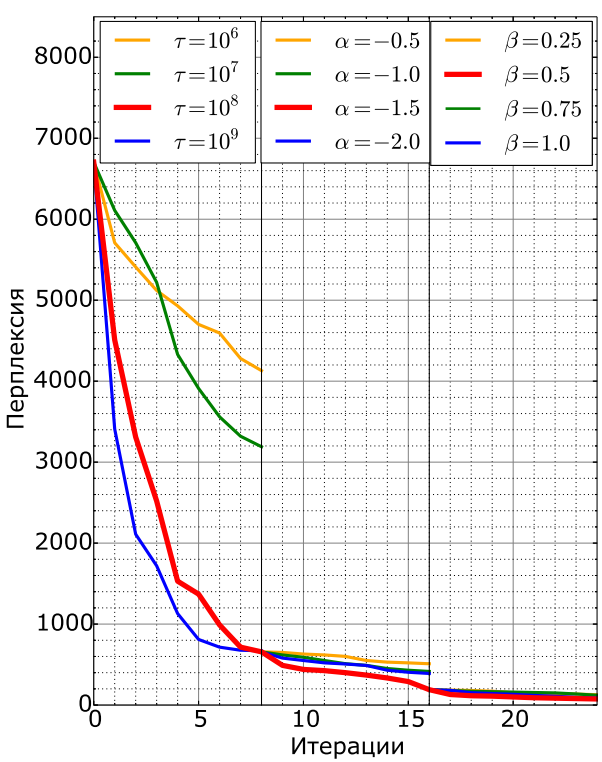
- **assessors:** результаты поиска, выполненного асессорами
- **TF-IDF, BM25:** сравнение документов по частотам слов
- **word2vec:** нетематические векторные представления слов
- **PLSA:** Probabilistic Latent Semantic Analysis (1999)
- **LDA:** Latent Dirichlet Allocation (2003)
- **ARTM:** тематическая модель с тремя регуляризаторами
- **hARTM:** двухуровневая иерархическая модель ARTM

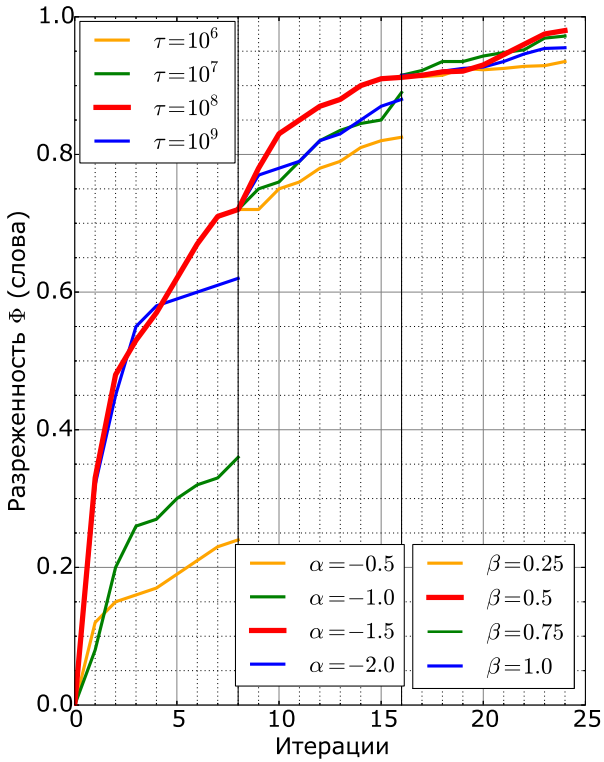
Задачи регуляризаторов в ARTM и hARTM:

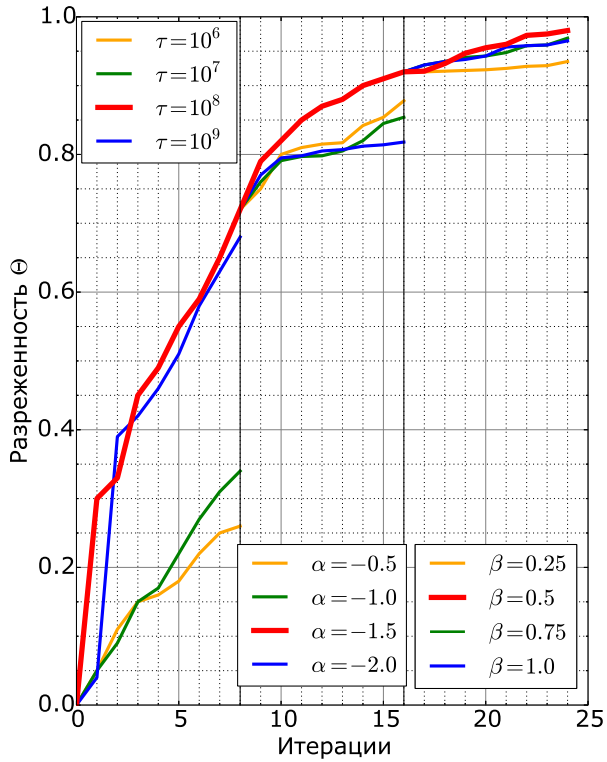
- сделать темы как можно более различными
- сделать векторы $p(t|d)$ как можно более разреженными
- не допустить вырожденности распределений $p(w|t)$

Последовательный подбор коэффициентов регуляризации

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

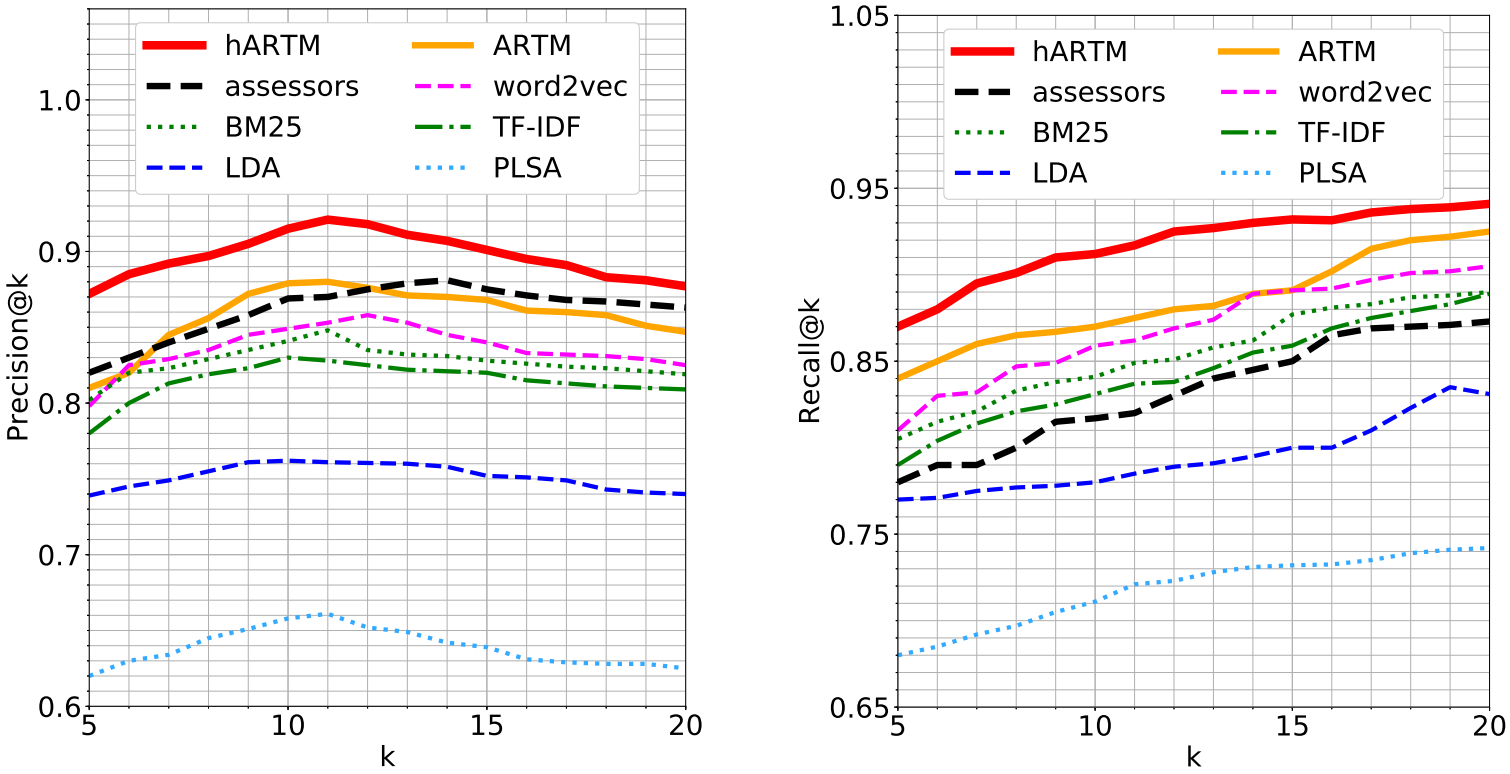






Сравнение с асессорами по качеству поиска

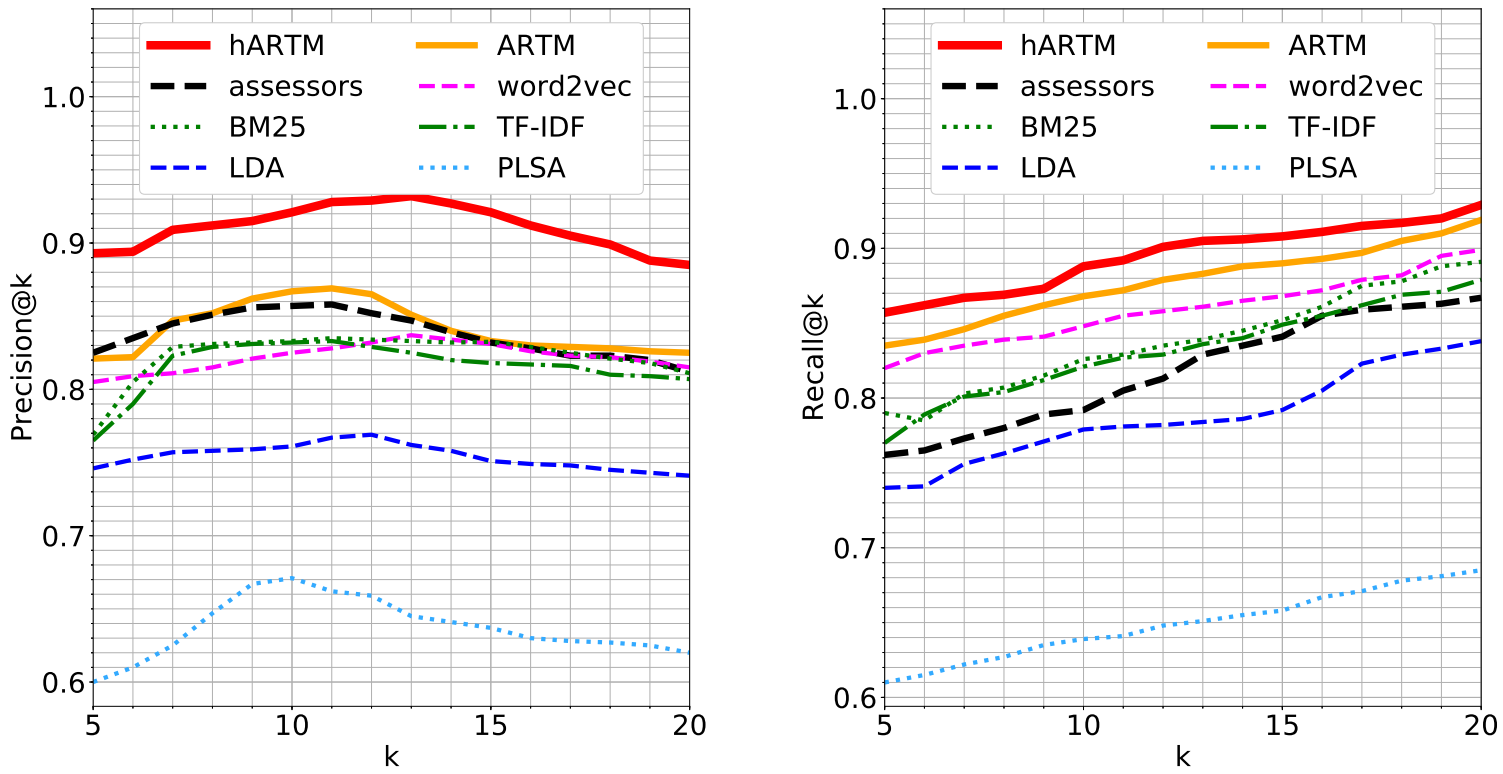
Точность и полнота по первым k позициям поисковой выдачи (коллекция Habrahabr.ru)



A.Ianina, K.Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. 2019.

Сравнение с асессорами по качеству поиска

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



A.Ianina, K.Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. 2019.

Выводы по результатам экспериментов

- Небольших асессорских данных хватает для оценивания тематических моделей, т. к. они обучаются *без учителя*
- Регуляризаторы, улучшающие интерпретируемость модели, повышают также и качество поиска
- Иерархия улучшает качество поиска (в основном точность) благодаря постепенному сужению области поиска
- Подбор траектории регуляризации и оптимизация коэффициентов регуляризации влияет на качество поиска
- При тщательной оптимизации тематический поиск превосходит как асессоров, так и конкурирующие модели

A.Ianina, K.Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. 2019.

Анализ транзакций розничных клиентов банка

Дано (Sberbank Data Science Contest):
D — множество клиентов (15 000)
W — категории = MCC-коды (Merchant Category Code) (328)
n_{dw} — сумма транзакций клиента d по категории w

Найти: темы — типы потребительского поведения клиентов
$\phi_{wt} = p(w|t)$ — структура потребления для темы t
$\theta_{td} = p(t|d)$ — типы потребления клиента d

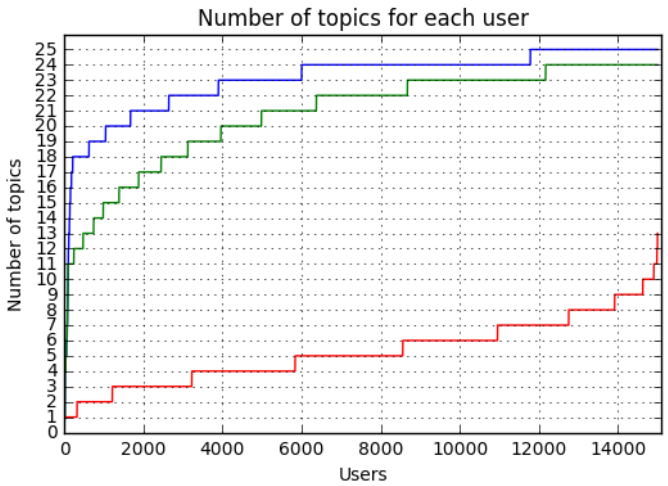
Регуляризаторы:

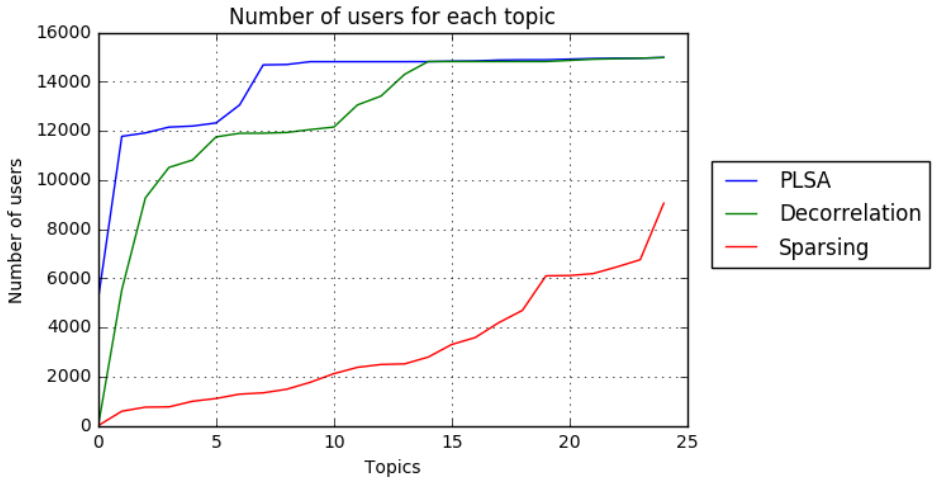
- повышение различности тем
- разреживание $p(t|d)$
- учёт модальностей времени, типа транзакции, терминала

Egorov E., Nikitin F., Goncharov A., Alekseev V., Vorontsov K. Topic Modelling for Extracting Behavioral Patterns from Transactions Data // IC-AIAI 2019.

Построение модели ARTM, 25 тем

- 30 итераций PLSA — без регуляризаторов
- 10 итераций — декоррелирование тем
- 10 итераций — разреживание $p(t|d)$





Декоррелирование Φ и разреживание Θ определяют минимальное число типов экономического поведения каждого клиента, достаточное для описания его расходов.

Пользуюсь картой только чтобы снять наличные

ϕ_{wt},%	MCC-код (категория расходов)
72	Финансовые институты — снятие наличности вручную
27	Финансовые институты — снятие наличности автоматически
0.23	Денежные переводы MasterCard MoneySend
0.1	Денежные переводы
0.012	Финансовые институты — снятие наличности вручную
0.0055	Легковой и грузовой транспорт: продажа, сервис, ремонт, лизинг
0.0027	Магазины игрушек

Наличные + авто, спорт, компьютеры

ϕ_{wt},%	MCC-код (категория расходов)
55	Финансовые институты — снятие наличности автоматически
44	Денежные переводы
0.111	Станции техобслуживания
0.105	Автозапчасти и аксессуары
0.094	Компьютерная сеть/информационные услуги
0.043	Спортивная одежда, одежда для верховой езды и езды на мотоцикле
0.024	Финансовые институты — снятие наличности вручную
0.020	СТО общего назначения
0.018	Горючее топливо — уголь, нефть, разжиженный бензин, дрова
0.015	Магазины мужской и женской одежды
0.015	Финансовые институты — снятие наличности вручную
0.013	Магазины спорттоваров
0.012	Садовые принадлежности (в том числе для ухода за газонами) в розницу
0.011	Паркинги и гаражи
0.011	Бакалейные магазины, супермаркеты
0.010	Различные магазины одежды и аксессуаров

Цивилизованный потребитель: разные магазины, связь, авто

ϕ_{wt},%	MCC-код (категория расходов)
27	Станции техобслуживания
20	Различные продовольственные магазины, рынки, полуфабрикаты
15	Звонки с использованием телефонов, считывающих магнитную ленту
12	Финансовые институты — снятие наличности автоматически
4.7	Горючее топливо — уголь, нефть, разжиженный бензин, дрова
4.1	Универсальные магазины
3.4	Автозапчасти и аксессуары
1.4	Аптеки
1.2	Магазины с продажей спиртных напитков на вынос
1.1	Бакалейные магазины, супермаркеты
0.57	Автошины
0.37	Прямой маркетинг — торговля через каталог
0.35	Товары для дома
0.33	Универмаги
0.32	Плавательные бассейны — распродажа
0.21	Места общественного питания, рестораны

Всего 24 категории с $\phi_{wt} > 0.1\%$; 61 категория с $\phi_{wt} > 0.01\%$

Продвинутые мамки

ϕ_{wt},%	MCC-код (категория расходов)
56	Бакалейные магазины, супермаркеты
8.6	Финансовые институты — снятие наличности автоматически
5.4	Аптеки
4.0	Звонки с использованием телефонов, считывающих магнитную ленту
2.2	Рестораны, закусочные
1.8	Обувные магазины
1.5	Различные продовольственные магазины — рынки, полуфабрикаты
1.4	Магазины спорттоваров
1.4	Детская одежда, включая одежду для самых маленьких
1.3	Магазины игрушек
1.3	Места общественного питания, рестораны
1.1	Магазины мужской и женской одежды
1.1	Магазины с продажей спиртных напитков на вынос
1.1	Магазины косметики
1.0	Садовые принадлежности в розницу
0.73	Одежда для всей семьи

Всего 41 категория с $\phi_{wt} > 0.1\%$; 95 категорий с $\phi_{wt} > 0.01\%$

Бизнес-леди: забыла про наличку — всё по карте

ϕ_{wt},%	MCC-код (категория расходов)
12	Магазины мужской и женской одежды
7.3	Оборудование, мебель и бытовые принадлежности
7.0	Места общественного питания, рестораны
5.6	Магазины по продаже часов, ювелирных изделий и изделий из серебра
5.3	Обувные магазины
4.7	Магазины косметики
4.6	Одежда для всей семьи
3.8	Универмаги
3.2	Готовая женская одежда
2.8	Практикующие врачи, медицинские услуги
1.8	Прямой маркетинг — торговля через каталог
1.5	Салоны красоты и парикмахерские
1.3	Детская одежда, включая одежду для самых маленьких
1.3	Аптеки
1.0	Изготовление и продажа меховых изделий
1.0	Центры здоровья

Всего 70 категорий с $\phi_{wt} > 0.1\%$; 134 категории с $\phi_{wt} > 0.01\%$

Бизнес-класс: авиа, отели, казино, рестораны, ценные бумаги

ϕ_{wt},%	MCC-код (категория расходов)
28	Авиалинии, авиакомпании
19	Финансовые институты — торговля и услуги
9.5	Отели, мотели, базы отдыха, сервисы бронирования
8.6	Транзакции по азартным играм (плюс)
5.2	Финансовые институты — торговля и услуги
3.2	Места общественного питания, рестораны
3.1	Не-финансовые институты: ин.валюта, переводы, дорожн.чеки, квази-кэш
2.2	Пассажирские железнодорожные перевозки
1.7	Бизнес-сервис
1.4	Жилье — отели, мотели, курорты
1.3	Галереи/учреждения видеоигр
1.3	Транзакции по азартным играм (минус)
0.6	Ценные бумаги: брокеры/дилеры
0.5	Туристические агентства и организаторы экскурсий
0.3	Лимузины и такси
0.3	Беспошлинные магазины Duty Free

Всего 50 категорий с $\phi_{wt} > 0.1\%$; 103 категории с $\phi_{wt} > 0.01\%$

Провинциальный малый бизнес

ϕ_{wt},%	MCC-код (категория расходов)
27	Финансовые институты — снятие наличности автоматически
8.5	Лесо- и строительный материал
8.4	Бытовое оборудование
6.6	Плавательные бассейны — распродажа
5.5	Продажа электронного оборудования
4.1	Бакалейные магазины, супермаркеты
3.3	Универсальные магазины
3.0	Садовые принадлежности в розницу
2.6	Телекоммуникационное оборудование, включая продажу телефонов
2.4	Легковой и грузовой транспорт: продажа, сервис, ремонт, лизинг
2.2	Товары для дома
2.1	Пассажирские железнодорожные перевозки
1.5	Оборудование, мебель и бытовые принадлежности
1.3	Скобяные товары в розницу
1.2	Магазины спорттоваров
1.1	Аптеки

Всего 54 категории с $\phi_{wt} > 0.1\%$; 104 категории с $\phi_{wt} > 0.01\%$

Анализ транзакций корпоративных клиентов банка

Данные:
лесная отрасль, 2016 г., 10.7M транзакций, 1M компаний.
Транзакция — это тройка $\langle$покупатель, продавец, текст$\rangle$.

Некоторые *тексты* платёжных поручений (далеко не все!) содержат названия товаров и услуг.

Документ — это история транзакций одной компании

Семь модальностей:

- компании: поставщики / покупатели
- слова в платёжных поручениях: поставщики / покупатели
- ОКВЭДы данной компании
- ОКВЭДы контрагентов: поставщики / покупатели

Примеры тем — видов деятельности компаний

покупка	продажа
0.11: услуга	0.12: лдсп
0.07: классик	0.08: дсп
0.05: дрова	0.03: мдф
0.05: пиловочник	0.03: поставка
0.05: материал	0.02: услуга
0.03: порода	0.02: охранный
0.03: лесоматериал	0.02: ламинировать
0.03: сертум	0.02: хдф
0.02: хвойный	0.02: материал
0.01: дерево	0.01: накл
0.01: транспортный	0.01: товар

покупка	продажа
0.19: право	0.16: арендный
0.17: сбис	0.10: часть
0.16: использование	0.08: плата
0.03: аккаунт	0.04: минимальный
0.02: электронный	0.04: участок
0.02: лицевой	0.04: использование
0.02: устный	0.02: земля
0.01: устройство	0.02: лесничество
0.01: генерация	0.02: земельный
0.01: хранение	0.01: фонд
0.01: ключевой	0.01: федеральный

Примеры тем — видов деятельности компаний

покупка		продажа
0.09: ткань		0.16: мебель
0.09: поставка		0.05: плёнка
0.02: мебельный		0.04: стул
0.02: деревянный		0.03: кресло
0.02: транспортный		0.03: изделие
0.02: фанера		0.02: краска
0.02: поролон		0.02: фанера
0.01: механизм		0.01: лкм
0.01: плата		0.01: лакокрасочный
0.01: частичный		0.01: лак
		0.01: материал
		0.01: клеить

покупка		продажа
0.06: лдсп		0.37: товар
0.05: фурнитура		0.15: мебель
0.02: плёнка		0.04: поставка
0.02: материал		0.04: накладный
0.02: мебельный		0.03: накл
0.02: стекло		0.03: рубль
0.02: мдф		
0.02: кромка		
0.01: транспортный		
0.01: клеить		
0.01: профиль		
0.01: пвх		

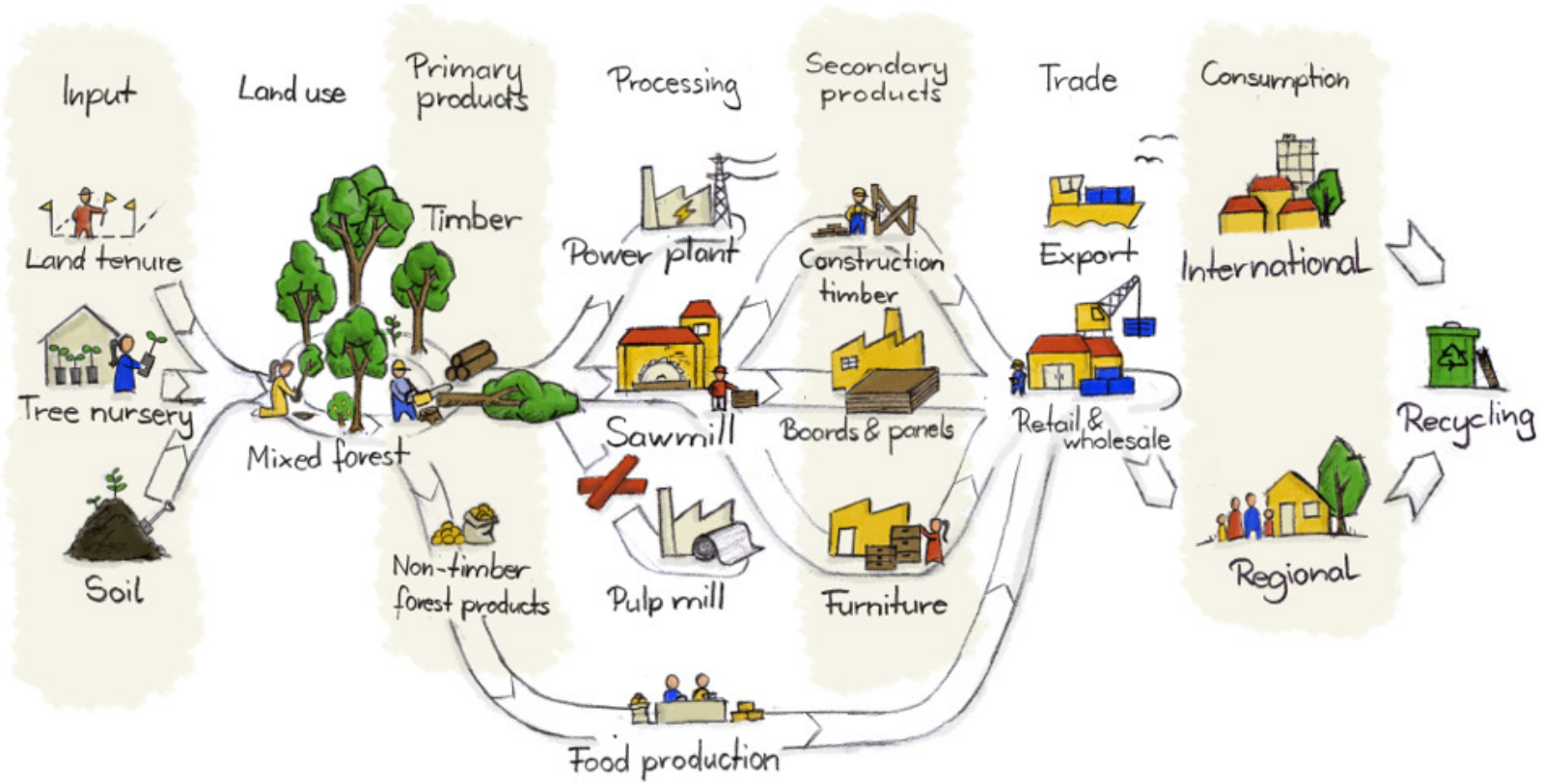
Примеры тем — видов деятельности компаний

покупка		продажа
0.52: гсм		0.14: вывоз
0.43: далее		0.09: тбо
		0.04: мусор
		0.03: отход
		0.02: утилизация
		0.01: тко

покупка		продажа
0.19: налог		0.11: бумага
0.06: услуга		0.08: гофроящик
0.04: макулатура		0.04: гофрокартон
0.03: поставка		0.03: гофрокороб
0.03: транспортный		0.03: поставка
0.02: лесопродукция		0.03: фактура
0.02: автоуслуга		0.02: гофропродукция
0.01: перевозка		0.02: гофротару
0.01: плата		0.02: гофрирование
		0.02: гофролоток
		0.02: товар
		0.01: лоток

Цели тематического моделирования банковских данных

- Получение векторных представлений компаний
- Поиск схожих и конкурирующих компаний
- Восстановление структуры товарных потоков отрасли



Резюме

- ARTM — это «много поучительных регуляризаторов»
- ARTM существенно проще методов байесовского вывода
- Теория ARTM оказалась «теорией одной Леммы»
- С её помощью легко выводятся разнообразные модели:
  — двухматричные (классика)
  — трёхматричные (не было времени показать)
  — мультимодальные (для разнородных данных)
  — гиперграфовые (для транзакционных данных)
  — одноматричные (для быстрой векторизации документов)
- **Открытая проблема:** исследовать свойства сходимости итерационного процесса $\omega_j = \text{norm}\left(\omega_j \otimes \frac{\partial f}{\partial \omega_j}\right)$

Воронцов К.В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM. http://www.MachineLearning.ru, 2020.

http:bigartm.org BigARTM — эффективная открытая реализация

http:github.com/machine-intelligence-laboratory/TopicNet
TopicNet — обёртка над BigARTM для подбора и анализа моделей