Задачи и технологии вероятностного тематического моделирования (к обсуждению возможностей их применения в биоинформатике)

Воронцов Константин Вячеславович
д.ф.-м.н., профессор РАН, руководитель научной группы
«Машинное обучение и семантический анализ»

Институт искусственного интеллекта
МГУ

совместный семинар научных групп
«Машинное обучение и семантический анализ»
и «ИИ в биоинформатике и медицине»
• 17 мая 2023 •

Содержание

1. **Задачи тематического моделирования**
   - Математическая постановка задачи
   - Интерпретируемость и цели моделирования
   - Примеры задач
2. **Методы и инструменты**
   - Аддитивная регуляризация (ARTM)
   - Библиотека BigARTM
   - Средства визуализации
3. **Приложения тематического моделирования**
   - Анализ текстов и информационный поиск
   - Нетекстовые приложения
   - Задачи в области медицины и биоинформатики

Тематическое моделирование: «о чём все эти тексты?»

Дано:

- коллекция текстовых документов D, словарь W
- n_{dw} — частота слов (термов) $w \in W$ в документе $d \in D$
- $|T|$ — сколько тем хотим найти в коллекции D

Найти:

- $p(w|t) = \phi_{wt}$ — вероятности слов w в каждой теме t
- $p(t|d) = \theta_{td}$ — вероятности тем t в каждом документе d
- $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$ — тематическую языковую модель

Критерий: правдоподобие предсказания слов w в документах d с дополнительными критериями-регуляризаторами $R_i(\Phi, \Theta)$:

$$\sum_{d \in D} \sum_{w \in d} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

Критерий 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

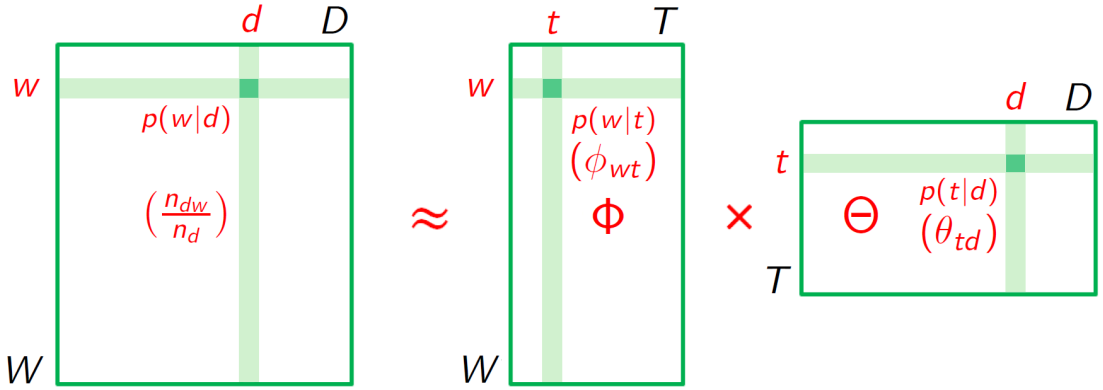
$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

при ограничениях неотрицательности и нормировки столбцов (такие матрицы Φ, Θ называются *стохастическими*)

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

Три интерпретации задачи тематического моделирования

1. **Мягкая кластеризация** документов по кластерам-темам
2. Низкоранговое стохастическое **матричное разложение:**



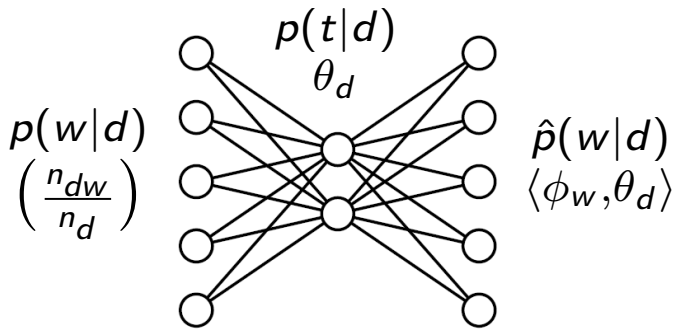
3. **Автокодировщик** документов в тематические эмбединги:

кодировщик $f_\Phi : \frac{n_{dw}}{n_d} \to \theta_d$

декодировщик $g_\Phi : \theta_d \to \Phi\theta_d$

задача реконструкции текстов:

$$\sum_d \mathrm{KL}\left(\frac{n_{dw}}{n_d} \,\middle\|\, \langle \phi_w, \theta_d \rangle\right) \to \min_{\Phi,\Theta}$$



Свойство интерпретируемости тематических моделей

Тематическая модель формирует тематические векторы:

- $p(t|d) = \theta_{td}$ для каждого документа d
- $p(t|w) = \dfrac{p(w|t)p(t)}{p(w)} = \phi_{wt}\dfrac{n_t}{n_w}$ для каждого терма w
- $p(t|d,w)$ для каждого локального контекста (d,w)

Интерпретируемость тематических векторов:

- каждая тема t описывается *семантическим ядром* — частотным словарём слов $\left\{w\colon p(w|t) > \gamma p(w)\right\}$
- тема может «рассказать о себе» словами или фразами
- любой объект x с вектором $p(t|x)$ описывается частотным словарём слов $\left\{w\colon p(w|x) = \sum\limits_{t\in T} p(w|t)p(t|x) > \gamma p(w)\right\}$

Цели и не-цели тематического моделирования

Цели:

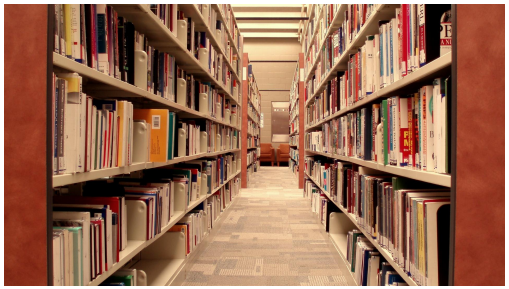
- Выяснить тематическую кластерную структуру текстовой коллекции, сколько в ней тем и о чём они
- Получать интерпретируемые тематические векторные представления (эмбединги) документов, фрагментов, слов $p(t|d)$, $p(t|w)$, $p(t|d,w)$ и нетекстовых объектов $p(t|x)$
- Решать задачи поиска, категоризации, сегментации, суммаризации с помощью тематических эмбедингов

Не-цели:

- Угадывать следующие слова (TM слабы как модели языка)
- Генерировать связный текст
- Понимать смысл текста

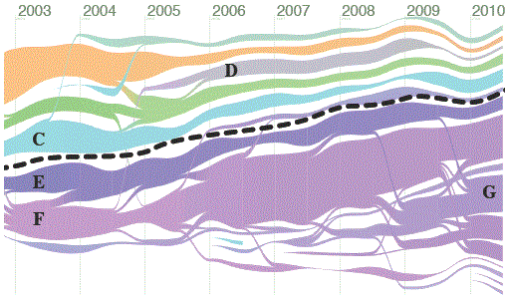
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

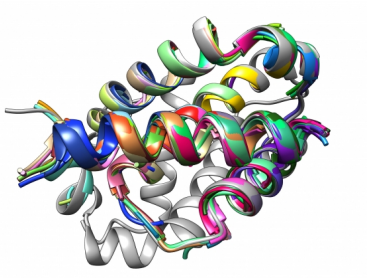


поиск тематического контента в соцсетях

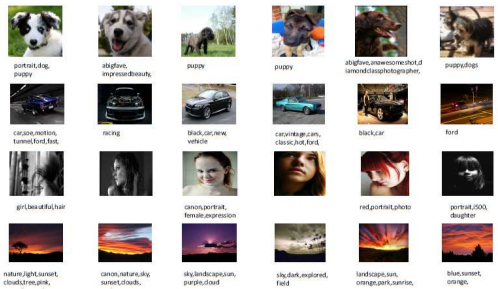
выявление и отслеживание цепочек новостей



паттерны биологических последовательностей



мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

J.Boyd-Graber, Yuening Hu, D.Mimno. Applications of Topic Models. 2017.

H.Jelodar et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. 2019.

Пример 1. Мультиязычная тематическая 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.

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

Пример 1. Мультиязычная тематическая модель Википедии

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

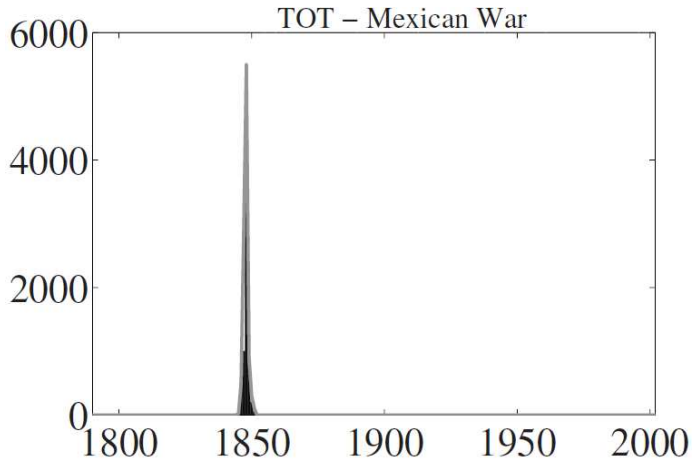
Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	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

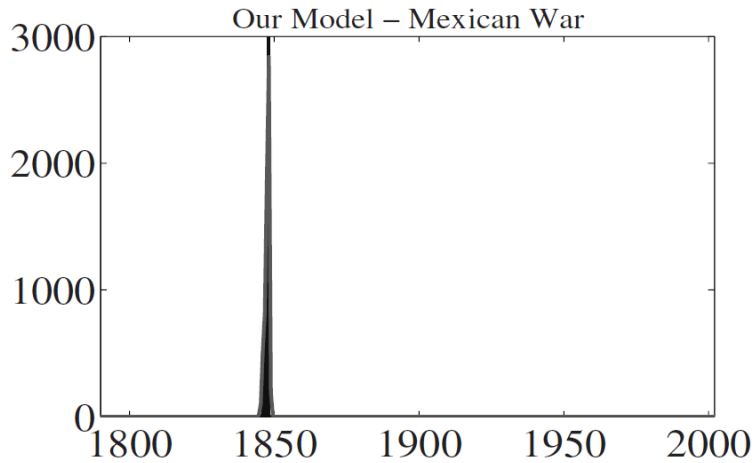
Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример 2. Совмещение темпоральной и n-граммной модели

По коллекции выступлений президентов США





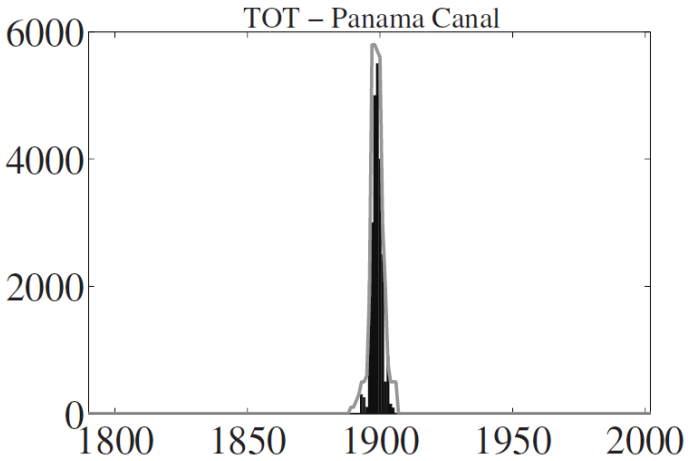
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

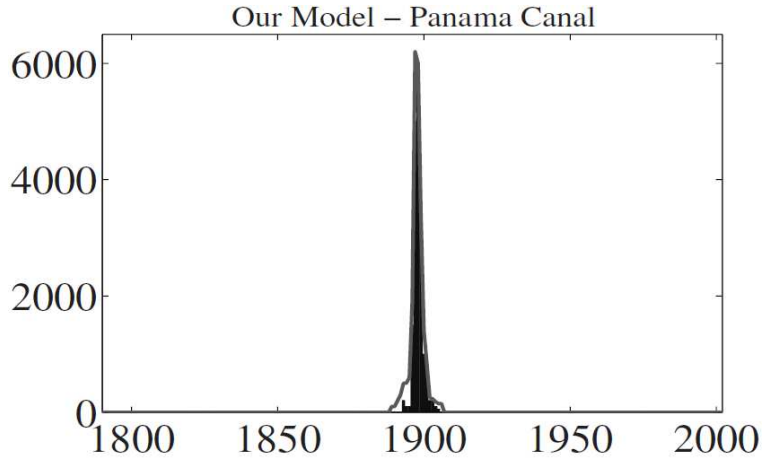
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Пример 2. Совмещение темпоральной и n-граммной модели

По коллекции выступлений президентов США





1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Модель PLSA (Probabilistic Latent Semantic Analysis)

Максимизация log-правдоподобия для стохастических матриц:

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

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

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_d n_{dw} p_{tdw}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_w n_{dw} p_{tdw}\right) \end{cases}$$

где $\underset{t\in T}{\text{norm}}(x_t) = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — операция нормировки вектора.

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Модель LDA (Latent Dirichlet Allocation)

Максимизация log-правдоподобия + байесовская регуляризация с априорными распределениями Дирихле на столбцы Φ, Θ:

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

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

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

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

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

Аддитивная Регуляризация Тематических Моделей (ARTM)

Максимизация log правдоподобия с регуляризатором R:

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

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

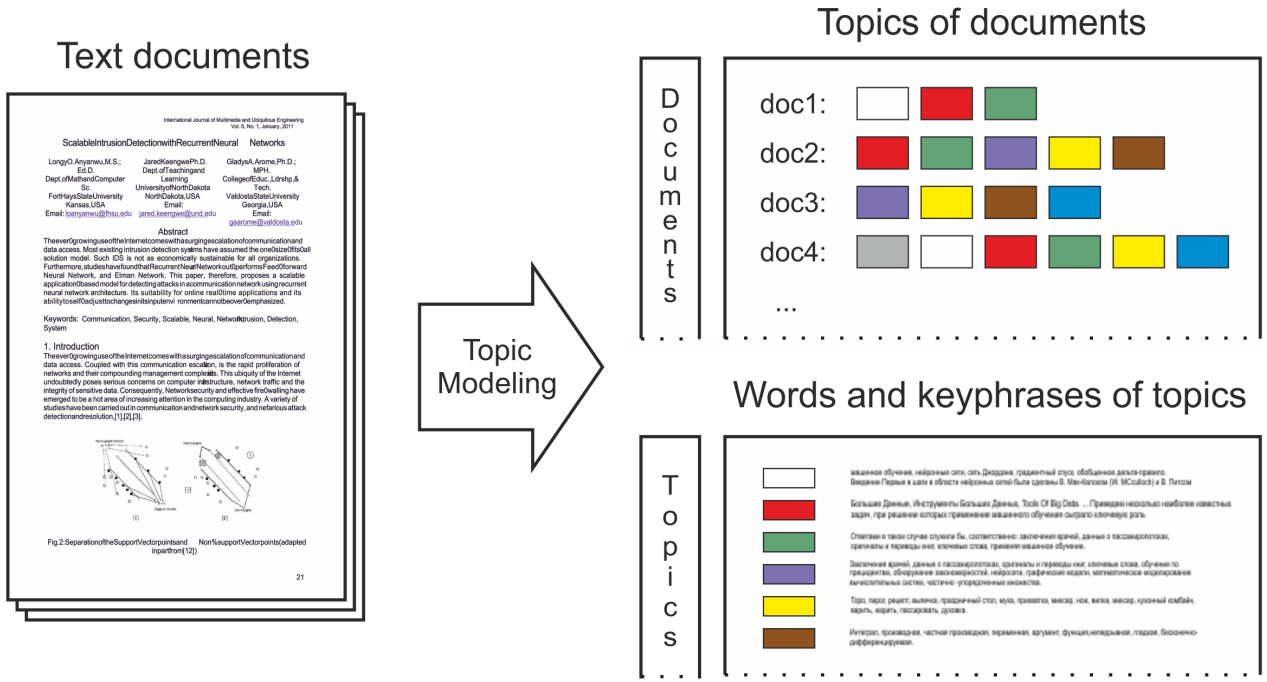
$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

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

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

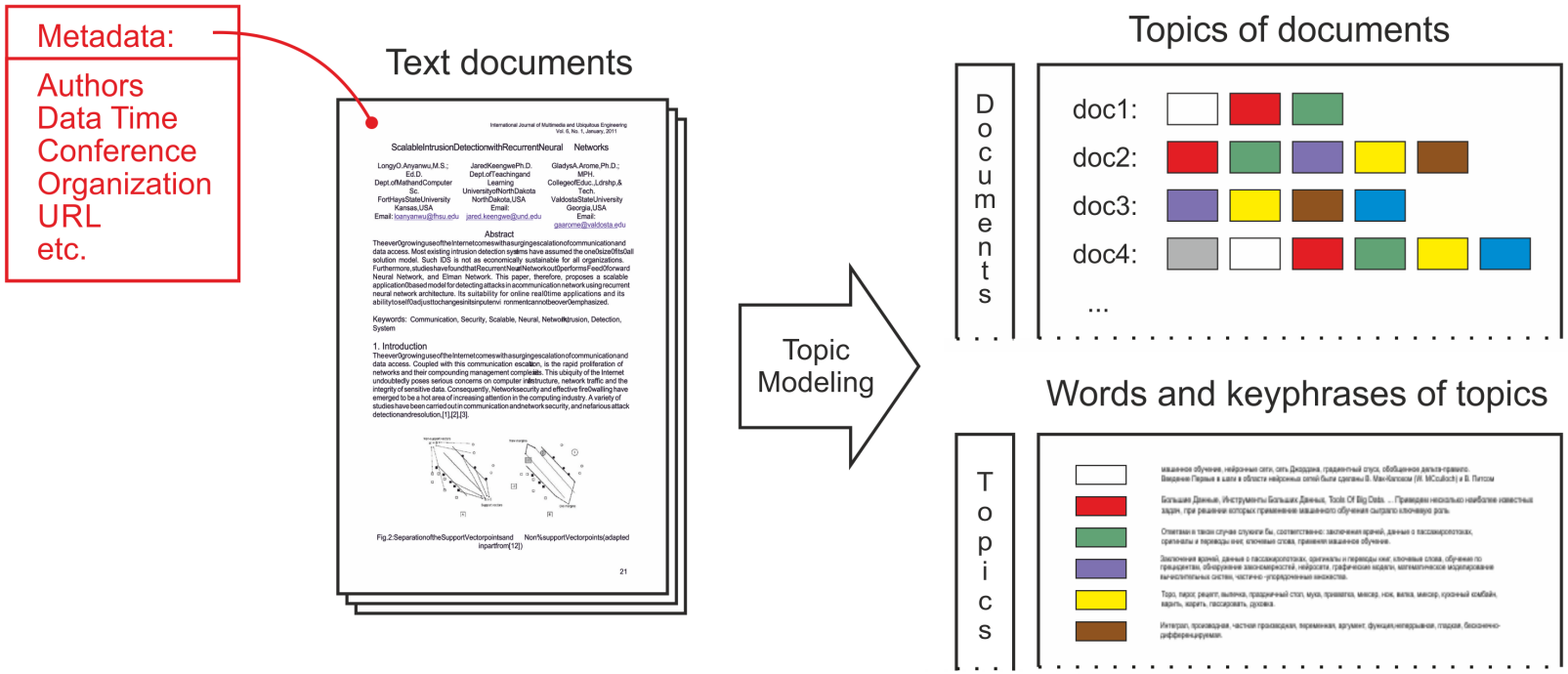
Мультимодальная тематическая модель

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



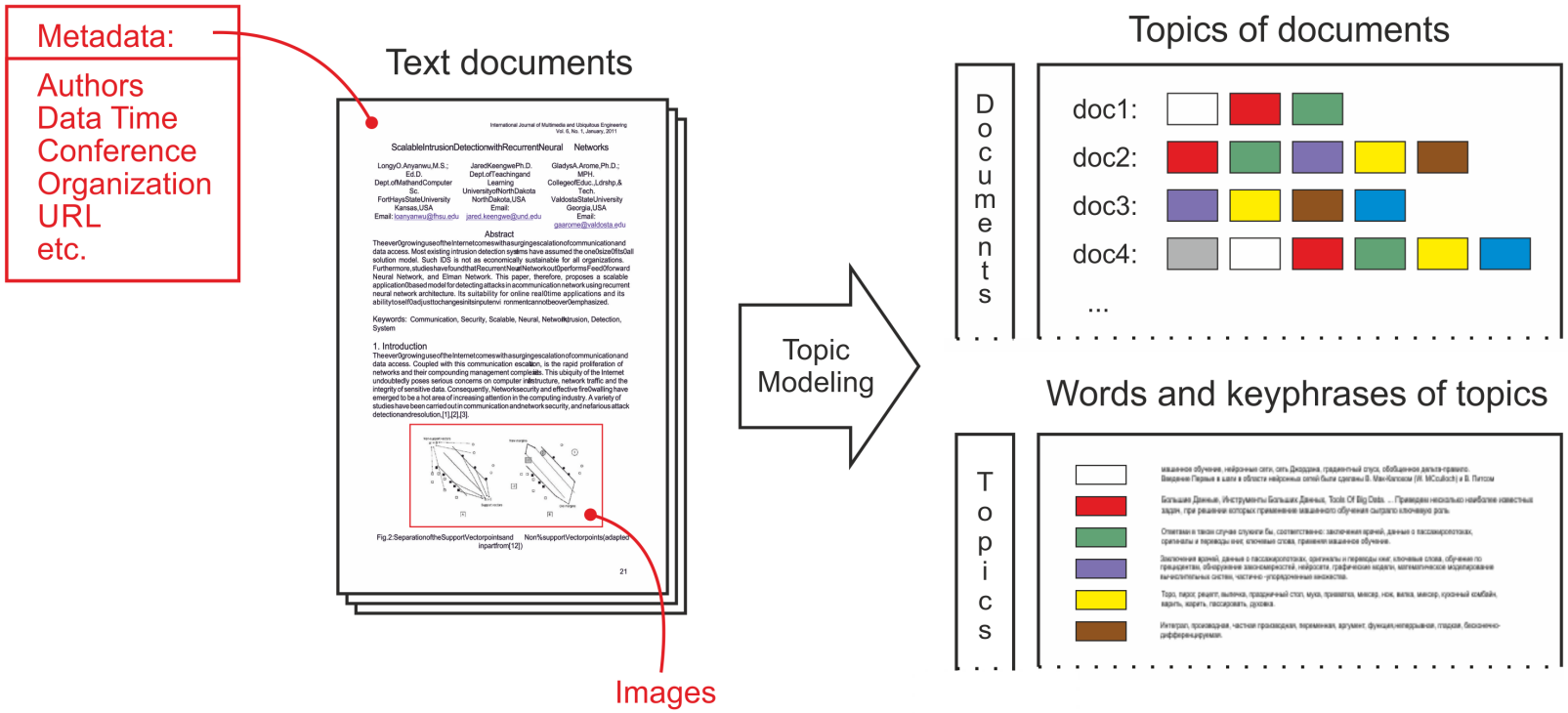
Мультимодальная тематическая модель

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



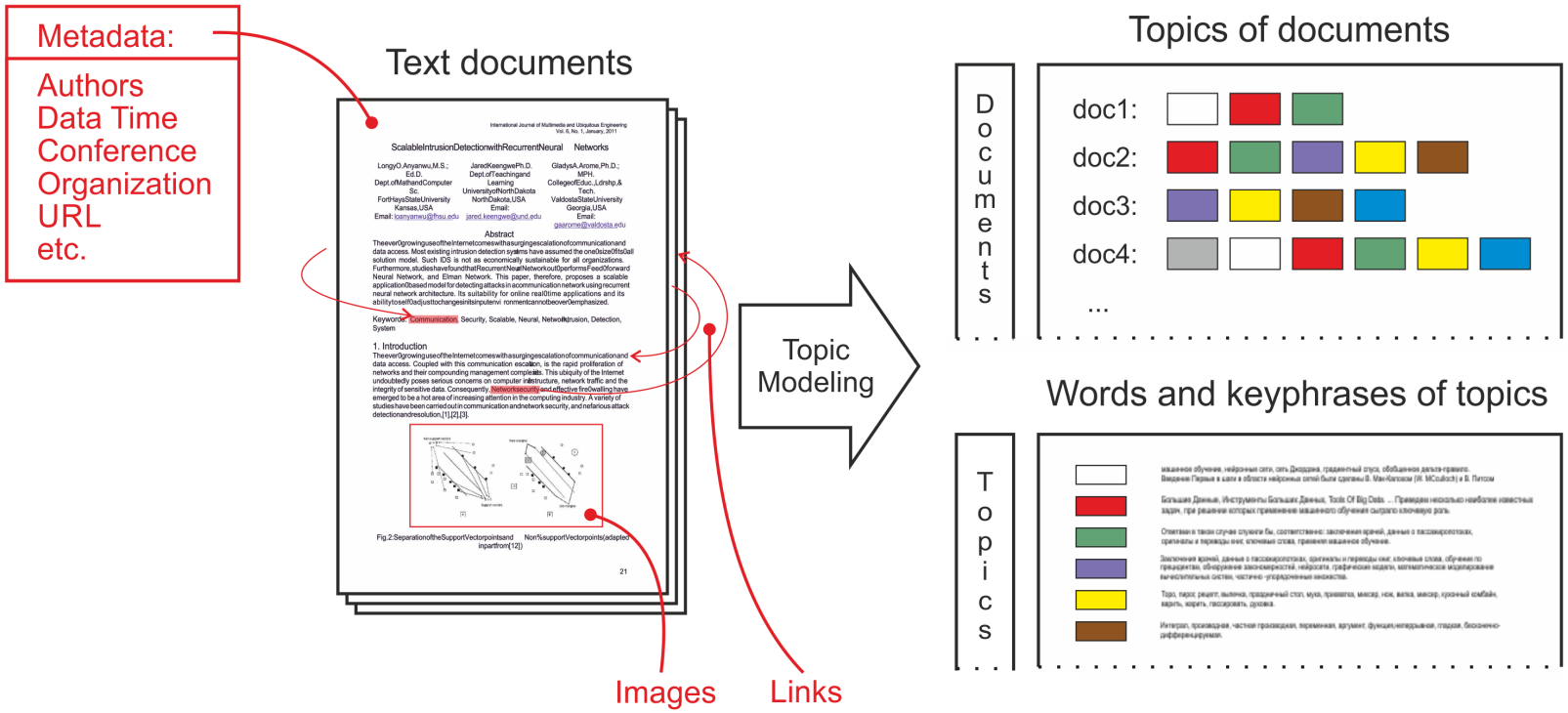
Мультимодальная тематическая модель

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



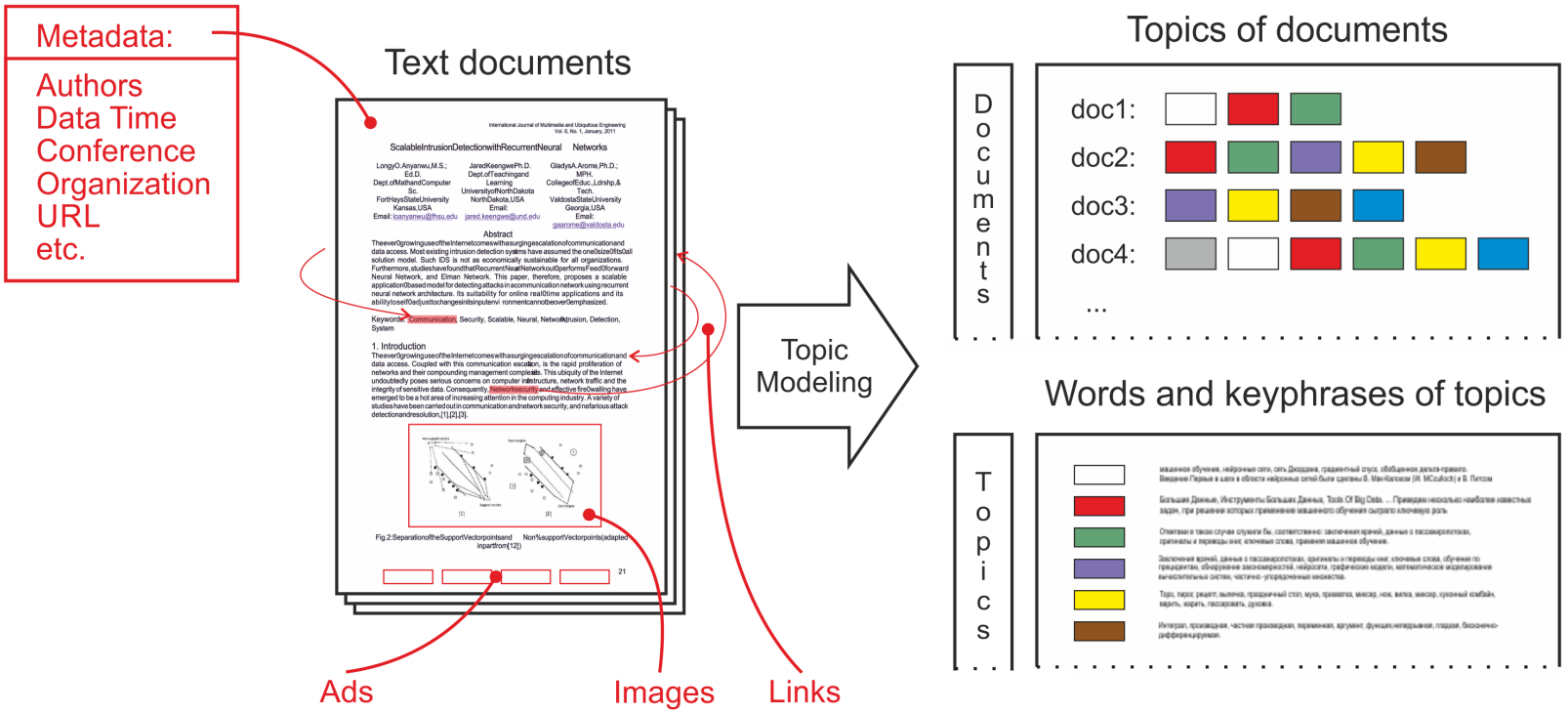
Мультимодальная тематическая модель

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



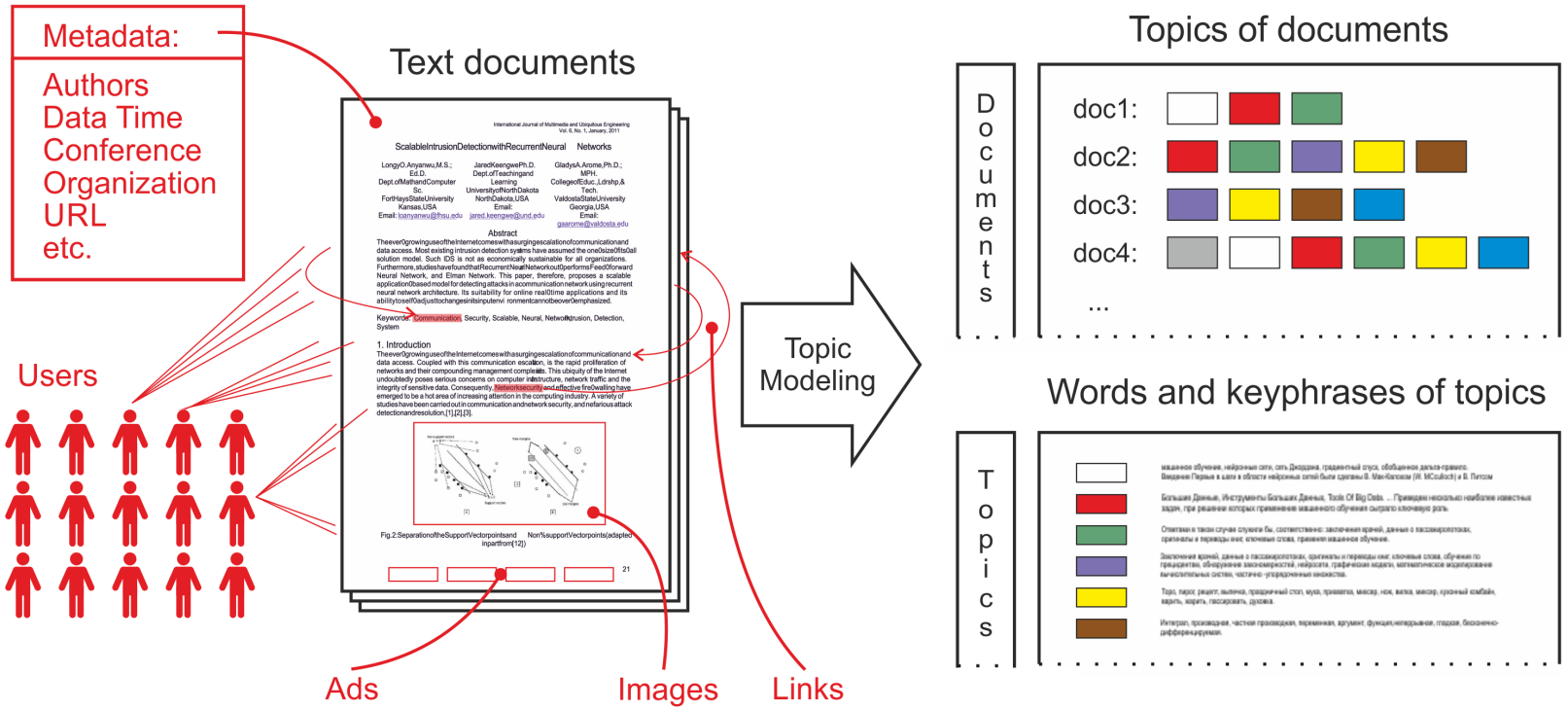
Мультимодальная тематическая модель

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



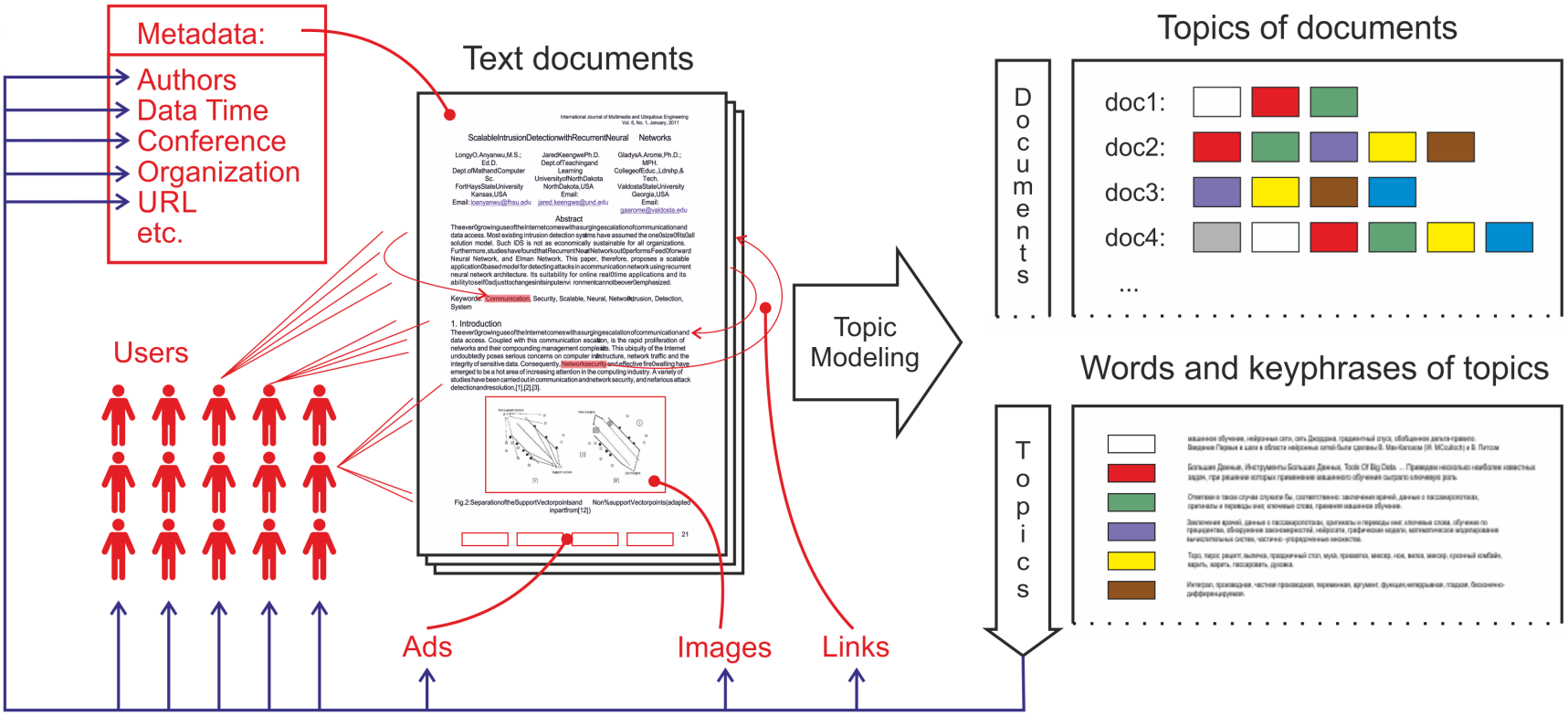
Мультимодальная тематическая модель

Тема t может содержать 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 t может содержать 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.

Транзакционные данные

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

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

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

Гиперграфовая транзакционная ARTM

n_{kdx} — частота транзакции (d, x), $x \subset W$ типа k в выборке E_k

Максимизация суммы log-правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

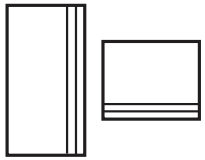
$$\text{E-шаг:} \quad p_{tdx} = \operatorname*{norm}_{t \in T} \Big(\theta_{td} \prod_{v \in x} \phi_{vt} \Big)$$

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

K.Vorontsov. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization // Springer Optimization and Its Applications. 2023

Регуляризаторы для улучшения интерпретируемости тем

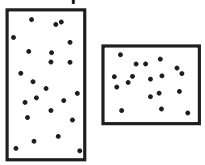
background



Сглаживание фоновых тем $B \subset T$:

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

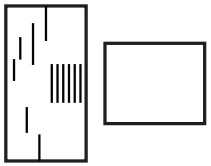
sparse



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

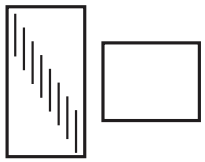
$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d} \sum_{t \in S} \alpha_t \ln \theta_{td}$$

seed words



Сглаживание для выделения релевантных тем
с помощью словаря «затравочных» ключевых слов

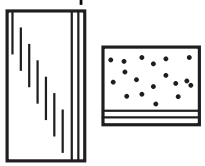
decorrelated



Декоррелирование для повышения различности тем:

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

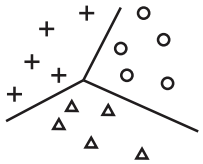
interpretable



Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

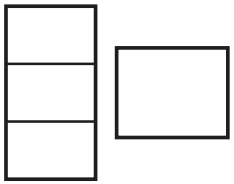
Регуляризаторы для мультимодальных тематических моделей

supervised



Модальности меток классов или категорий для задач классификации и категоризации текстов.

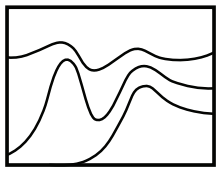
multilanguage



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

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

temporal



Темпоральные модели с модальностью времени i:

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

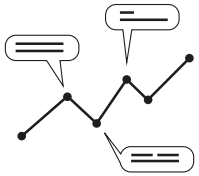
geospatial

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

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

Регуляризаторы для учёта дополнительной информации

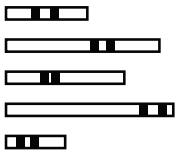
regression



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

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

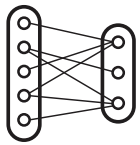
biterm



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

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

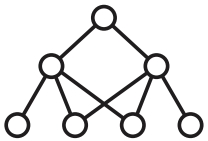
relational



Связи или ссылки между документами:

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

hierarchy



Связи родительских тем t с дочерними подтемами s:

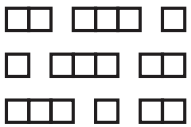
$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st}$$

Регуляризаторы для моделирования последовательного текста

sentence

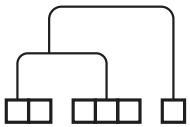
Тематические модели, учитывающие границы предложений, абзацев и секций документов

n-gram



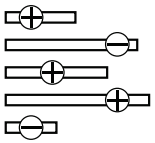
Модели с модальностями *n*-грамм, коллокаций, именованных сущностей (используем TopMine)

syntax



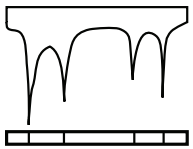
Модели, учитывающие результаты автоматического синтаксического разбора (используем UDPipe)

sentiment



Модели выделения мнений на основе тональностей, фактов, семантических ролей именованных сущностей

segmentation



Тематические модели сегментации с автоматическим определением границ сегментов

Модульный подход к синтезу моделей с заданными свойствами

Для построения композитных моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».

Этапы моделирования	Bayesian TM	ARTM	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная модель порождения данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Единый регуляризованный EM-алгоритм для любых моделей и их композиций	
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
- Онлайновый параллельный мультимодальный ARTM
- Встроенная библиотека регуляризаторов и метрик 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

Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

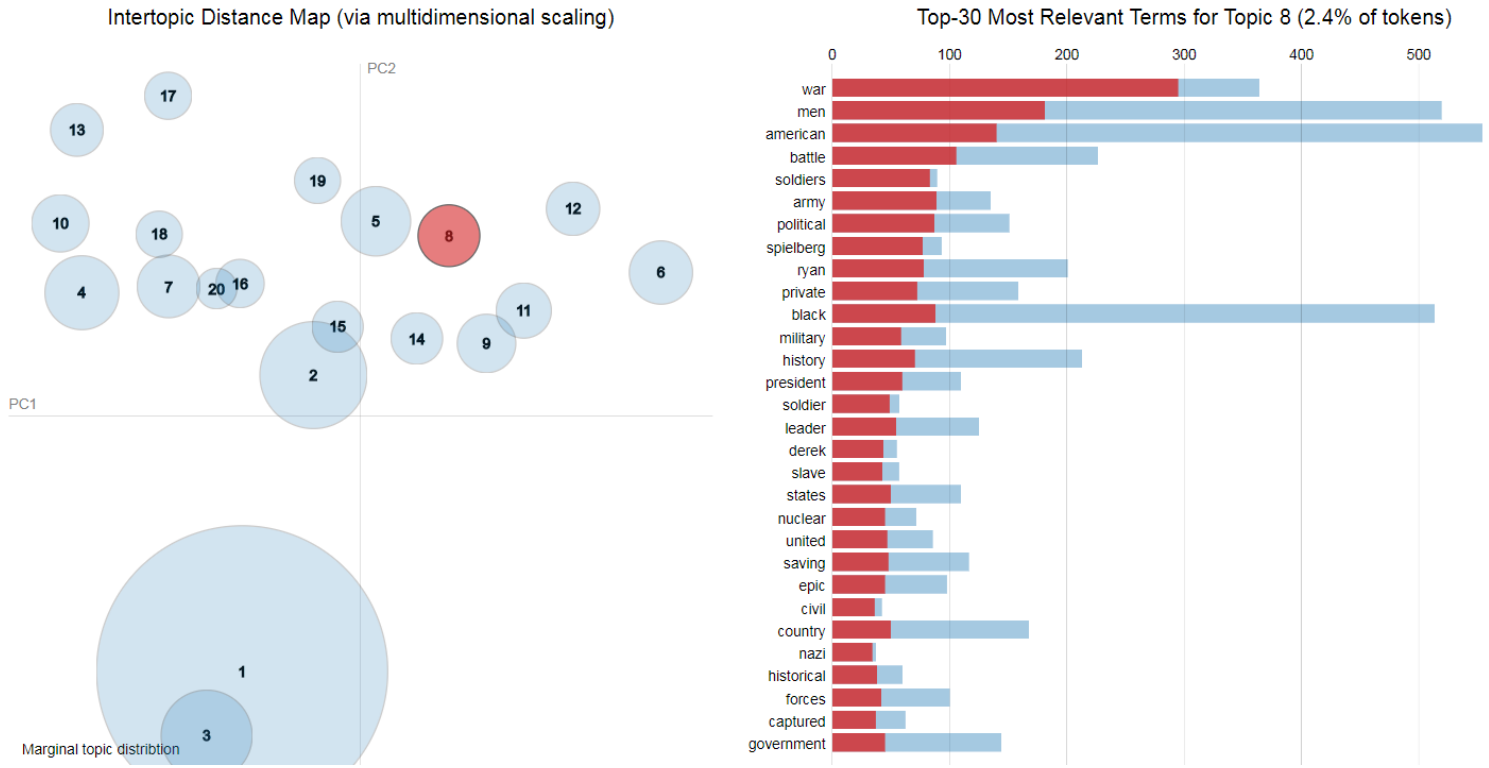
3.7M статей Википедии, 100K слов: время min (перплексия)

проц.	$\|T\|$	Gensim	Vowpal Wabbit	BigARTM	BigARTM асинхрон
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov.
Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

Система LDAvis

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

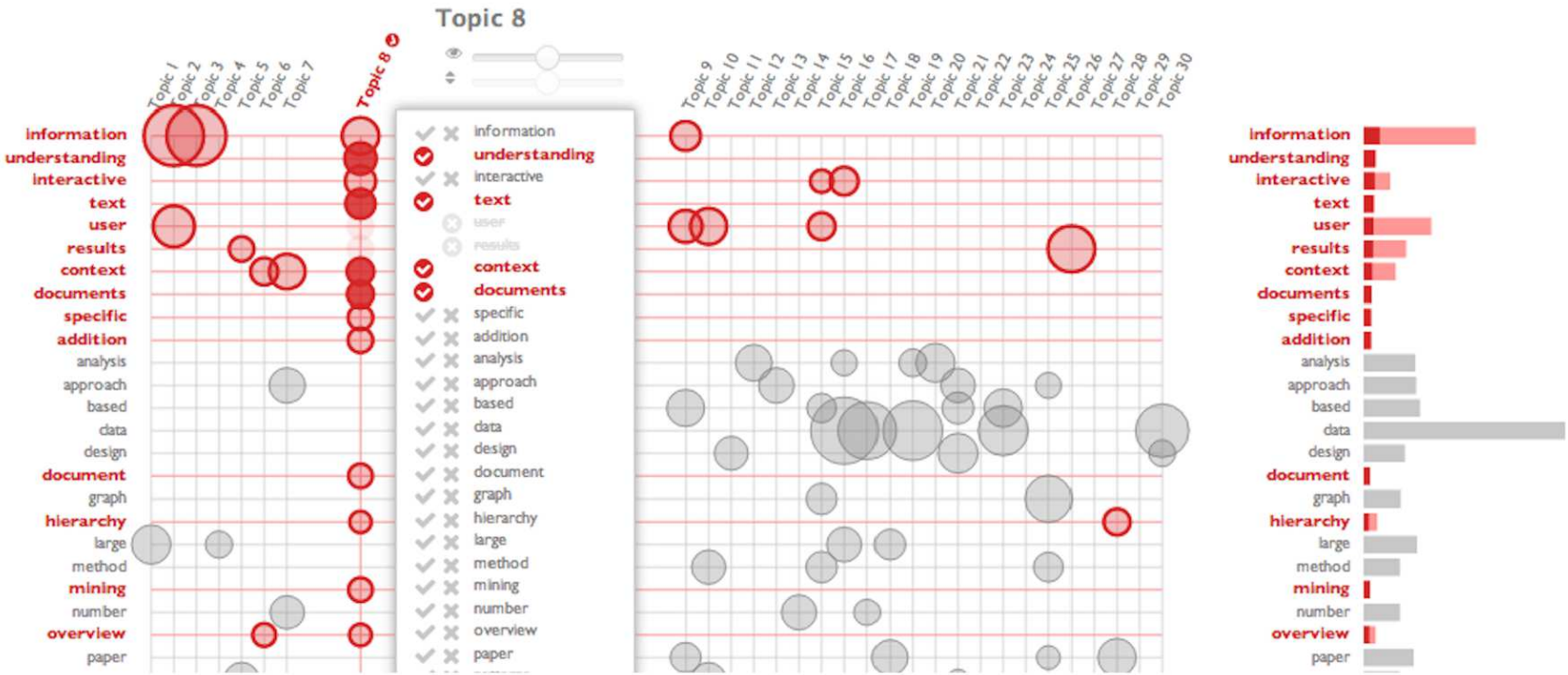


https://github.com/cpsievert/LDAvis

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

Система Termite

Интерактивная визуализация матрицы Φ и сравнение тем:

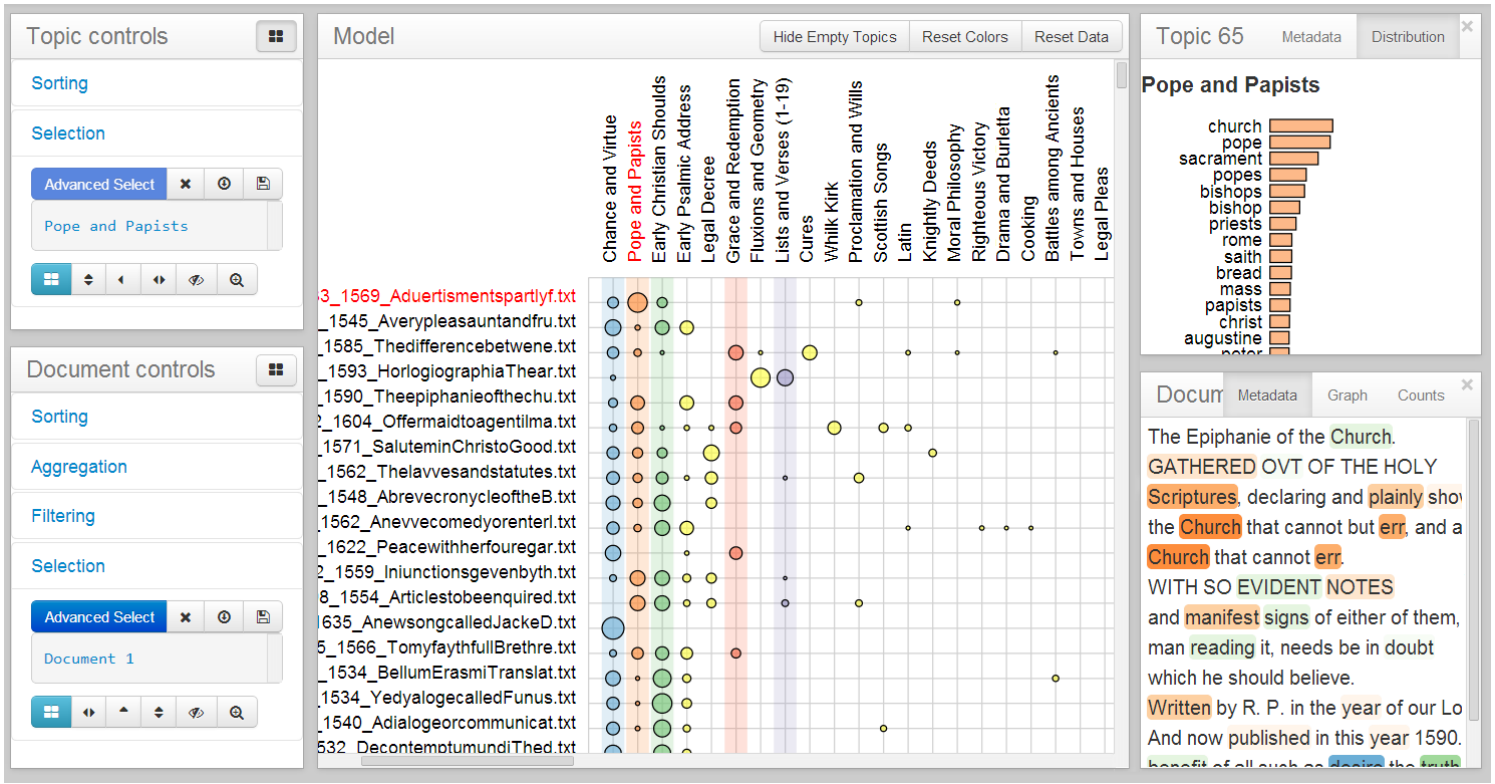


https://github.com/uwdata/termite-visualizations

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

Система Serendip

Визуализация матриц Φ, Θ и тематики слов в текстах:



http://vep.cs.wisc.edu/serendip

E.Alexander, J.Kohlmann, R.Valenza, M.Witmore, M.Gleicher. Serendip: Topic Model-Driven Visual Exploration of Text Corpora. IEEE VAST 2014.

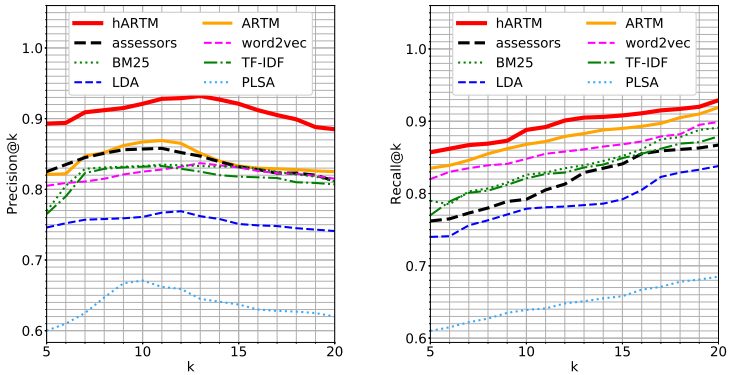
Источники вдохновения: http://textvis.lnu.se

Интерактивный обзор 440 средств визуализации текстов

Shixia Liu, Weiwei Cui, Yingcai Wu, Mengchen Liu. A survey on information visualization: recent advances and challenges. 2014.

Разведочный поиск в технологических блогах

Цель: поиск документов
по длинным текстовым запросам
— Habr.ru (175K документов),
— TechCrunch.com (760K док.).



Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{hierarchy}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) \to \max$$

Результаты:

- Точность и полнота 93%, превосходит асессоров и другие методы (tf-idf, BM25, word2vec, PLSA, LDA, ARTM).
- Увеличилась оптимальная размерность векторов:
  200 → 1400 (Habr.ru), 475 → 2800 (TechCrunch.com).

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

Поиск и классификация этно-релевантных тем в соцсетях

Цель: выявление как можно большего числа тем о национальностях и межнациональных отношениях (по словарю из 300 этнонимов).

(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, сообщать, океан, станция, хатико, район, правительство, атомный,
(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,
(венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадуро, боливия, фидель, глава, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,
(китайцы): китайский, россия, производство, китай, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,
(азербайджанцы): русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, москва, страна, армянин, слово, рынок,
(грузины): грузинский, спецназ, военный, август, баташева, российский, спецназовец, миротворец, операция, румын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвале,
(осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алахай, российский, население, конфликт,
(цыгане): наркотик, цыган, цыганка, хороший, место, страна, деньга, время, работать, жизнь, жить, рука, дом, цыганский, наркоманка,

Регуляризаторы:

$$\mathscr{L}\left(\underset{\Phi}{\square}\,\underset{\Theta}{\square}\right) + R\left(\square\right) + R\left(\square\right) + R\left(\square\right) + R\left(\square\right) + R\left(\square\right) + R\left(\square\right) + R\left(\square\right) \to \max$$

PLSA; seed words; interpretable; multimodal; temporal; geospatial; sentiment

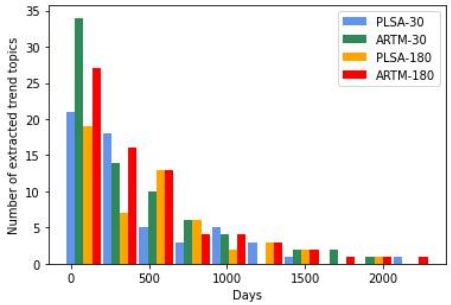
Результаты: число релевантных тем: 45 (LDA) $\to$ 83 (ARTM).

M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI, 2016.

Mining ethnic content online with additively regularized topic models. 2016.

Выявление трендов в коллекции научных публикаций

Цель: раннее обнаружение трендовых тем с начальным экспоненциальным ростом; проверка модели на трендах в области AI/ML 2009–2021 гг.



Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{interpretable}\right) + R\left(\text{dynamic}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) \to \max$$

Результаты:

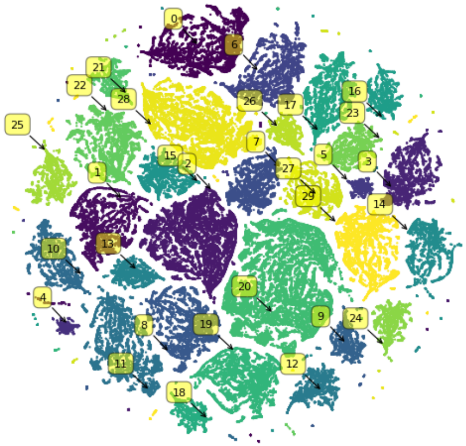
- выделение 90 из 91 тренда в области машинного обучения
- 63% тем выделяется за год, 79% за два года

Н.Герасименко, А.Чернявский, М.Никифорова, М.Никитин, К.Воронцов.
Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях. Доклады РАН, 2022.

Тематическая модель банковских транзакционных данных

Цель: Выявление паттернов потребительского поведения клиентов банка, причём

- документы → клиенты,
- слова → MCC-коды продавцов.



Регуляризаторы:

$$\mathscr{L}\left(\underset{\Phi\ \Theta}{\text{PLSA}}\right) + R\left(\underset{}{\text{interpretable}}\right) + R\left(\underset{}{\text{multimodal}}\right) + R\left(\underset{}{\text{supervised}}\right) \to \max$$

Результаты:

- темы — паттерны потребительского поведения
- предсказание пола, возраста, достатка клиентов

E.Egorov, F.Nikitin, A.Goncharov, V.Alekseev, K.Vorontsov. Topic modelling for extracting behavioral patterns from transactions data. 2019.

Поиск обсуждений болезней в социальных сетях

Цель: поиск сообщений в Twitter о болезнях, симптомах, способах лечения, побочных эффектах.

Регуляризаторы:

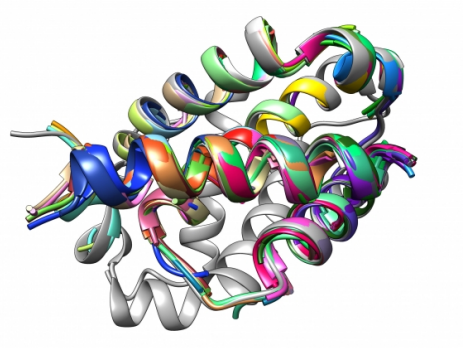
$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{seed words}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{hierarchy}\right) + R\left(\text{temporal}\right) + R\left(\text{geospatial}\right) \to \max$$

Модель ATAM (Ailment Topic Aspect Model) похожа на поиск этно-релевантных тем и легко реализуема в BigARTM

M.J.Paul, M.Dredze. Discovering Health Topics in Social Media Using Topic Models, 2014

Обработка последовательностей нуклеотидов или аминокислот

Цель: поиск мотивов и предсказание функций по нуклеотидным или аминокислотным последовательностям.



Регуляризаторы (гипотеза):

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{seed words}\right) + R\left(\text{interpretable}\right) +$$
$$+ R\left(\text{multimodal}\right) + R\left(\text{hierarchy}\right) + R\left(\text{n-gram}\right) + R\left(\text{segmentation}\right) \to \max$$

Такая модель легко реализуема в BigARTM.

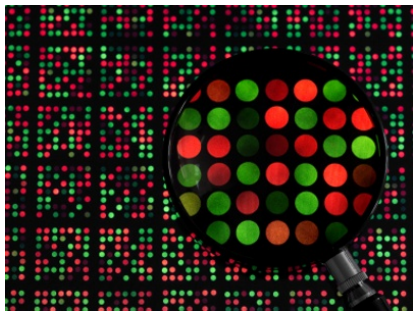
J.B.Gutierrez, K.Nakai. A study on the application of topic models to motif finding algorithms. 2016.

Lin Liu, Lin Tang, Libo He, Shaowen Yao, Wei Zhou. Predicting protein function via multi-label supervised topic model on gene ontology. 2017.

Lin Liu, Lin Tang, Xin Jin, Wei Zhou. A multi-label supervised topic model conditioned on arbitrary features for gene function prediction. 2019

Обработка данных ДНК-микрочипов

Цель:
кластеризация или классификация генов без предположения о функциональной независимости между генами.



Регуляризаторы (гипотеза):

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{supervised}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{hierarchy}\right) + R\left(\text{n-gram}\right) \to \max$$

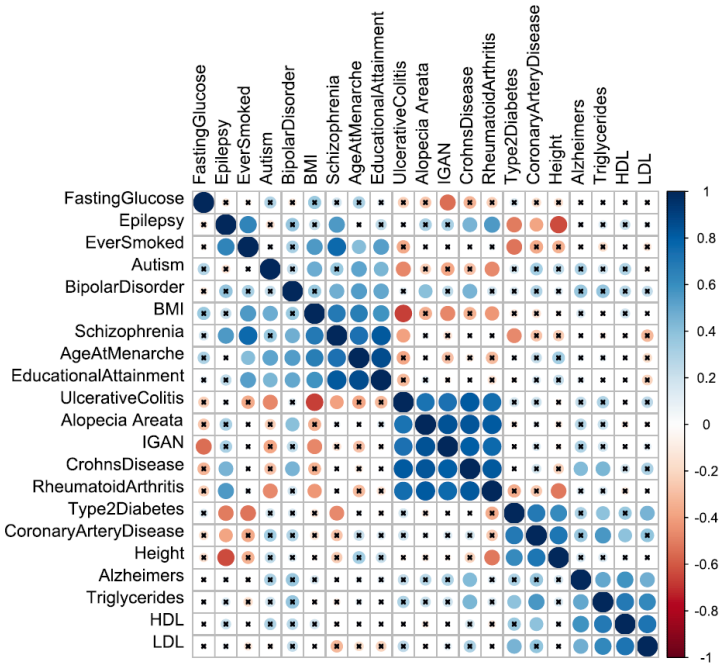
Такая модель легко реализуема в BigARTM.

M.Bicego, P.Lovato, et al. Investigating Topic Models' Capabilities in Expression Microarray Data Classification. 2012.

Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. 2016.

Функциональная аннотация генома человека

Цель: прогноз тканеспецифических функций некодирующих генетических вариантов для каждой позиции в геноме человека в 127 различных тканях и типах клеток.



Регуляризаторы (гипотеза):

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{supervised}\right) +$$

$$+ R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{hierarchy}\right) + R\left(\text{n-gram}\right) \to \max$$

Такая модель легко реализуема в BigARTM.

D.Backenroth et al. FUN-LDA: a latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. 2018

Резюме

- *Вероятностное тематическое моделирование* — это интерпретируемая кластеризация нескольких множеств, элементы которых взаимодействуют между собой.
- Принято считать, что это анализ текстовых коллекций, но, скорее, это синтез латентных векторных представлений вершин графа по наблюдаемым данным о рёбрах.
- **Теория ARTM** позволяет комбинировать регуляризаторы для построения моделей с требуемыми свойствами.
- **Библиотека BigARTM** — модульная реализация ARTM, «лего-конструктор» тематических моделей.
- **Обсуждение:** возможности применения теории ARTM и библиотеки BigARTM в задачах биоинформатики.
- В биоинформатике недалеко уходят от устаревшей LDA.

К.Воронцов. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. 2023. (для изд-ва URSS)
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf