Сценарии использования BigARTM: тематический анализ текстовых и мультимодальных данных

Воронцов Константин Вячеславович
voron@mlsa-iai.ru
зав. лаб. Машинного интеллекта и семантического анализа Института ИИ МГУ; проф., зав. каф. ММП ВМК МГУ; проф., зав. каф. МОЦГ МФТИ; г.н.с. ФИЦ ИУ РАН

Форум «Открытые данные»
Томск • 10–11 ноября 2023

Содержание

Что такое «тема» в коллекции текстовых документов?

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

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

Имея коллекцию текстовых документов, хотим узнать:

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

Пример. Мультиязычная модель Википедии

216 175 русско-английских пар статей.
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. Мультиязычная модель Википедии

216 175 русско-английских пар статей.
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.

Свойство интерпретируемости тематических моделей

Тематическая модель формирует тематические векторы:

- $p(t|d)$ для каждого документа d
- $p(t|w)$ 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 $\left\{w\colon p(w|t) > \gamma p(w)\right\}$
- и способна «рассказать о себе» словами и фразами языка
- 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:

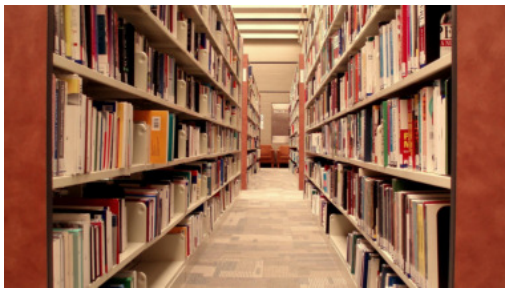
- Разведочный анализ текстовых данных
- Выяснить тематическую кластерную структуру текстовой коллекции, сколько в ней тем, и о чём они
- Получать интерпретируемые тематические векторные представления (*эмбединги*) документов, фрагментов, слов $p(t|d)$, $p(t|w)$, $p(t|d,w)$ и нетекстовых объектов $p(t|x)$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

Не-цели:

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

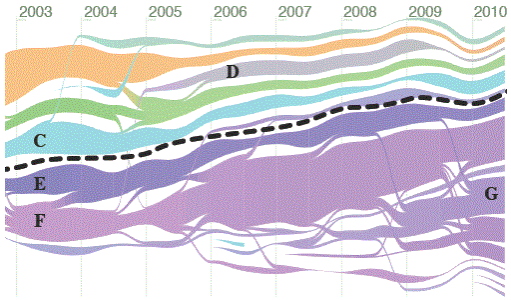
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

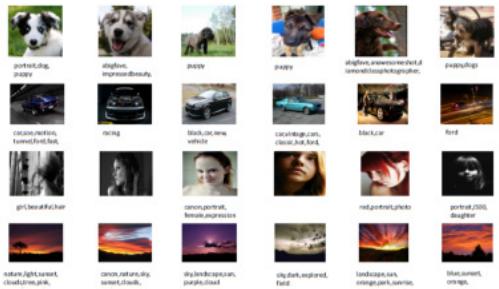


поиск тематического контента в соцсетях

выявление и отслеживание цепочек новостей

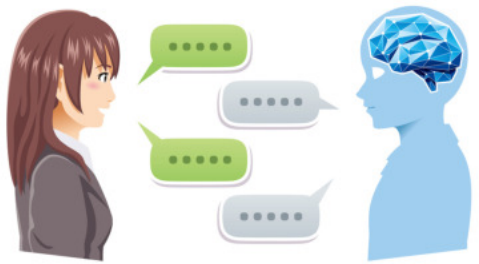


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

управлением диалогом в разговорном интеллекте



Постановка задачи тематического моделирования

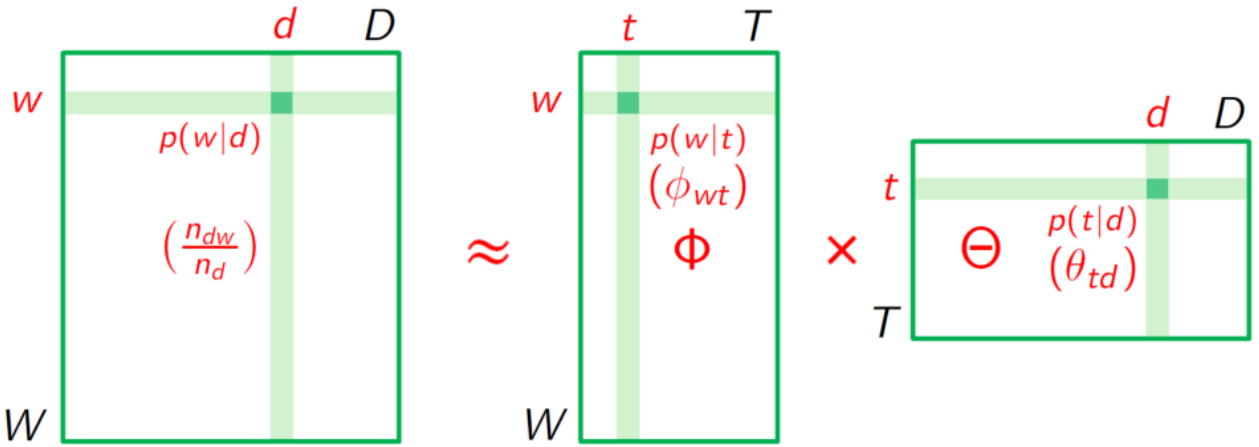
Дано: коллекция текстовых документов

- 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

Это задача стохастического матричного разложения:



ARTM: аддитивная регуляризация тематических моделей

Критерий: максимум лог-правдоподобия с регуляризатором

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

EM-алгоритм: метод простой итерации для системы уравнений

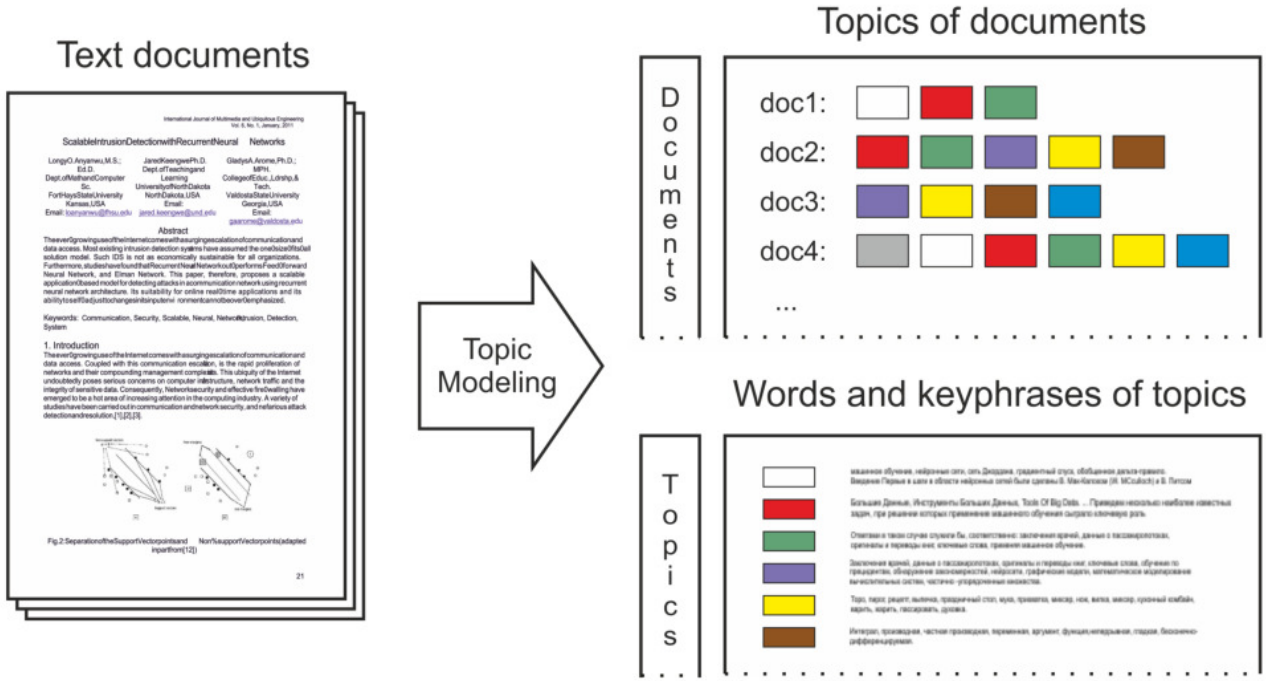
$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

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

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

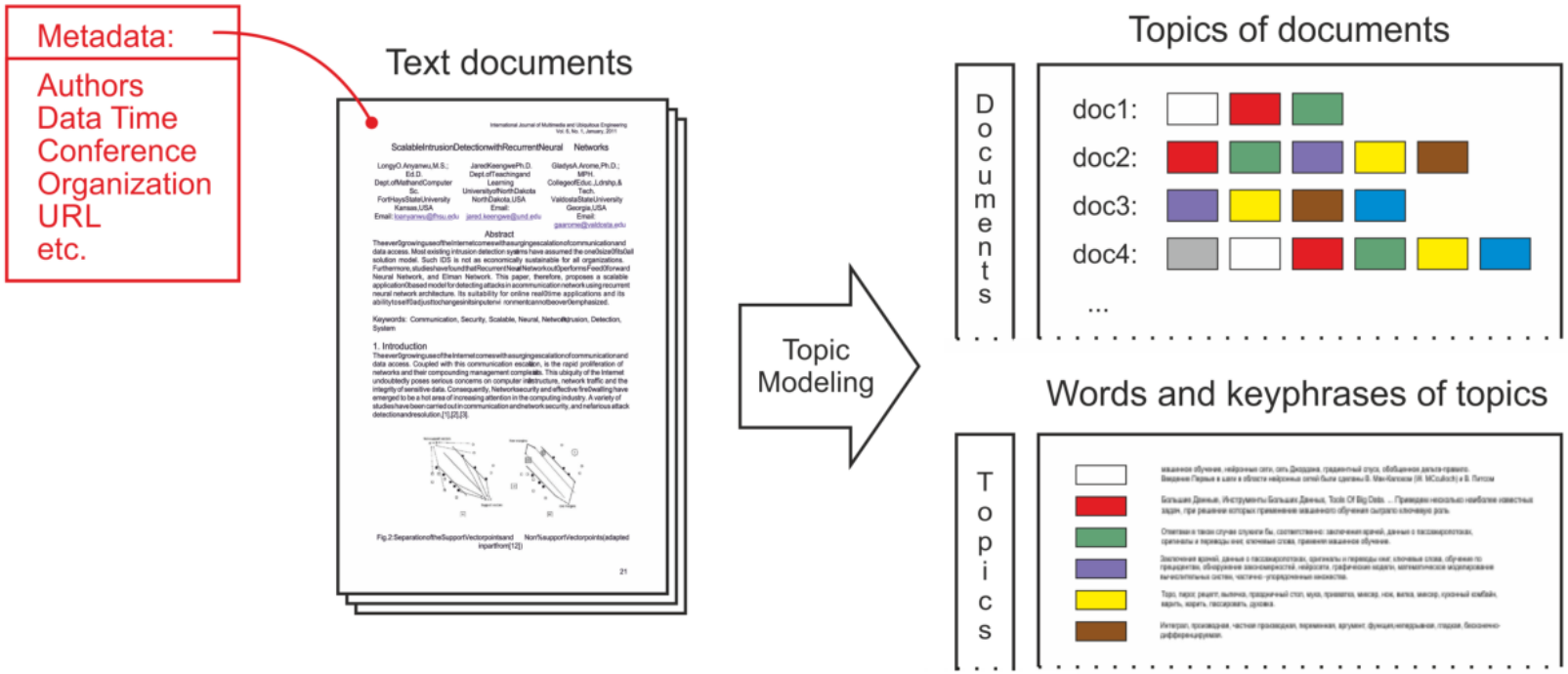
Мультимодальная тематическая модель

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



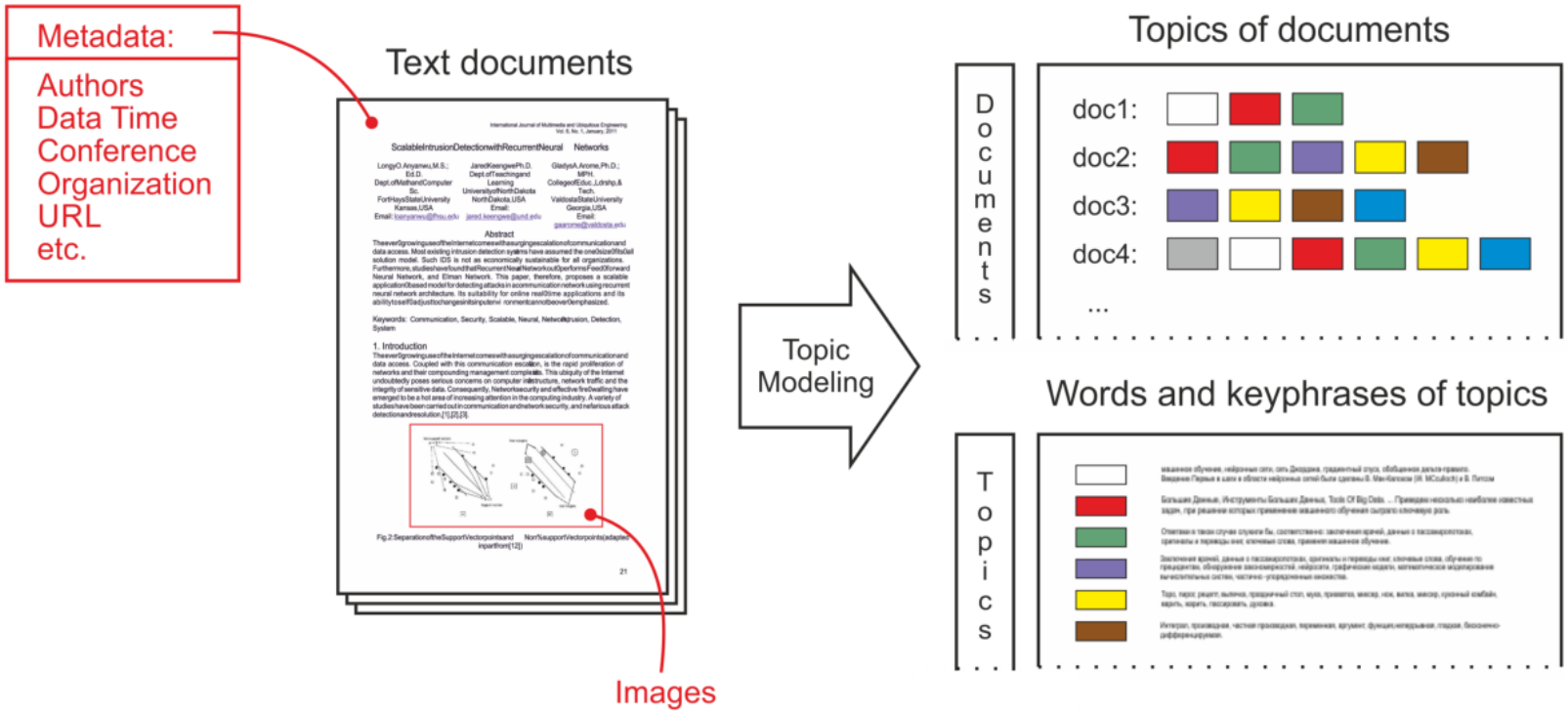
Мультимодальная тематическая модель

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



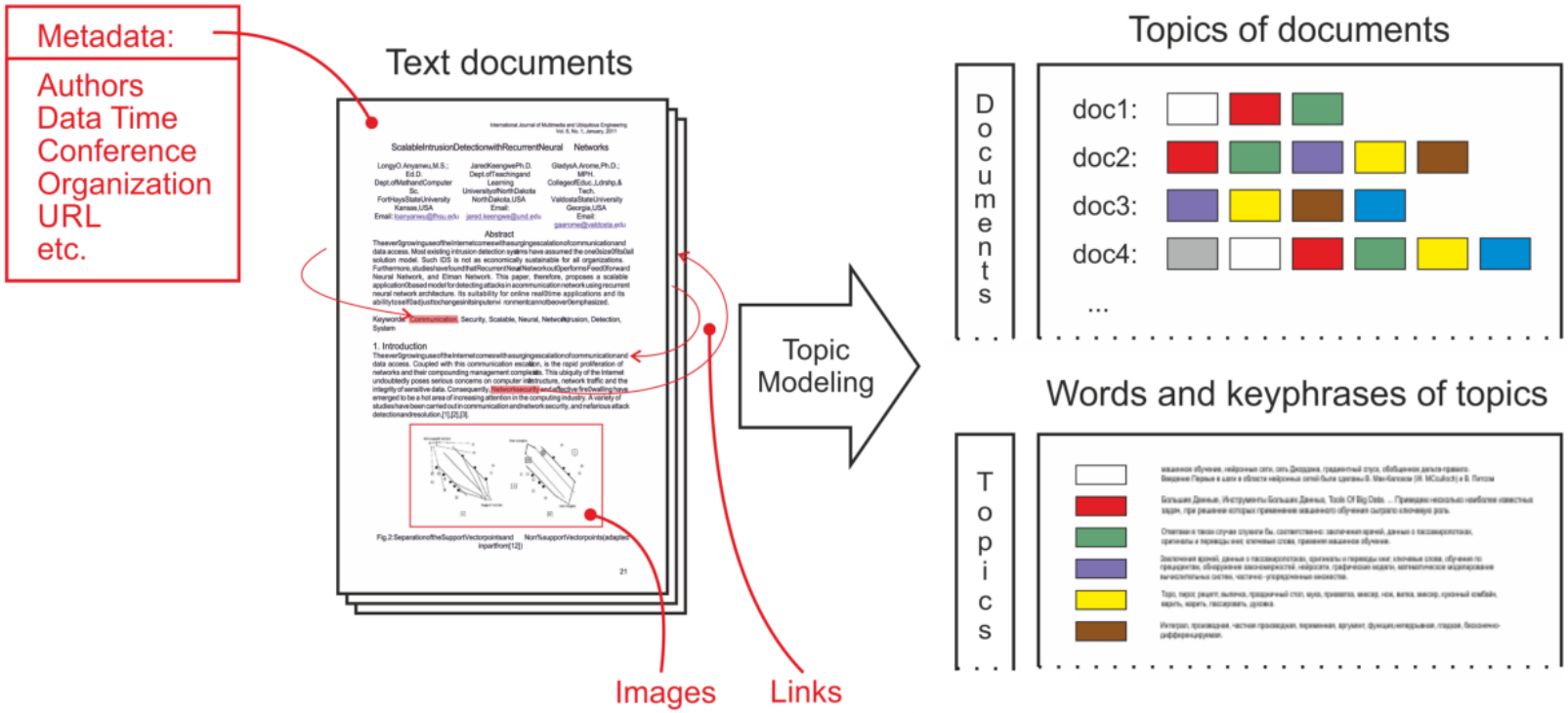
Мультимодальная тематическая модель

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



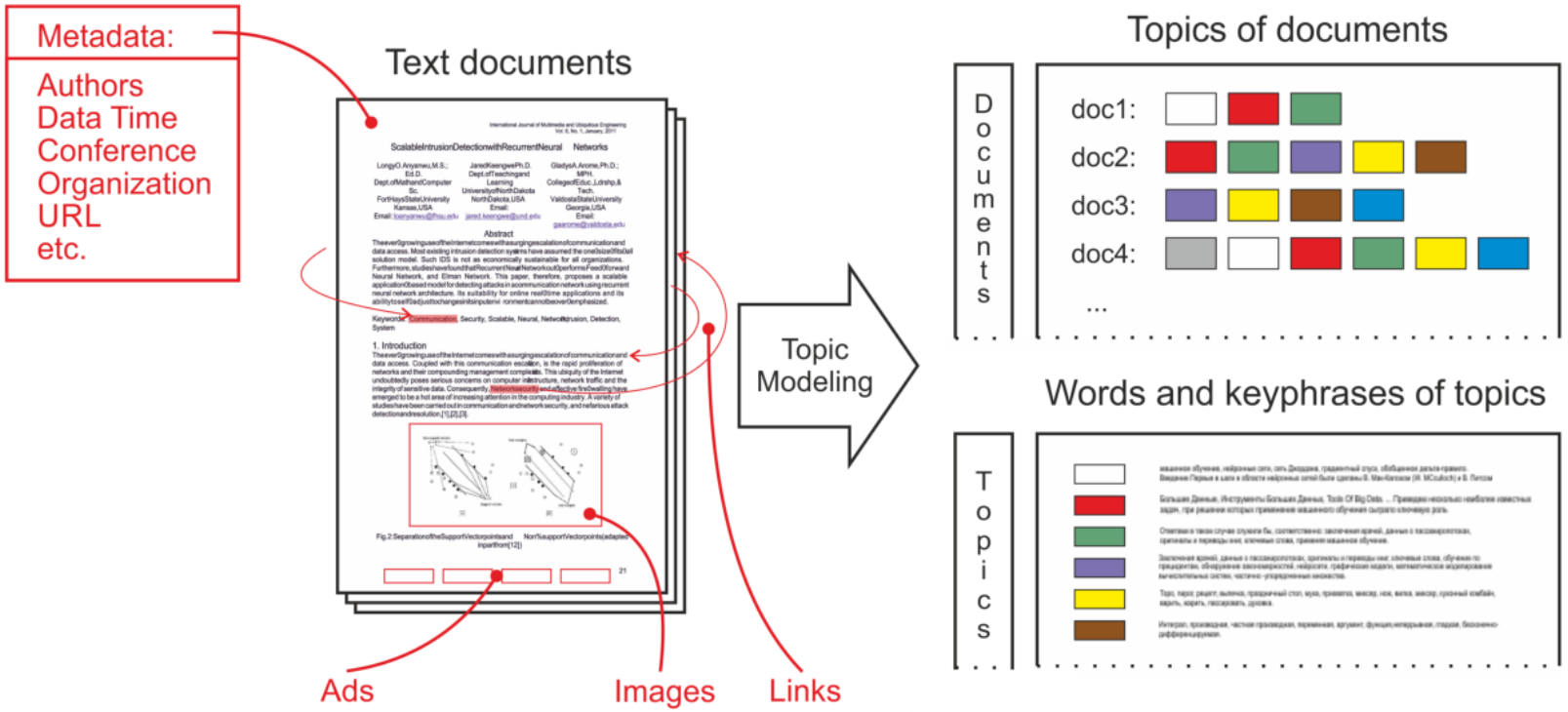
Мультимодальная тематическая модель

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



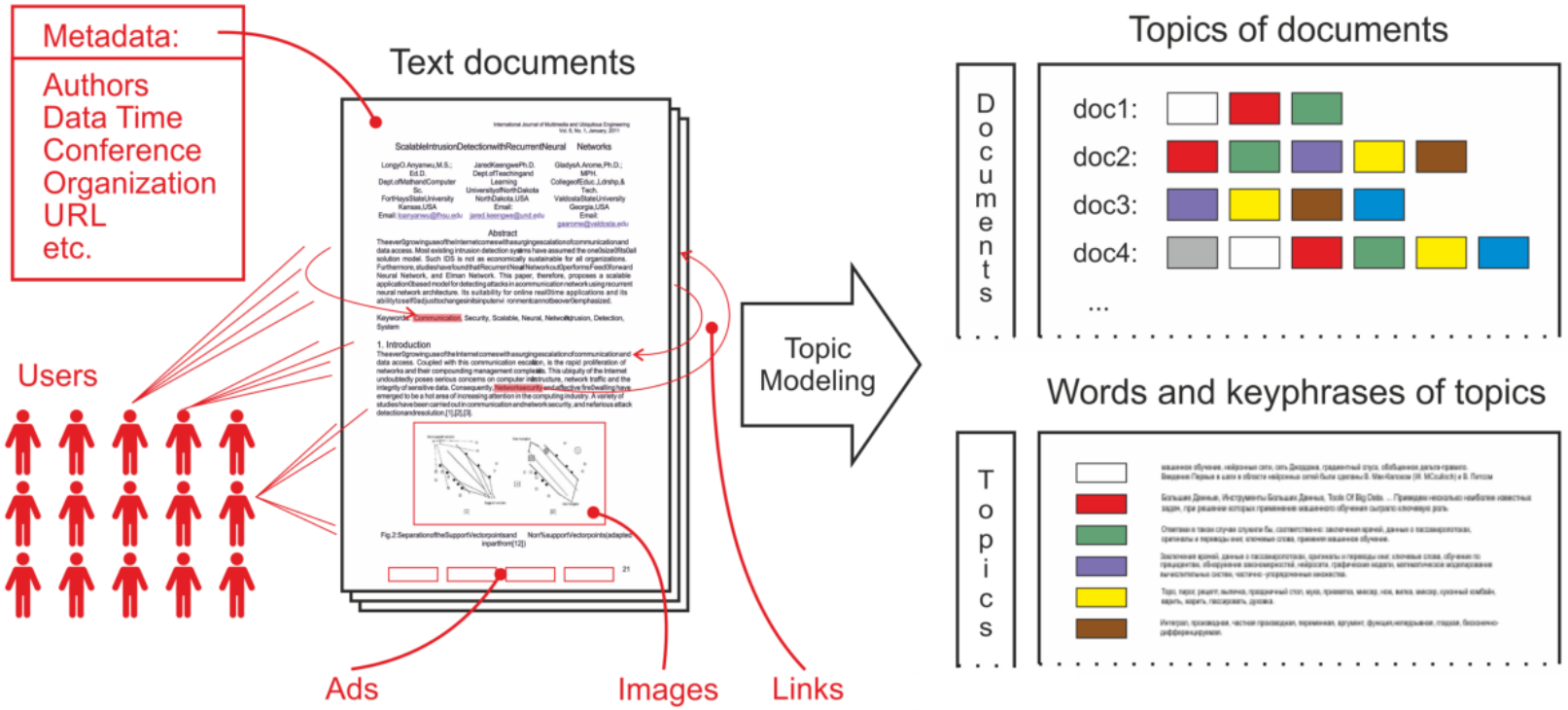
Мультимодальная тематическая модель

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



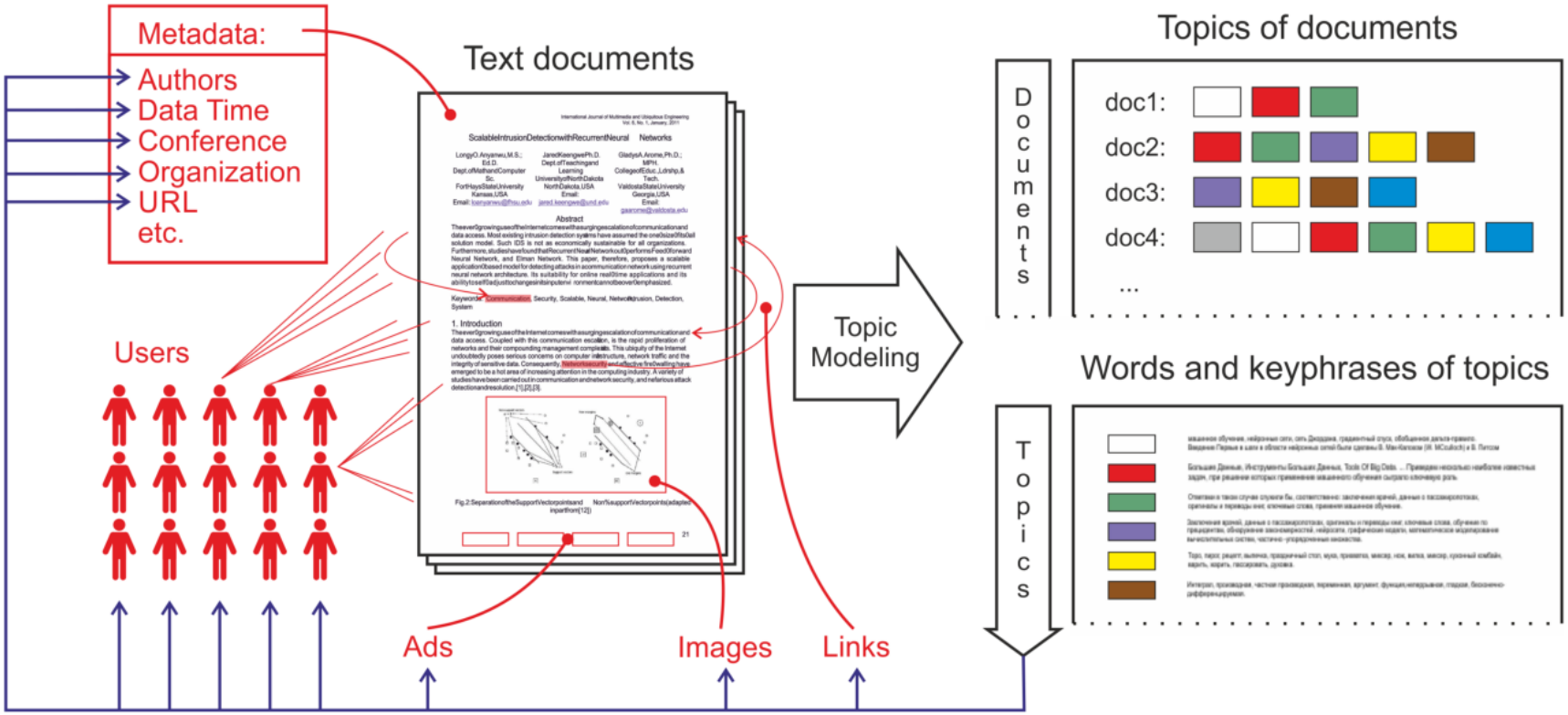
Мультимодальная тематическая модель

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



Мультимодальная тематическая модель

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



Мультимодальная ARTM

W_m — словарь термов m-й модальности, $m \in M$

Максимизация суммы log-правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

BigARTM: библиотека тематического моделирования

Ключевые возможности:

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

Сообщество:

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

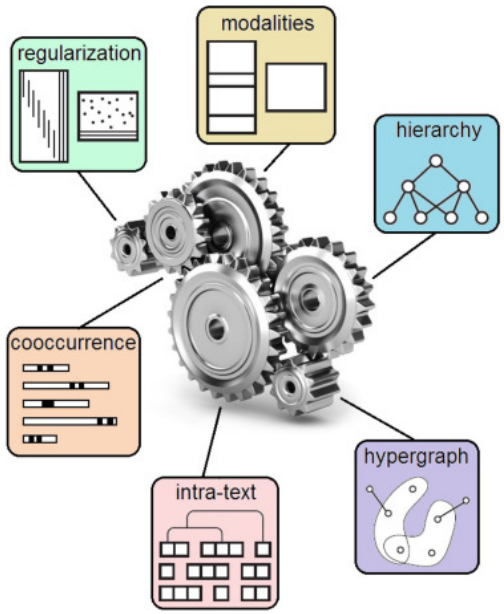
Лицензия и среда разработки:

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

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 результатов тематического моделирования

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

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.

ARTM — модульный подход к синтезу требуемых моделей

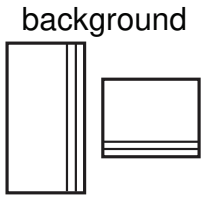
Для построения композитных моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».

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

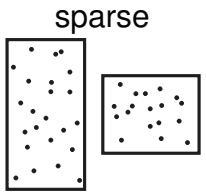
-- стандартизуемые этапы

Регуляризаторы для улучшения интерпретируемости тем



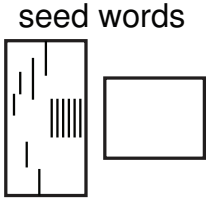
Сглаживание фоновых тем $B \subset T$:

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

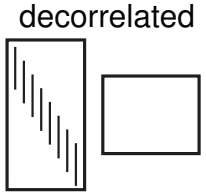


Разреживание предметных тем $S = T \setminus B$:

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

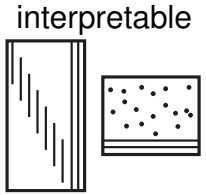


Сглаживание для выделения релевантных тем
с помощью словаря «затравочных» ключевых слов



Декоррелирование для повышения различности тем:

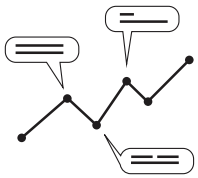
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

Регуляризаторы для учёта дополнительной информации

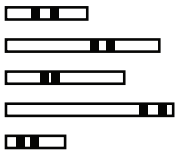
regression



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

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

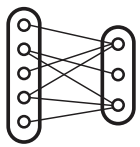
biterm



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

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

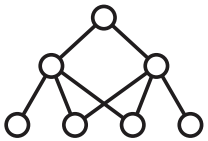
relational



Связи или ссылки между документами:

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

hierarchy

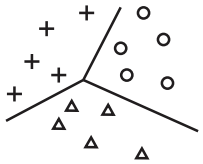


Связи родительских тем t с дочерними подтемами s:

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

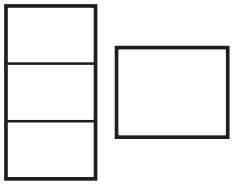
Регуляризаторы для мультимодальных тематических моделей

supervised



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

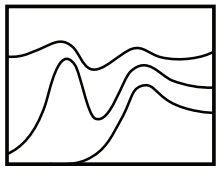
multilanguage



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:

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

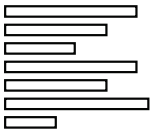
geospatial

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

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

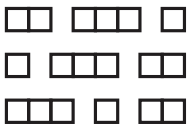
Регуляризаторы для моделирования последовательного текста

sentence



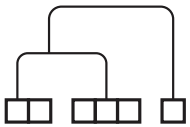
Тематические модели, учитывающие границы предложений, абзацев и секций документов

n-gram



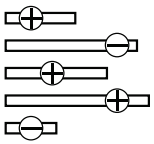
Модели с модальностями *n*-грамм, коллокаций, именованных сущностей (используем TopMine)

syntax



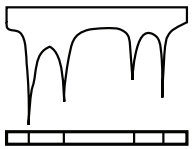
Модели, учитывающие результаты автоматического синтаксического разбора (используем UDPipe)

sentiment



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

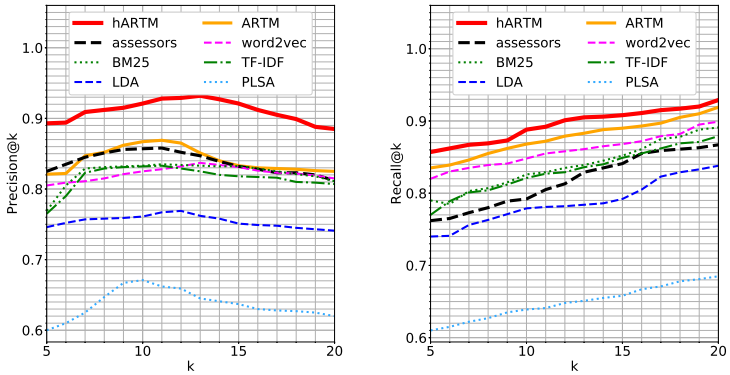
segmentation



Тематические модели сегментации с автоматическим определением границ сегментов

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 (затравка — словарь 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, боления, фидель, глаза, 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:

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

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.

Аналогичные исследования по выделению узкой тематики

Задачи «поиска и классификации иголок в стоге сена»

- поиск и кластеризация новостей [1]
- поиск в социальных медиа информации, связанной с болезнями, симптомами и методами лечения [2]
- поиск чатов, связанных с преступностью и экстремизмом [3, 4]
- поиск выступлений о правах человека в ООН [5]

---

1. *J.Jagarlamudi, H.Daumé III, R.Udupa.* Incorporating lexical priors into topic models. 2012.

2. *M.Paul, M.Dredze.* Discovering health topics in social media using topic models. 2014.

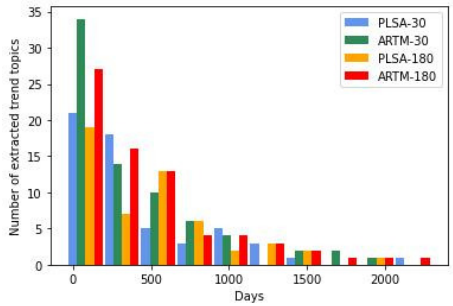
3. *M.A.Basher, A.Rahman, B.C.M.Fung.* Analyzing topics and authors in chat logs for crime investigation. 2014.

4. *A.Sharma, M.Pawar.* Survey paper on topic modeling techniques to gain useful forecasting information on violant extremist activities over cyber space. 2015.

5. *Kohei Watanabe, Yuan Zhou.* Theory-driven analysis of large corpora: semisupervised topic classification of the UN speeches. 2022.

Выявление трендов в коллекции научных публикаций

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



Регуляризаторы:

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

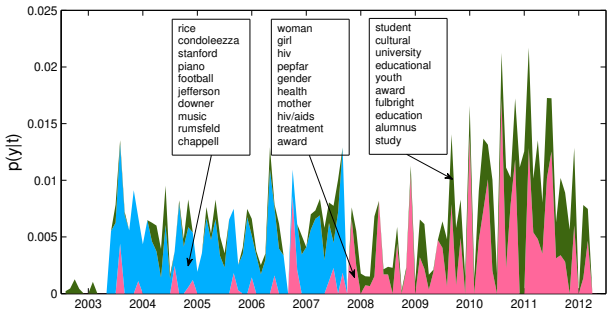
Результаты:

- 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.

Выявление динамики тем в новостных потоках

Цель: выделение тем в коллекции пресс-релизов МИДов 4х стран, с привязкой ко времени.



Регуляризаторы:

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

Результаты:

- разделение тем на событийные и перманентные
- когерентность тем: 5.5 $\to$ 6.5

Н.Дойков. Адаптивная регуляризация вероятностных тематических моделей. ВКР бакалавра, ВМК МГУ, 2015.

Выделение поляризованных мнений в политических новостях

Цель: найти признаки, по которым событийная тема разделяется на кластеры-мнения

Modalities	*Pr*	*Rec*	*F1*
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	**0.77**	**0.97**	**0.86**

Регуляризаторы:

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

Результаты:

- выделение мнений внутри тем: F1-мера = 0.86%
- совместное использование трёх модальностей:
  - факты как триплеты «субъект–предикат–объект»
  - семантические роли слов по Филлмору
  - тональности именованных сущностей

D.Feldman, T.Sadekova, K.Vorontsov. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. Dialogue 2020.

Анализ программ развития российских вузов

Цель — выявить закономерности в стратегиях развития вузов, не читая всех этих документов (Distant Reading)

- **Дано:**
  программам развития ВУЗов: 396 файлов, 284 вуза
- **Найти:**
  полный тематический спектр направлений развития
- **Критерий:**
  интерпретируемость тем;
  чёткого количественного критерия нет :(
  чётких целей исследования нет — какие именно закономерности ожидает найти заказчик? :(

Пример интерпретации темы

(слова): инновационный исследование результат региональный предприятие проведение основа среда внедрение уровень рамка сфера исследовательский научно научно-исследовательский участие приоритетный специалист цель выполнение международный прикладной ведущий взаимодействие

(биграммы): научный_исследование инновационный_деятельность приоритетный_направление научно_исследовательский исследование_разработка развитие_инновационный фундаментальный_прикладной разработка_внедрение направление_развитие мировой_уровень научно_образовательный исследовательский_деятельность инновационный_развитие малое_инновационный инновационный_предприятие научный_инновационный модернизация_научно-исследовательский прикладной_исследование инновационный_проект развитие_научный инновационный_инфраструктура проведение_научный

(ИНТЕРПРЕТАЦИЯ): научные исследования и инновационное развитие

Пример интерпретации темы

(слова): международный число количество участие конференция зарубежный увеличение учёный академический мобильность конкурс сотрудничество грант иностранный аспирант совместный молодая ведущий специалист привлечение преподаватель исследование школа сотрудник семинар

(биграммы): увеличение_количество академический_мобильность увеличение_число международный_деятельность международный_сотрудничество международный_научный развитие_международный принять_участие российский_международный научный_мероприятие международный_образовательный участие_международный иностранный_студент количество_студент научный_проект университет_международный международный_уровень международный_академический количество_участник научный_конференция программа_академический участие_студент

(ИНТЕРПРЕТАЦИЯ): академическая мобильность и международное сотрудничество

Пример интерпретации темы

(слова): общежитие корпус здание ремонт площадь инфраструктура комплекс помещение строительство объект капитальный кампус имущественный спортивный реконструкция безопасность территория сооружение место оборудование современный замена учебно-лабораторный комфортный

(биграммы): учебный_корпус капитальный_ремонт имущественный_комплекс общий_площадь здание_сооружение студенческий_общежитие корпус_общежитие развитие_имущественный инфраструктура_университет создание_комфортный развитие_инфраструктура университетский_кампус комплекс_университет спортивный_комплекс студент_сотрудник объект_университет земельный_участок условие_проживание территория_университет объект_инфраструктура социальный_инфраструктура использование_имущественный строительство_новый ремонтный_работа общежитие_университет

(ИНТЕРПРЕТАЦИЯ): инфрастуктура, кампус, строительство

Интерпретация всех 50 тем

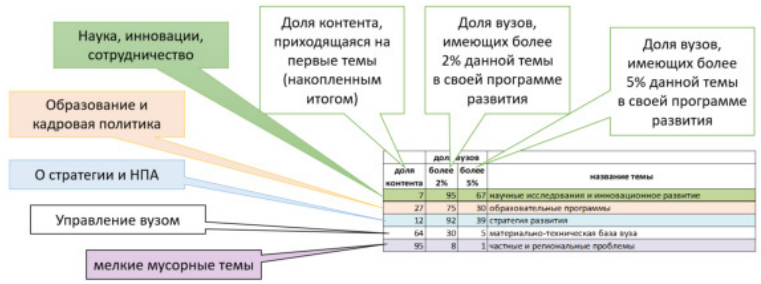
- Для интерпретируемости тем важны биграммы
- Модель построили примерно с 10-й попытки (подбирали число тем, регуляризацию, добивались различности тем)
- Интерпретация 50 тем заняла примерно 20 минут работы
- Иногда выделялись темы исследований и разработок, но для этого нужна более гранулированная модель
- Темы были сгруппированы вручную по 5 категориям:
  1. 16 тем про науку, инновации и сотрудничество
  2. 14 тем про образование и кадровый потенциал
  3. 11 тем про административное управление и хозяйство вуза
  4. 3 темы «юридические», о самой стратегии развития
  5. 6 тем «малые и мусорные», вместе не более 5% контента

Интерпретация всех 50 тем

	доля вузов		
доля контента	более 2%	более 5%	название темы
7	95	67	научные исследования и инновационное развитие
12	92	39	стратегия развития
15	84	23	академическая мобильность и международное сотрудничество
19	82	17	кадровой потенциал и кадровая политика
22	80	14	иностранные студенты
27	75	30	образовательные программы
30	75	13	повышение квалификации и переподготовка кадров
33	70	10	система управления вузом
36	68	16	учебный процесс
39	62	15	финансы и бюджет
43	62	21	бюрократия
45	56	3	подготовка высококвалифицированных кадров
48	47	9	инфрастуктура, кампус, строительство
50	44	4	меры повышения качества образования
52	42	4	влияние на экономику региона
54	41	8	молодежная политика
56	41	6	центры компетенций и технологического превосходства
58	40	6	отсылки к стратегическим документам и НПА
60	36	1	работа со школьниками и талантливой молодежью
62	34	7	ректорат и органы управления вузом
64	30	5	материально-техническая база вуза
65	29	2	связь с общественностью, имидж вуза
67	29	8	исследования сх, лес, химия, ит
69	29	1	публикационная активность и защиты диссертаций
71	29	2	взаимодействие с региональной властью

	доля вузов		
доля контента	более 2%	более 5%	название темы
72	27	1	образовательные программы, аккредитация, профстандарты
74	25	3	спортивная и культурная жизнь вуза
75	21	5	стратегия развития и региональная среда
77	20	1	образовательный процесс и образовательные технологии
78	19	1	международное сотрудничество и договорные отношения
79	19	2	цифровизация и цифровые технологии
81	18	2	медицинское обеспечение, обучение инвалидов
82	18	5	блоки мероприятий и показатели результативности
84	18	5	работа структурных подразделений вуза
85	17	2	выход в мировые рейтинги университетов
86	14	1	технологии транспорта и искусственного интеллекта
87	13	1	публикационная и издательская деятельность
88	12	1	финансовое и ресурсное обеспечение программы развития
89	11	1	мониторинг показателей эффективности
90	11	0	сетевые образовательные программы, ворлдскиллс
92	11	1	региональные особенности приёма и рынка труда
93	10	1	приём абитуриентов
93	10	0	исследования в экологии и медицине
94	9	1	образовательные программы (частные вопросы)
95	8	1	частные и региональные проблемы
96	8	2	авиационные технологии
97	8	0	смесь тем
98	7	0	образовательные программы & урбанистика и туризм (смесь тем
99	7	1	смесь тем
100	7	1	частные юридические вопросы

16 тем — наука и инновации
14 тем — образование и кадры
11 тем — управление и хозяйство
3 темы — о стратегии развития
6 тем — мелкие мусорные



Тематическая карта вузов

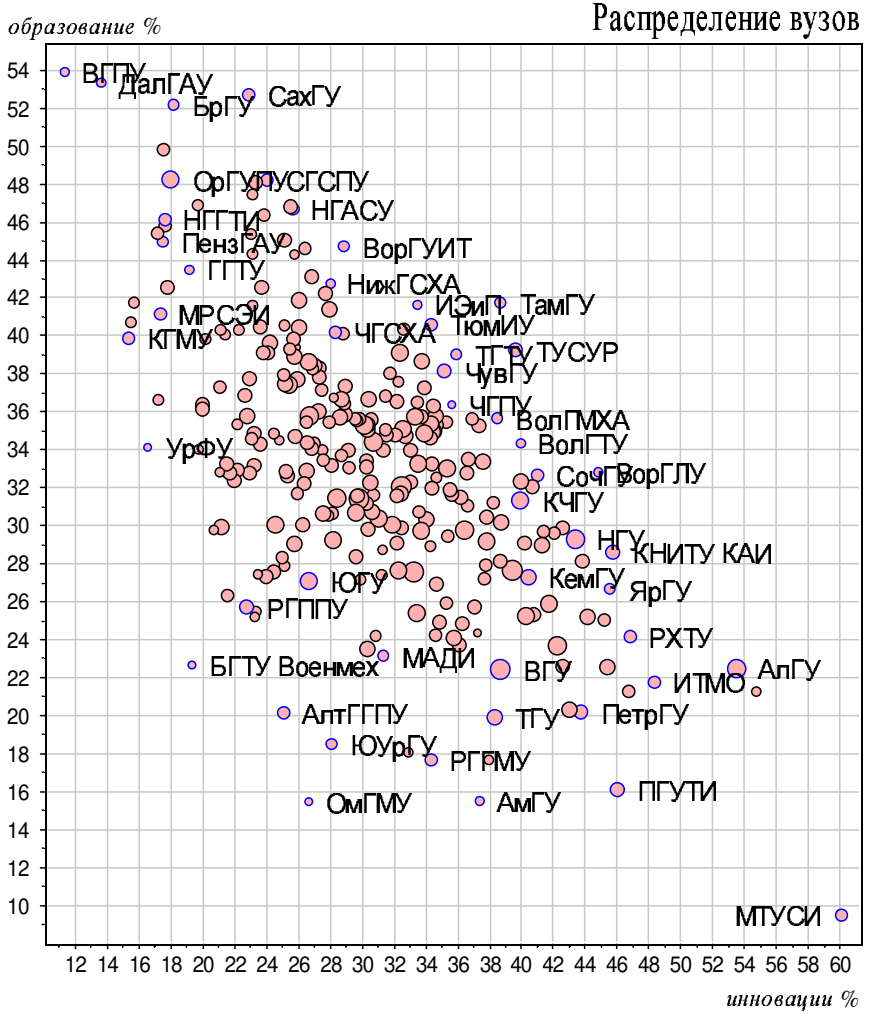
По осям:

объёмная доля тем
— про инновации
— про образование

Вывод:

объёмные доли тем, возможно, показывают баланс приоритетов развития

...хотя... это похоже на оценивание научного отчёта толщиной в сантиметрах :)



Проекты Школы Прикладного Анализа Данных (ноябрь, 2022)

Исходные данные ВКонтакте (через сервисы Крибрум)

- Анализ социального влияния на формирование образа правильного питания у студентов г. Томска
- Анализ научной и публикационной активности сотрудников университета или научной организации
- Анализ практик участия в читательских сообществах, формирующихся вокруг авторов или жанров
- Анализ социального и политического взаимодействия сетевых сообществ в регионах ресурсного типа
- Анализ туристической активности и оценка портрета потенциального туриста — путешественника по Камчатке
- Анализ корпуса текстов образовательных дисциплин: программа курса + материалы курса + отчёты студентов
- Анализ научной педагогической литературы для построения карт компетенций

TM в исторических исследованиях: газетные архивы

[1] Корпус *Pennsylvania Gazette* 1728–1800, 25M слов:
— выделение последовательности событийных тем;
— изучение синхронности событий;
— комбинирование автоматического анализа и ручного.

[2] *Газеты Техаса* от гражданской войны до наших дней:
— выделение всех тем, связанных с хлопком;
— построение серии моделей в скользящих окнах;
— важность качественной предобработки текстов.

[3] Газеты и периодика Финляндии (1854–1917):
— выделение тем о церкви, религии, образовании;
— тренды модернизации и секуляризации финского общества.

---

1. *D.Newman, S.Block.* Probabilistic topic decomposition of an eighteenth-century American newspaper. 2006.

2. *Tze-I Yang, A.J.Torget, R.Mihalcea.* Topic modeling on historical newspapers. 2011.

3. *J.Marjanen et al.* Topic modelling discourse dynamics in historical newspapers. 2021.

TM в исторических исследованиях: летописи и дневники

[1] Двуязычный корпус книг на английском и немецком:
— все темы, связанные с эпистемологией

[2] Корпус текстов на китайском языке (1644–1912):
— все темы, связанные с бандитизмом, преступлениями;
— необходим контекст для установления типа преступления;
— важность правильной токенизации для китайского языка.

[3] Дневник Martha Ballard (1735–1812), охватывает 27 лет:
— выделение событийных и перманентных тем;
— выделение персональных и исторических тем;
— специфичный английский XVIII века.

---

1. *M.Erlin.* Topic modeling, epistemology, and the English and German novel. 2017.
2. *Ian Matthew Miller.* Rebellion, crime and violence in Qing China, 2013.
3. *Cameron Blevins.*
http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary.

TM в исторических исследованиях: журнальная периодика

Статьи коллекции JSTOR доступны в виде «мешков слов».

[1] Научные журналы XX века:
— различия тематики на английском и немецком языках;
— особенно исследовались различия, связанные со 2MB;
— для объединения тем использовались интервики Википедии.

[2] Более 100 лет литературно-художественной периодики:
— как менялись темы;
— как менялись значения слов внутри каждой темы;
— как менялась тема насилия (violence, power, fear, blood, death, murder, act, guilt).

---

1. *D.Mimno.* Computational historiography: Data mining in a century of classics journals. 2012.

2. *A.Goldstone, T.Underwood.* The quiet transformations of literary studies: What thirteen thousand scholars could tell us. 2014.

TM в политологии: анализ публичных выступлений

[1] Выступления (210K) в Европарламенте, 1999–2014:
— выявление событийных тем и эволюции перманентных тем;
— как члены и комитеты ЕП влияют на формирование тем

[2] Модель контрастных мнений (Contrastive Opinion Modeling)
— выступления в Сенате США (www.votesmart.org);
— СМИ: New York Times, Xinhua News, The Hindu, 2009–2010

[3] Выступления в Совбезе США по Афганистану, 2001–2017:
— динамика отношения разных стран к проблемам Афганистана

---

[1] *D.Greene, J.P.Cross.* Unveiling the political agenda of the European Parliament plenary: a topical analysis. 2015.

[2] *Fang, Y., et al.* Mining contrastive opinions on political texts using cross-perspective topic model. 2012.

[3] *M.Schönfeld.* Discursive landscapes and unsupervised topic modeling in IR: a validation of text-as-data approaches through a new corpus of UN Security Council speeches on Afghanistan. 2018.

TM в политологии: анализ СМИ и социальных медиа

[1] Тематика изменения климата в СМИ Пакистана, 2010–2021
— выявление, группирование и динамика тем

[2] Выявление поляризации новостей (AYLIEN COVID-19)
— 1,5M новостей, 440 источников СМИ, 11.2019–07.2020

[3] Выявление политических взглядов пользователей Twitter

[4] Что пишет NYT о ядерных технологиях с 1945 по н/в

---

[1] *W.Ejaz et al.* Politics triumphs: A topic modeling approach for analyzing news media coverage of climate change in Pakistan. 2023

[2] *Zihao He.* Detecting polarized topics using partisanship-aware contextualized topic embeddings. 2021

[3] *R.Cohen, D.Ruths.* Classifying Political Orientation on Twitter: It's Not Easy! 2013.

[4] *C.Jacobi.* Quantitative analysis of large amounts of journalistic texts using topic modelling. 2015.

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

Проект «Тематизатор»: общие требования

Переход от библиотек (BigARTM, VisARTM, TopicNet)
к приложению «Тематизатор» для конечного пользователя —
аналитика в области цифровых гуманитарных исследований

1. Цели пользователя — разведочный анализ, понимание тематической структуры данных, «о чём эта коллекция»
2. Пользователь не обязан знать
   — форматы исходных данных и способы их предобработки
   — теорию TM и ARTM, виды регуляризаторов
   — методики подбора гиперпараметров
   — критерии качества моделей
   — библиотеку BigARTM
3. Интуитивная визуальная среда, веб-интерфейс
4. Пользователю должны быть доступны настройки
5. Дефолтные настройки должны работать на любых данных

Приложения и исследования, взятые для анализа требований

1. Поиск этно-релевантных тем в социальных медиа
2. Анализ программ развития российских вузов
3. Проекты Школы Прикладного Анализа Данных
4. Тематический поиск по длинному текстовому запросу
5. Составление тематических подборок
6. Поиск и рубрикация научных статей на 100 языках
7. Выявление трендов в коллекции научных публикаций
8. Тематизация научно-просветительского онлайн-журнала
9. Поиск похожих дел в актах арбитражных судов
10. Тематизация пресс-релизов внешнеполитических ведомств
11. Тематизация twitter о российско-украинских отношениях
12. Выявление событийных тем в новостных потоках

Функциональные требования (по приоритетности)

1. Визуализация множества всех тем и их характеристик
2. Визуализация каждой темы с её «рассказом о себе»
3. Возможность задавать словари затравок для (групп) тем
4. Определение динамики тем во времени
5. Выявление коротких тем-событий и долгих тем-трендов
6. Разбиение тем на подтемы иерархически
7. Возможность группирования тем вручную
8. Выявление связей тем по сочетаемости в документах
9. Возможность отбора и накопления «банка тем»
10. Тематическая фильтрация коллекции
11. Тематический поиск по документу или фрагменту
12. Рекомендательный поиск и построение подборок

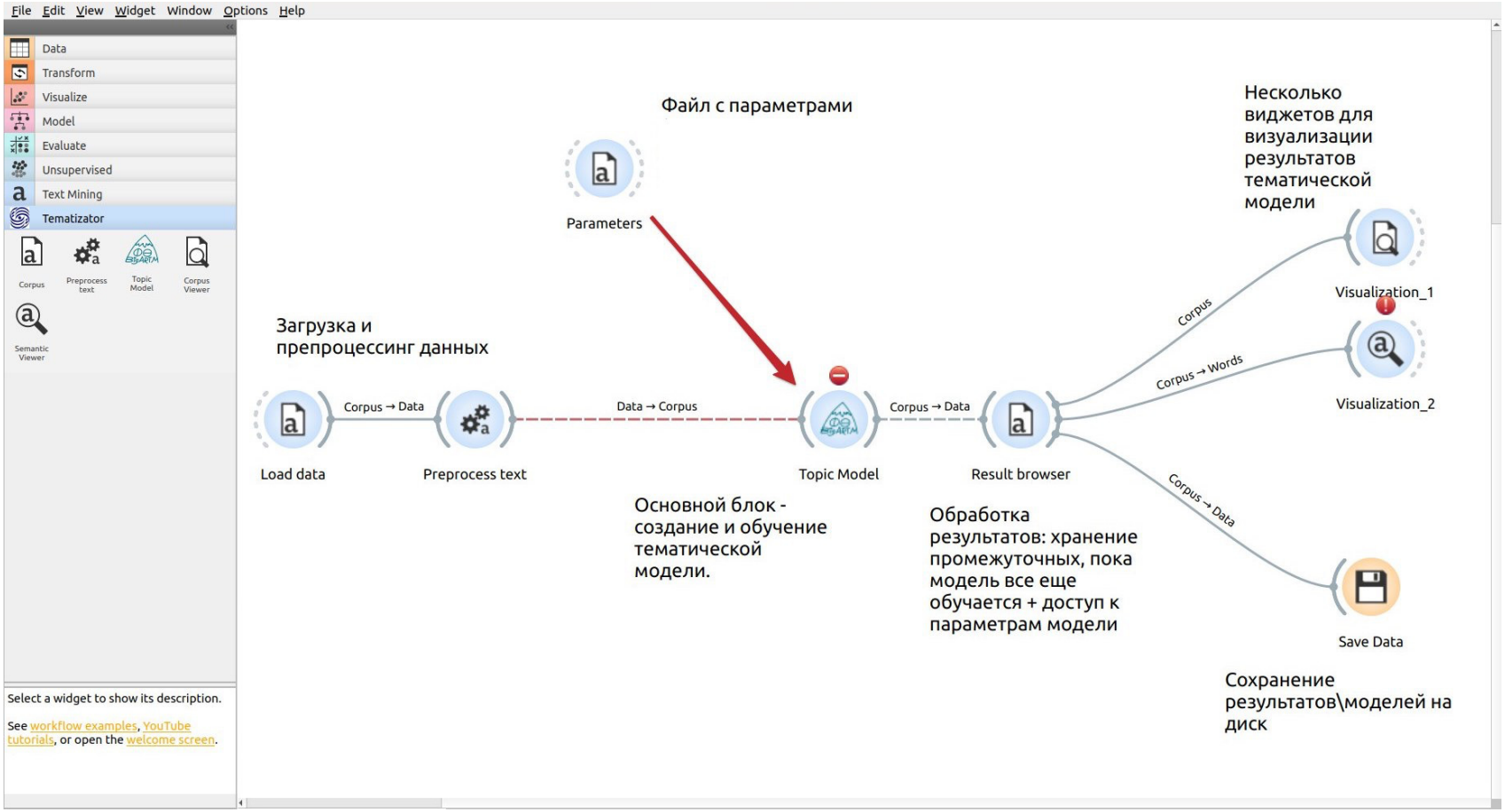
Требования к интерпретируемости (по приоритетности)

1. Доля интерпретируемых тем близка к 100%
2. Темы строятся более на терминах, чем на словах
3. Общая лексика выводится в отдельные фоновые темы
4. Нет мусорных тем, нет тем-дубликатов (декорреляция)
5. Решена проблема несбалансированности тем
6. Темы способны рассказать о себе словами и фразами
7. Нетекстовые термы способны рассказать о себе словами
8. Темы именуются автоматически
9. В иерархии имена дочерних тем уточняют родительские
10. Тематика слов согласуется с их локальными контекстами
11. Короткие тексты объяснимо наследуют тематику их слов
12. Длинные тексты разбиваются на тематические сегменты

Основной пользовательский сценарий (без детализации)

1. **Загрузка**
   - данные в различных «сырых» форматах
   - возможна дозагрузка данных порциями
2. **Предобработка**
   - автоматический выбор обработчиков на основании данных
   - выделение модальностей: языков, времени, терминов и т.д.
3. **Моделирование**
   - визуализация метрик качества в процессе обучения модели
   - возможность перехода к анализу, не прерывая обучения
4. **Визуализация**
   - каждая тема должна уметь «рассказать о себе»
   - много разных графиков (distant reading)
5. **Коррекция**
   - перебор моделей и накопление «банка тем»
   - пользовательские темы как подборки с рекомендациями

Встраивание модуля BigARTM в Orange



Требования к функциям Загрузки

1. Загрузка коллекций из различных сырых форматов
2. — txt, json, docx, odt, pdf и др.
3. — СМИ, соцмедиа, Википедия, статьи, патенты и др.
4. Представление метаданных и модальностей
5. Возможность загрузки как локально, так и из облака
6. Возможность дозагрузки данных из источника порциями
7. Текст как последовательность или как «мешок слов»
8. В одном файле один документ или много документов

Требования к функциям Предобработки

1. Автоматическая токенизация и лемматизация
2. Автоматическое исправление опечаток (соцсети)
3. Автоматическое выделение терминов n-грамм
4. Метаданные: авторы, время, категории, заголовки и др.
5. Модальности: онимы, теги, ссылки, пользователи и др.
6. Настройка шаблонов для выделения модальностей
7. Сортировка по времени и нарезка по пакетам
8. Автоматическое определение коротких текстов
9. Автоматическая редукция словарей (по необходимости)
10. Автоматическое определение языков
11. Машинный перевод для получения параллельных текстов
12. Предобработка не должна идти дольше тематизации

Требования к функциям Моделирования

1. Визуализация процесса обучения модели
2. Вывод метрик на графиках от #итерации, #пакета
3. Метрики перплексии, разреженности, вырожденности и др.
4. Автоматическая подстройка под короткие тексты
5. Автоматическая подстройка под длинные тексты
6. Темпоральная модель, если есть модальность времени
7. Подбор числа тем или построение иерархии тем
8. Автоматический подбор гиперпараметров, AutoML
9. Логирование информации о найденных аномалиях
10. Логирование данных о моделях, журнал экспериментов
11. Возможность перехода к анализу, не прерывая обучения
12. Возможность замены BigARTM на альтернативы

Требования к функциям Визуализации

1. Визуальная навигация по темам, документам, терминам
2. XY-график тем в осях свойств тем
3. XY-график документов/объектов в осях объёмов тем/групп
4. Построение спектра тем по семантической близости
5. XY-график документов в осях «время–спектр тем»
6. Визуализация связей между словами и понятиями темы
7. Визуализация динамики тем в осях «время–объём темы»
8. Визуализация иерархии тем
9. Визуализация связей тем по их сочетаемости в документах
10. Визуализация тематической структуры документа
11. Выбор характеристик тем для осей XY-графиков
12. Выбор характеристик объектов и документов для осей

Требования к функциям Коррекции

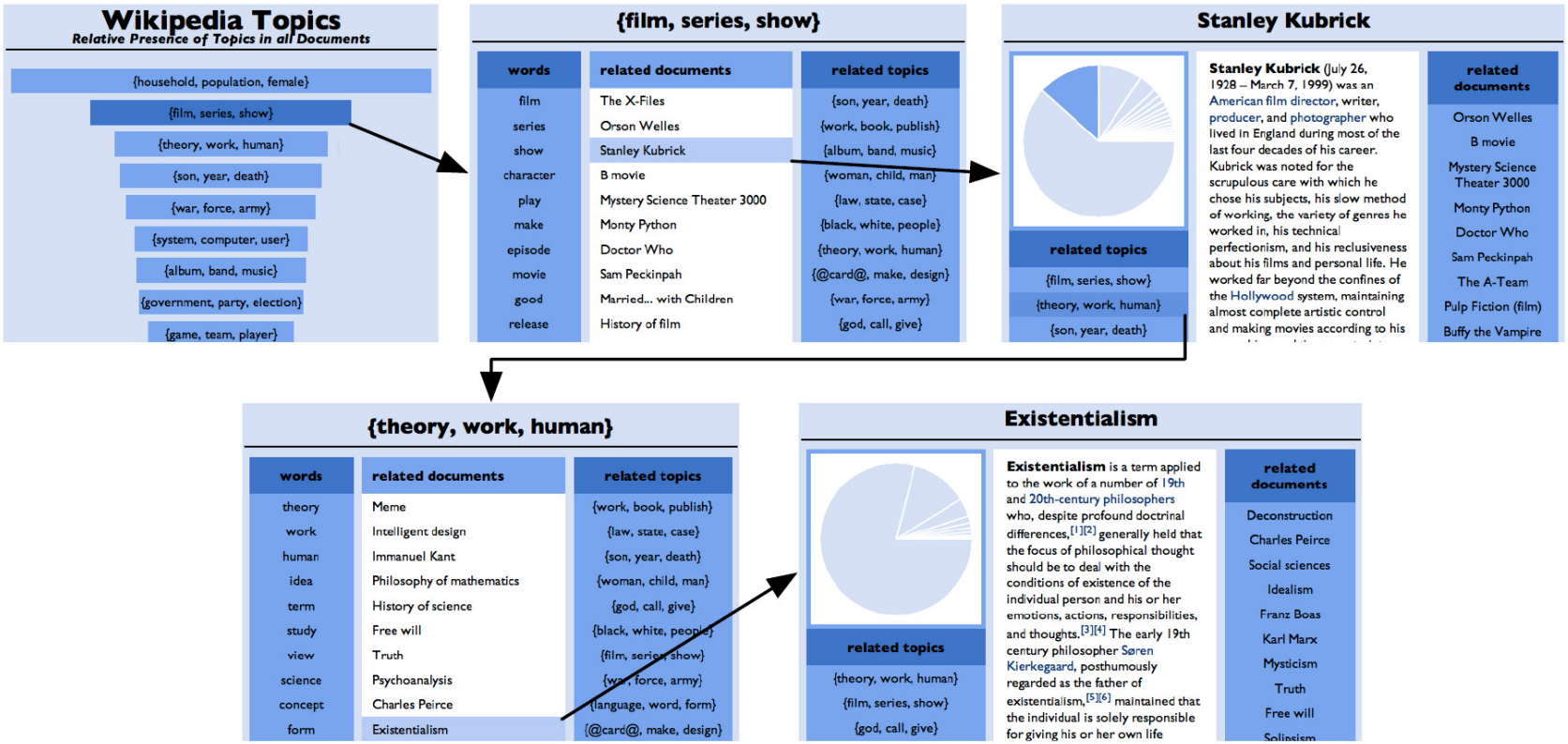
1. Разметка тем на релевантные, нерелевантные, мусорные
2. Разметка релевантных термов, документов в темах
3. Термы-затравки для «классификации иголок в стоге сена»
4. Обнаружение и расщепление неоднородных тем
5. Автоматический переход к тематической иерархии
6. Детекция новых событийных тем в темпоральных моделях
7. Накопление «банка тем» по множеству моделей
8. Многокритериальное оценивание качества моделей
9. Планирование экспериментов по улучшению моделей
10. Тематическая фильтрация коллекции и потока
11. Создание пользовательских тем — подборок документов
12. Ранжирование рекомендаций для пользовательских тем

Требования к рабочему пространству проекта пользователя

1. Настройки входных данных — коллекций и потоков
2. Настройки модулей предобработки
3. Структура и гиперпараметры сравниваемых моделей
4. Структура и гиперпараметры финальной модели
5. Визуализации процесса обучения модели
6. Визуализации количественных результатов моделирования
7. Визуализации качественных результатов (аннотации тем)
8. Банк тем — множество тем, отобранных из моделей
9. Пользовательские темы — подборки документов
10. Настройка подробности отчёта по проекту
11. Настройка комментариев к пунктам отчёта по проекту
12. Сгенерированный отчёт по проекту

Визуализация TMVE (Topic Model Visualization Engine)

Тематический навигатор с веб-интерфейсом:

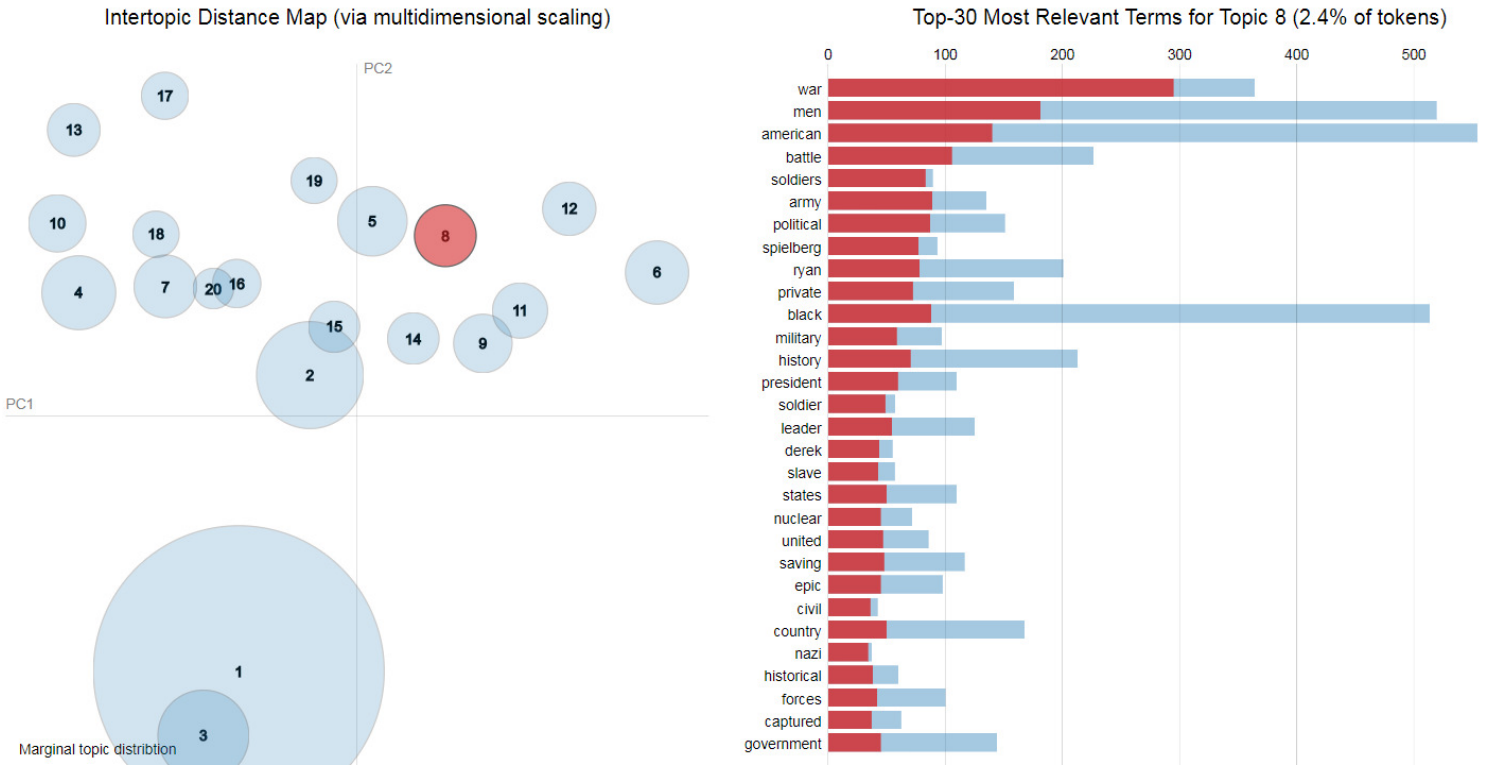


https://github.com/ajbc/tmv

Chaney A., Blei D. Visualizing Topic Models, 2012.

Система LDAvis

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

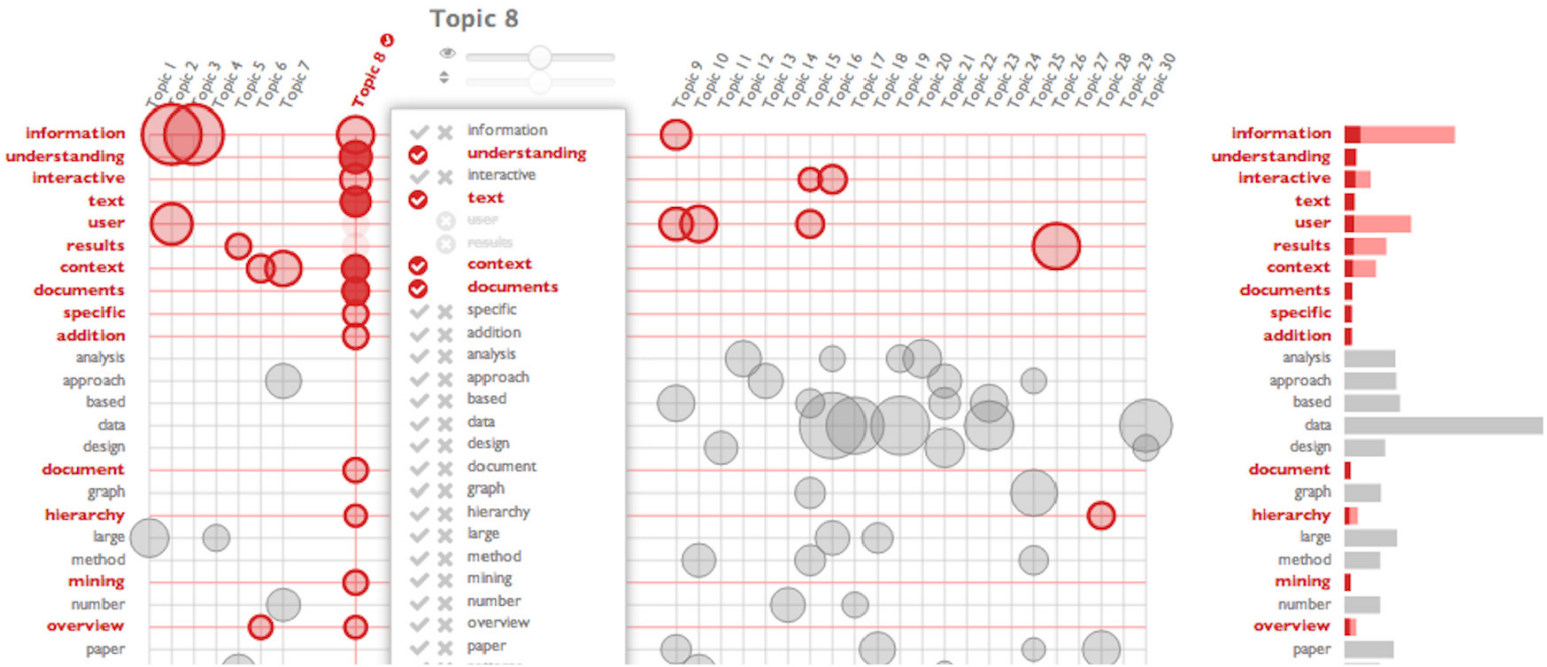


https://github.com/cpsievert/LDAvis

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.

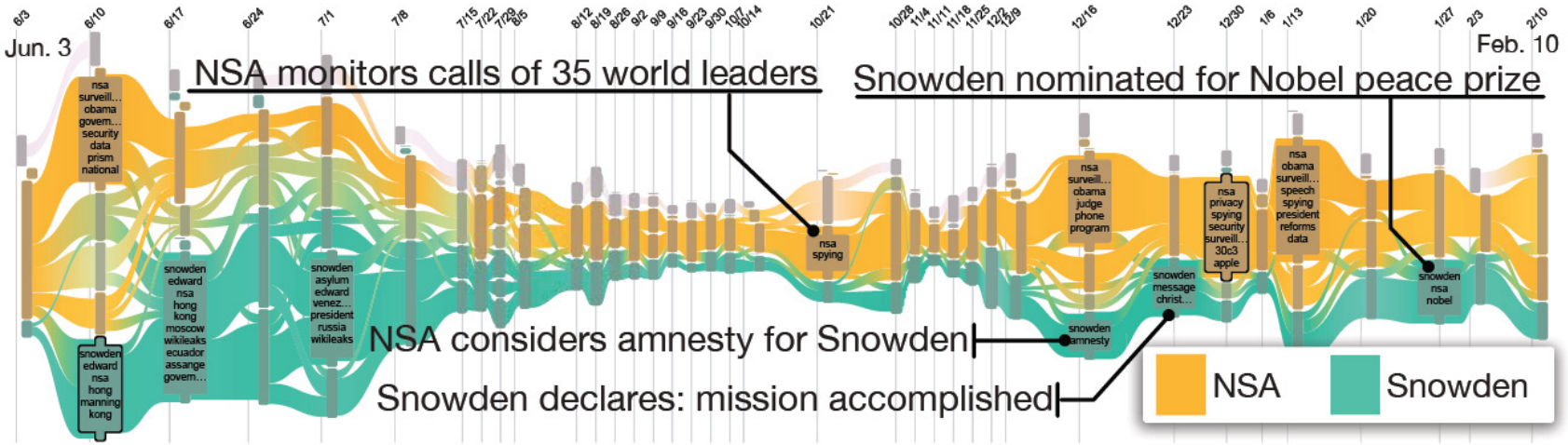
Система Serendip

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.

Динамика тем: эволюция предметной области



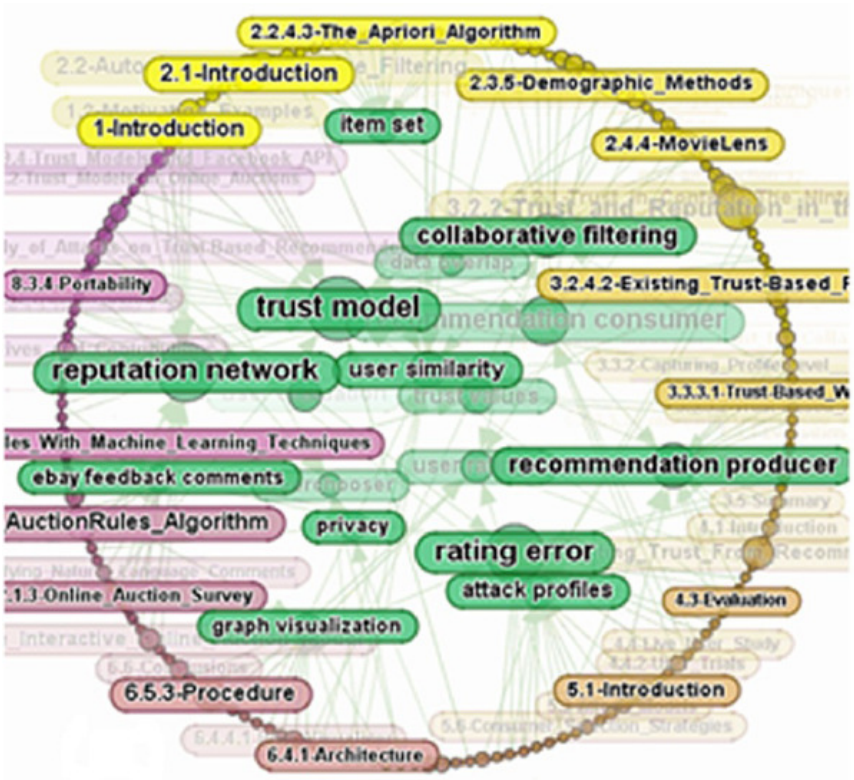
Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

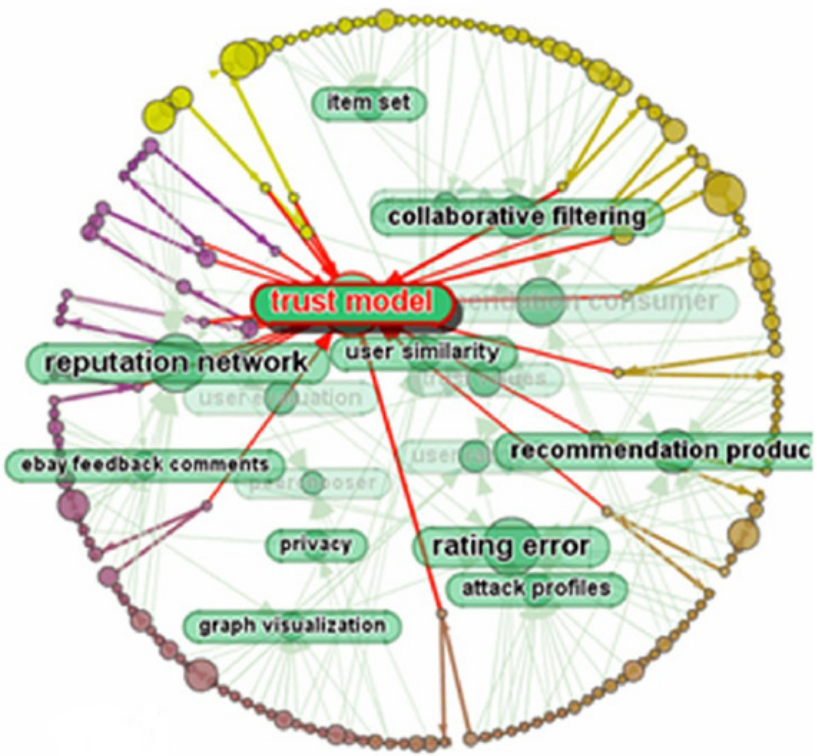
Стратегия визуализации в системах TextFlow и RoseRiver:

- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- получает сгенерированный отчёт с инфографикой.

Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. 2014.

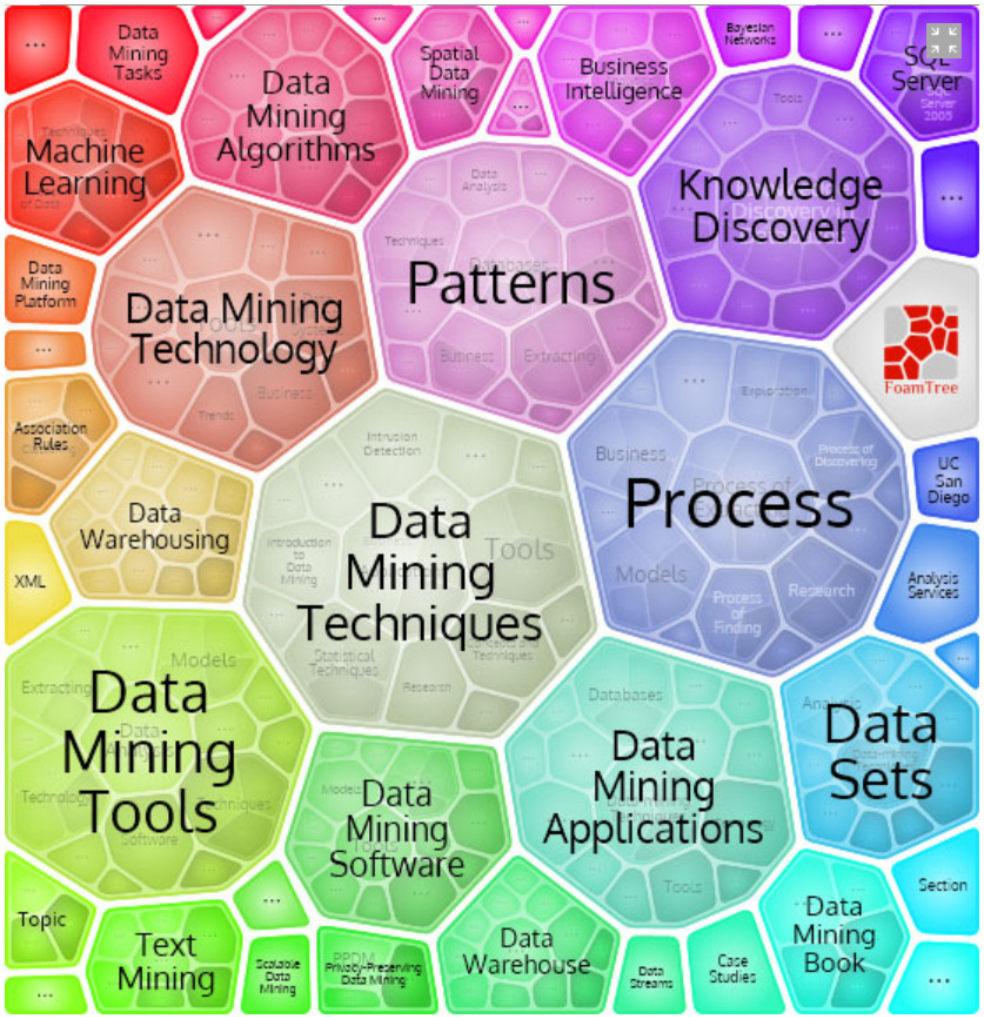
Динамика тем внутри документа: тематическая сегментация





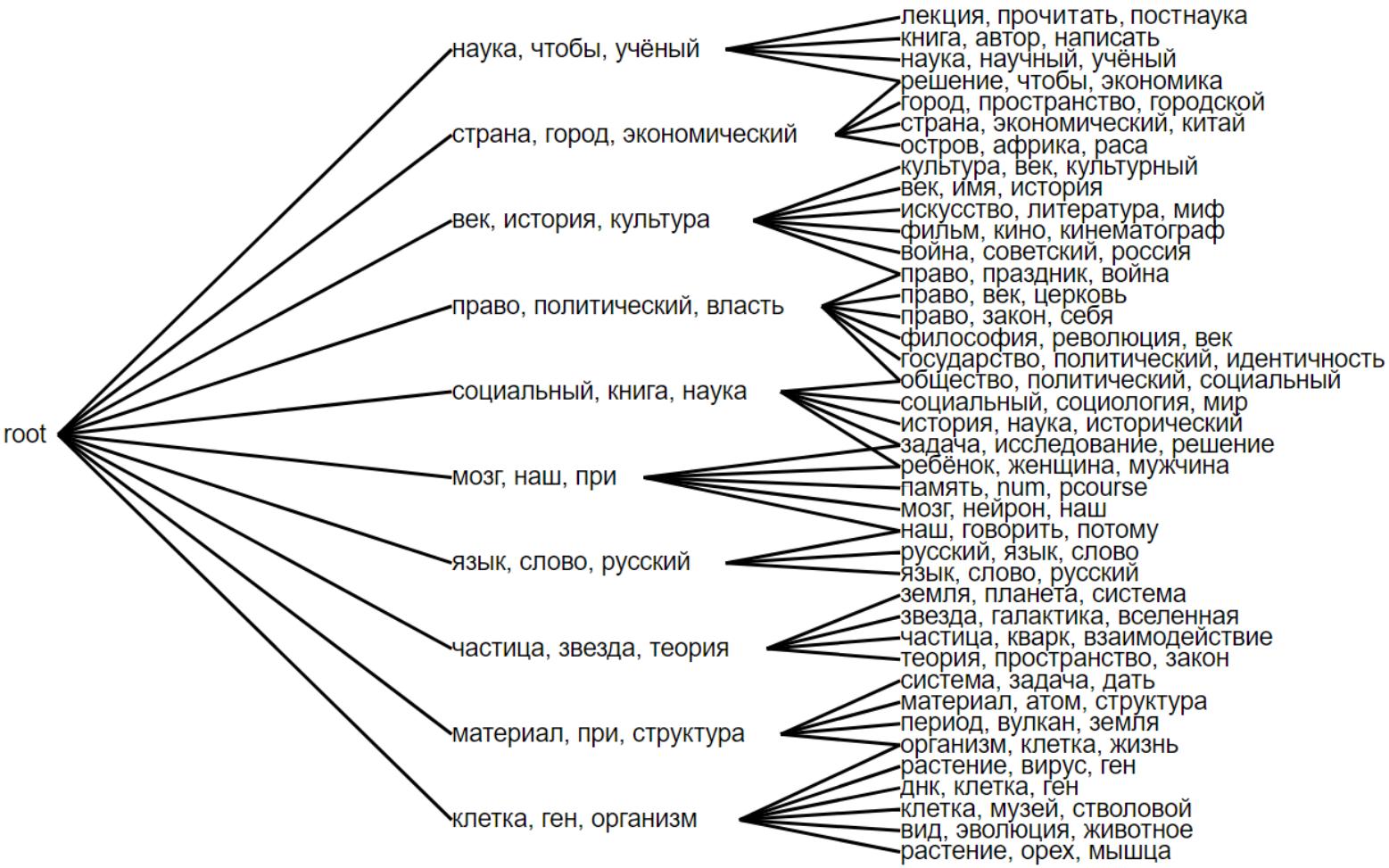
Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P. TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology. 2012.

Пример иерархической карты области *Data Mining*



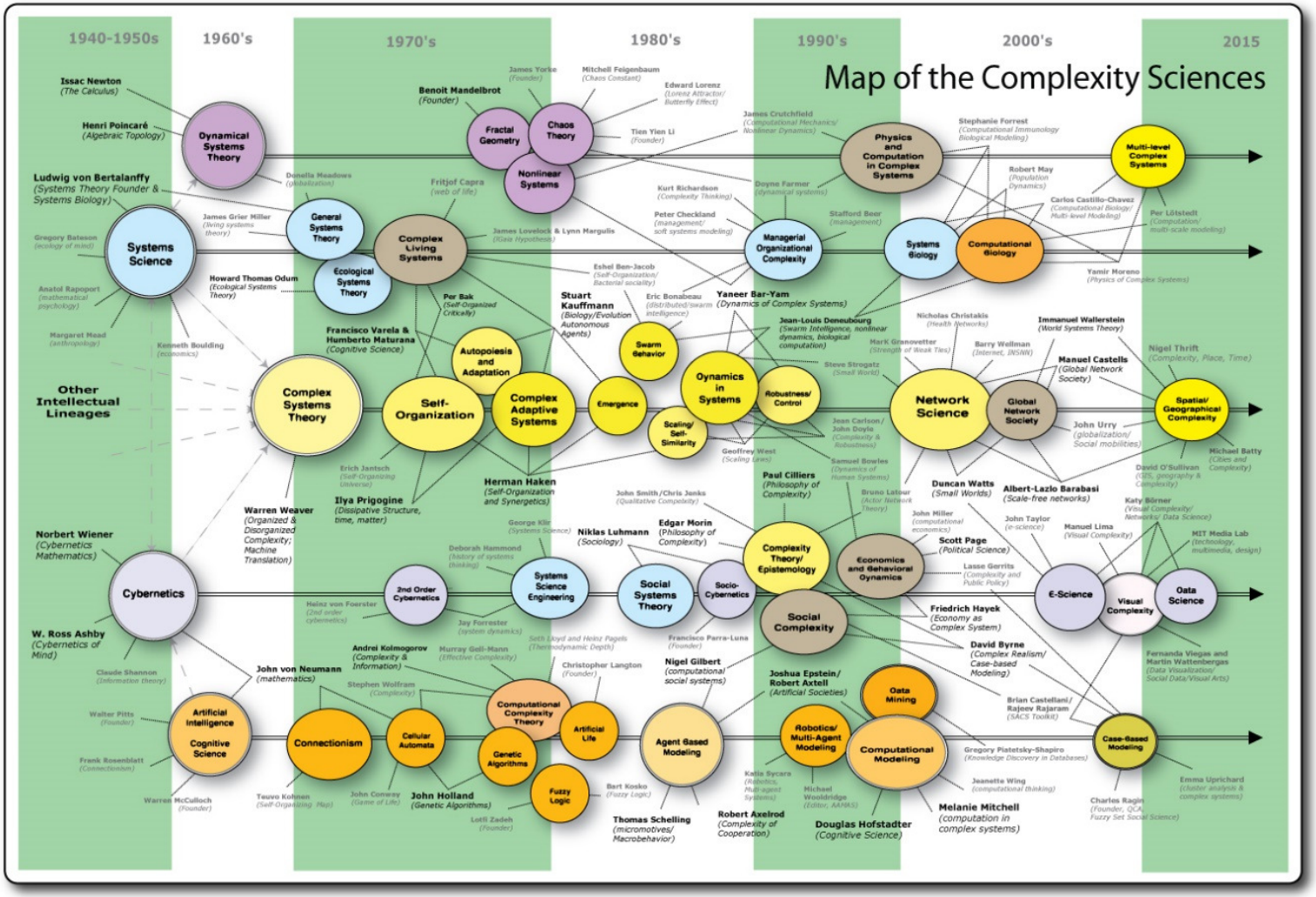
FoamTree: https://carrotsearch.com/foamtree

Иерархический спектр тем (коллекция postnauka.ru)



Дмитрий Федоряка. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация. МФТИ, 2017.

Пример карты предметной области, построенной вручную



http://www.theoryculturesociety.org/brian-castellani-on-the-complexity-sciences

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.

Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // JMLDA, 2015.

Резюме

- Тематические модели не теряют актуальности, несмотря на повальное увлечение БЯМами. Приложений много!
- ARTM — единственная теория для комбинирования и мультизадачного обучения тематических моделей
- BigARTM — эффективная параллельная реализация, не требующая графических карт
- Есть перспективная альтернатива: PyTorch + ARTM
- Давайте вместе доделаем оранжевый Тематизатор!

Воронцов Константин Вячеславович
voron@mlsa-iai.ru

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