• Вероятностные языковые модели • Лекция 8. Примеры прикладных задач и проект «Тематизатор»

Константин Вячеславович Воронцов
k.vorontsov@iai.msu.ru

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

ВМК МГУ • 20 апреля 2026

Содержание

Напоминание. Тематическая модель «мешка термов»

Дано: коллекция текстовых документов D, словарь W;
n_{dw} — частота терма $w \in W$ в документе $d \in D$.

Найти: вероятностную языковую модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$
с параметрами $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

Напоминание. Тематическая модель локальных контекстов

Дано: последовательность $w_1, \ldots, w_n$ термов словаря W;
$C_i \subset \{1, \ldots, n\}$ — локальный контекст терма w_i, $1, \ldots, n$;
α_{ci} — коэффициент внимания, вес терма w_c из C_i для w_i.

Найти: вер. языковую модель $p(w|C_i) = \sum_{t \in T} \phi_{tw} \frac{p(w)}{p(t)} p(t|C_i)$
с параметрами $\phi_{tw} = p(t|w)$

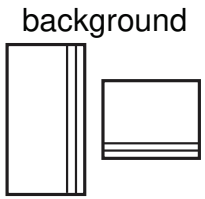
Критерий: $\sum_{i=1}^{n} \ln \sum_{t \in T} \phi_{tw_i} \frac{p(w_i)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} + R(\Phi) \to \max_{\Phi}$

EM-алгоритм (после некоторых насильственных упрощений):

$$\text{E-шаг:} \begin{cases} p_{ti} = \underset{t \in T}{\text{norm}} \left(\frac{\phi_{tw_i}}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} \right), & p(t) = \sum_{w \in W} \phi_{tw} p(w) \end{cases}$$

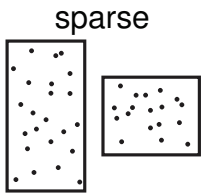
$$\text{M-шаг:} \begin{cases} \phi_{tw} = \underset{t \in T}{\text{norm}} \left(n_{tw} + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} \right), & n_{tw} = \sum_{i=1}^{n} p_{ti} [w_i = w] \end{cases}$$

Регуляризаторы для улучшения интерпретируемости тем



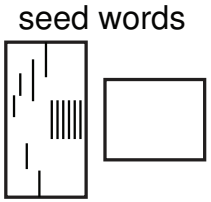
Сглаживание фоновых тем $B \subset T$:

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

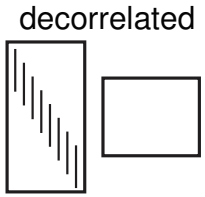


Разреживание предметных тем $S = T \setminus B$:

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

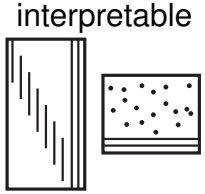


Сглаживание для выделения релевантных тем
с помощью словаря «затравочных» ключевых слов



Декоррелирование для повышения различности тем:

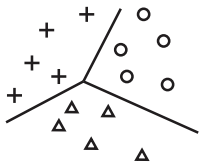
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

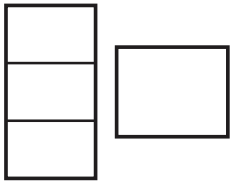
Регуляризаторы для мультимодальных тематических моделей

supervised



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

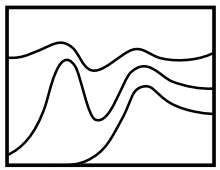
multilanguage



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:

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

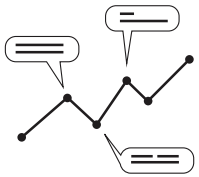
geospatial

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

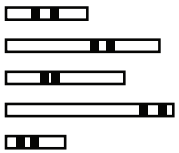
regression



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

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

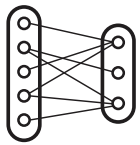
biterm



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

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

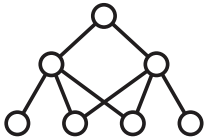
relational



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:

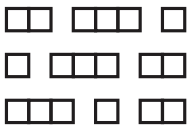
$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st}$$

Регуляризаторы для моделирования последовательного текста

sentence

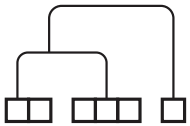
Тематические модели, учитывающие границы предложений, абзацев и секций документов

n-gram



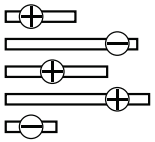
Модели с модальностями *n*-грамм, коллокаций, именованных сущностей (используем TopMine)

syntax



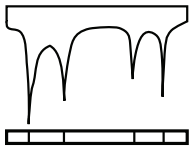
Модели, учитывающие результаты автоматического синтаксического разбора (используем UDPipe)

sentiment



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

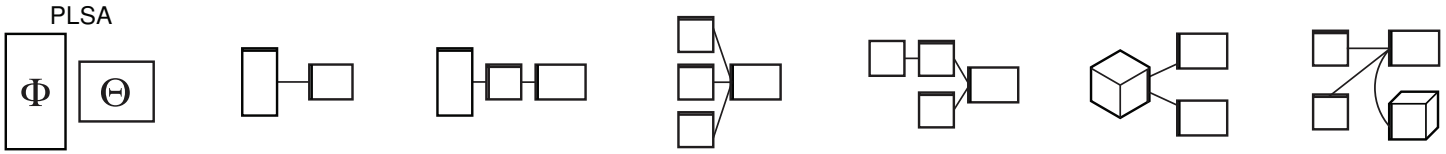
segmentation



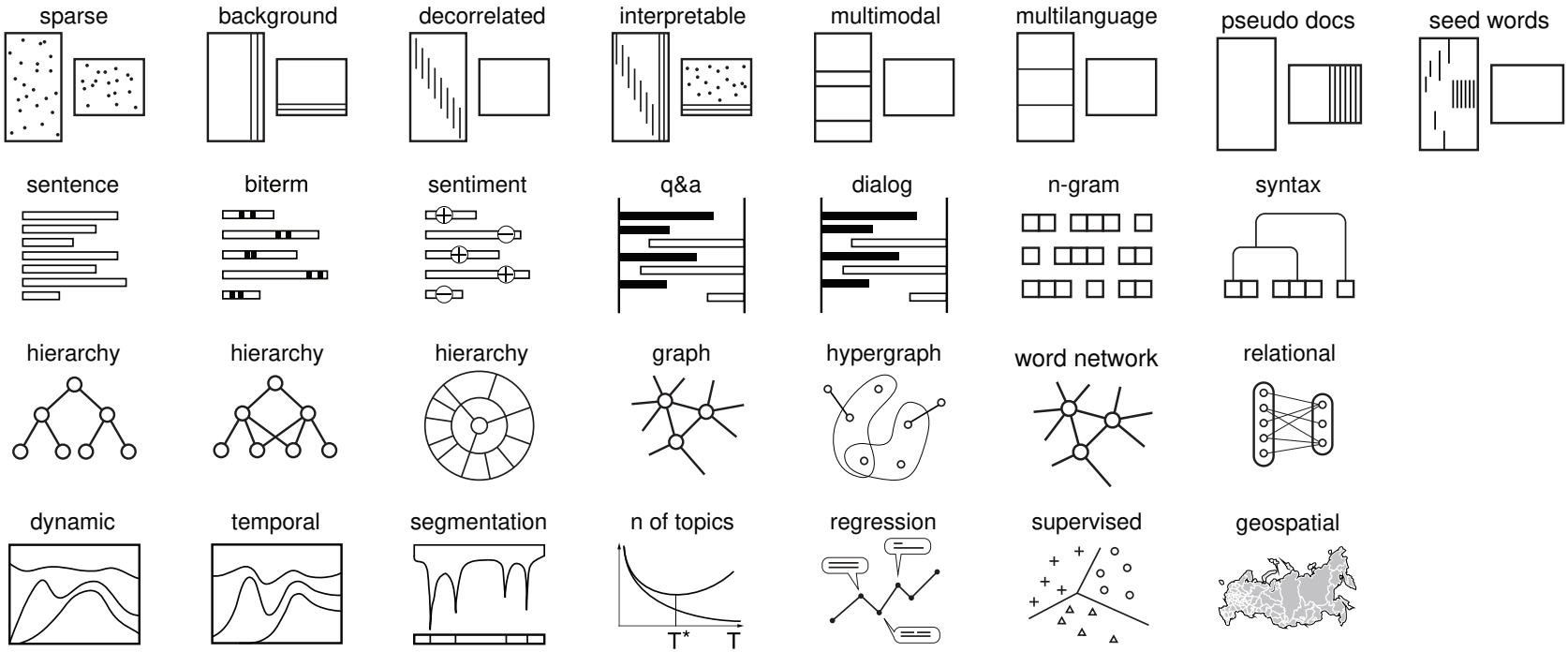
Тематические модели сегментации с автоматическим определением границ сегментов

ARTM: конструктор моделей с заданными свойствами

Структуры матричных разложений в вероятностных моделях:



Регуляризаторы — дополнительные критерии и ограничения:



ARTM: модульный подход к тематическому моделированию

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 мер качества

Сообщество:

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

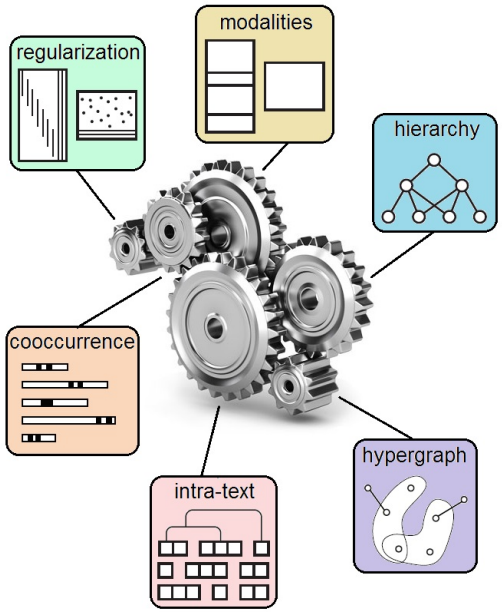
Лицензия и среда разработки:

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

Ключевые возможности библиотек BigARTM и TopicNet

BigARTM

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



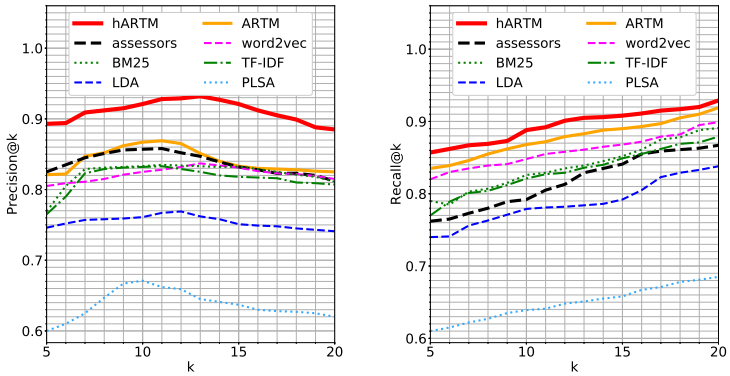
TopicNet

- 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

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:
  $200 \to 1400$ (Habr.ru), $475 \to 2800$ (TechCrunch.com).

А.Янина. Тематические и нейросетевые модели языка для разведочного информационного поиска // диссертация к.ф.-м.н. МФТИ, 2022.

Поиск и рубрикация научных публикаций на 100 языках

Цель: мультиязычный поиск и классификация научных публикаций по рубрикаторам УДК, ГРНТИ, ОЭСР, ВАК

модель	ср.ч. УДК	ср.% УДК	ср.ч. ГРНТИ	ср.% ГРНТИ
Базовая TM	0.558	0.165	0.536	0.220
XLM-RoBERTa	0.835	0.179	0.832	0.288
ARTM	0.995	0.225	0.852	0.366

Регуляризаторы:

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

Результаты:

- точность мультиязычного поиска 94%
- сокращение модели 128 Гб $\to$ 4.8 Гб при редукции словарей (BPE-токенизация) до 11K токенов на каждый язык.

Н.А.Герасименко, П.С.Потапова, А.О.Янина, К.В.Воронцов. Применение вероятностного тематического моделирования в четырёх задачах разведочного информационного поиска // Информационный бюллетень РБА, 2022, №98, С.43–48.

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, 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:

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

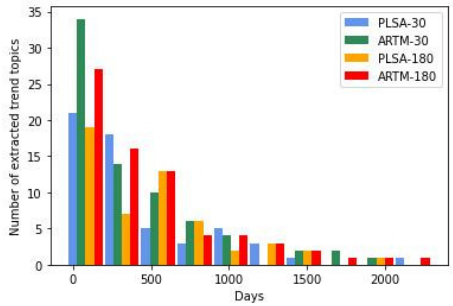
Результаты: число релевантных тем: 45 (LDA) → 83 (ARTM).

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 в области AI/ML 2009–2021 гг.



Регуляризаторы:

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

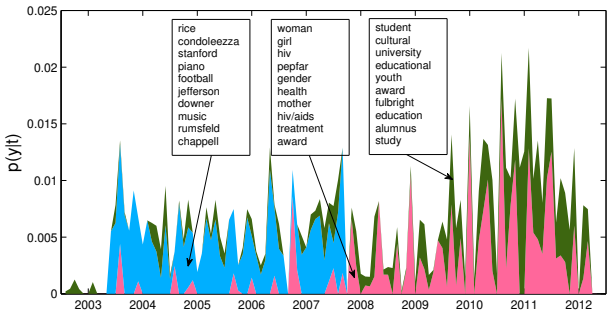
Результаты:

- 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:

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

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:

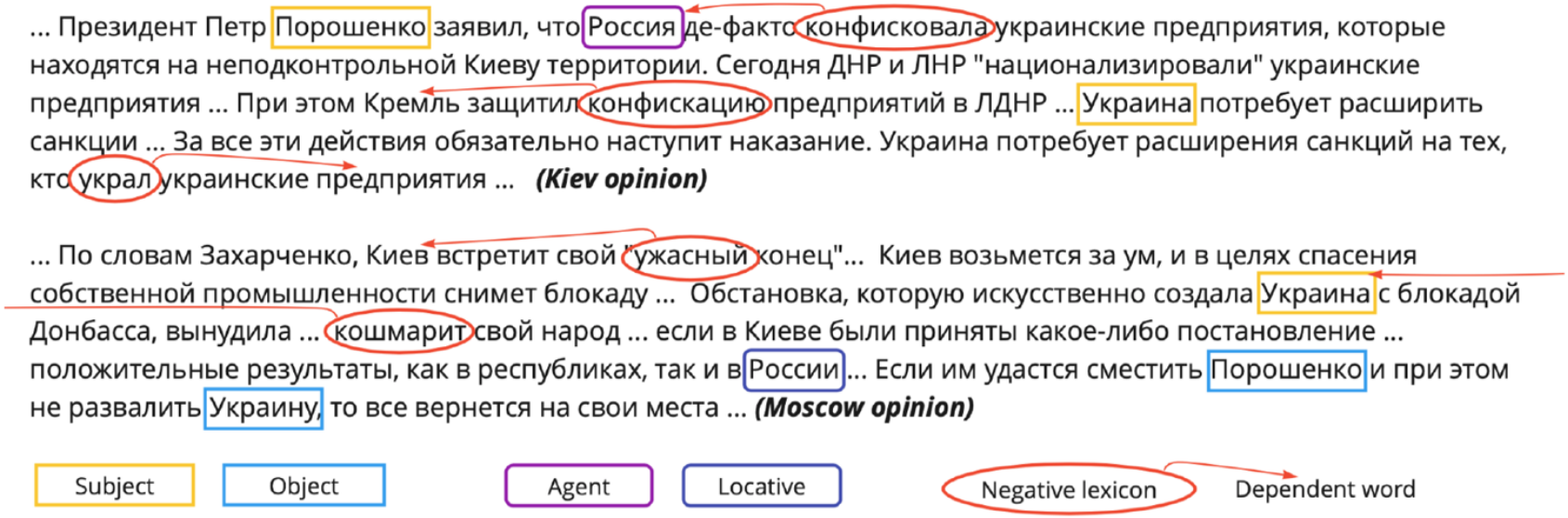
$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) + R\left(\text{syntax}\right) \to \max$$

Результаты:

- выделение мнений внутри тем: F1-мера = 0.86%
- совместное использование трёх модальностей:
  - SPO — факты как триплеты «субъект–предикат–объект»
  - FR — семантические роли слов по Филлмору
  - Sent — 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.

Выделение поляризованных мнений в политических новостях

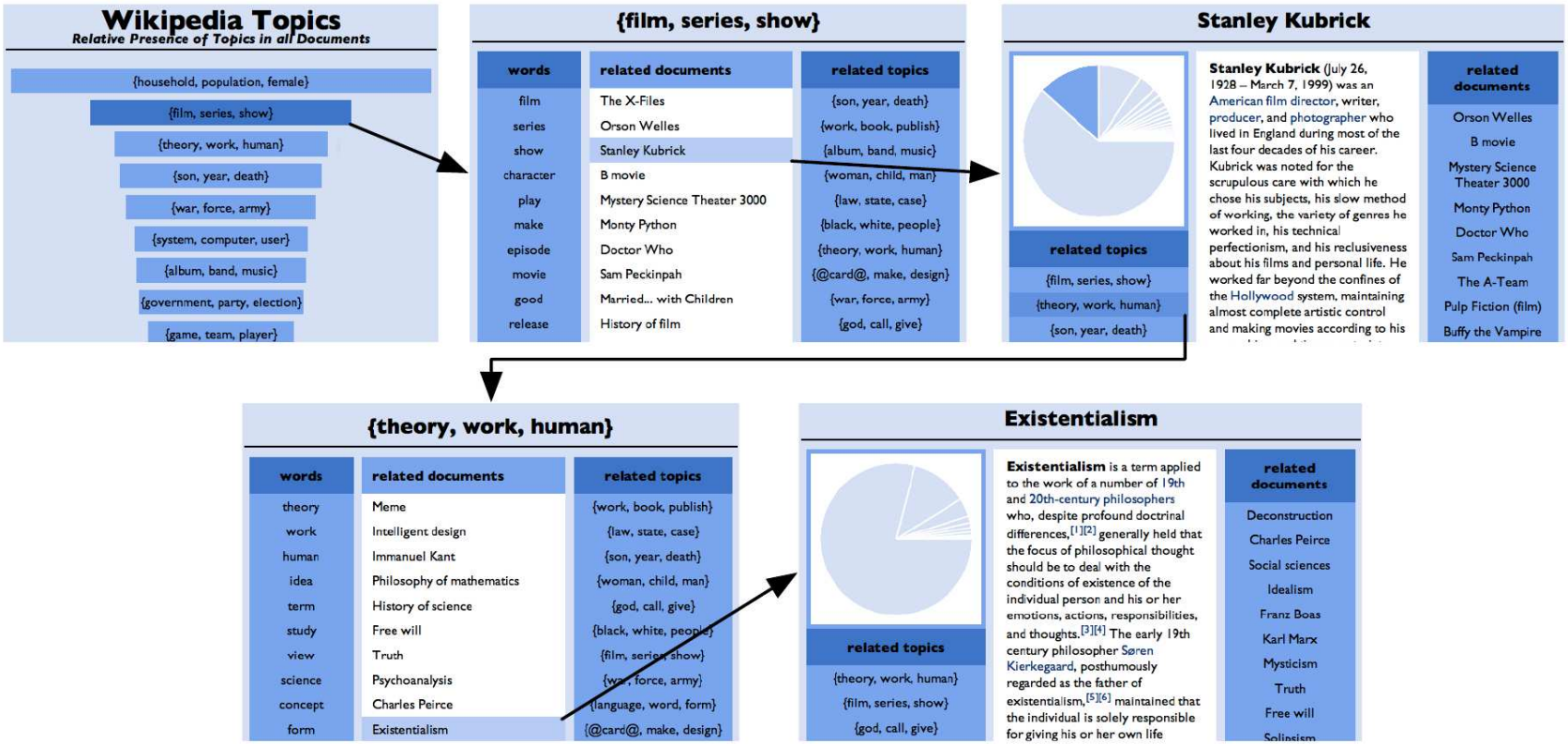


Слова «Порошенко», «Россия», «Украина» встречаются в тексте-1 и тексте-2 одинаково часто, однако:

- «Порошенко» — субъект в тексте-1 и объект в тексте-2;
- «Россия» — агенс в тексте-1 и локация в тексте-2;
- негативная тональность: «Россия», «Кремль» в тексте-1, «Киев», «Украина» в тексте-2.

Визуализация TMVE (Topic Model Visualization Engine)

Тематический навигатор с веб-интерфейсом:

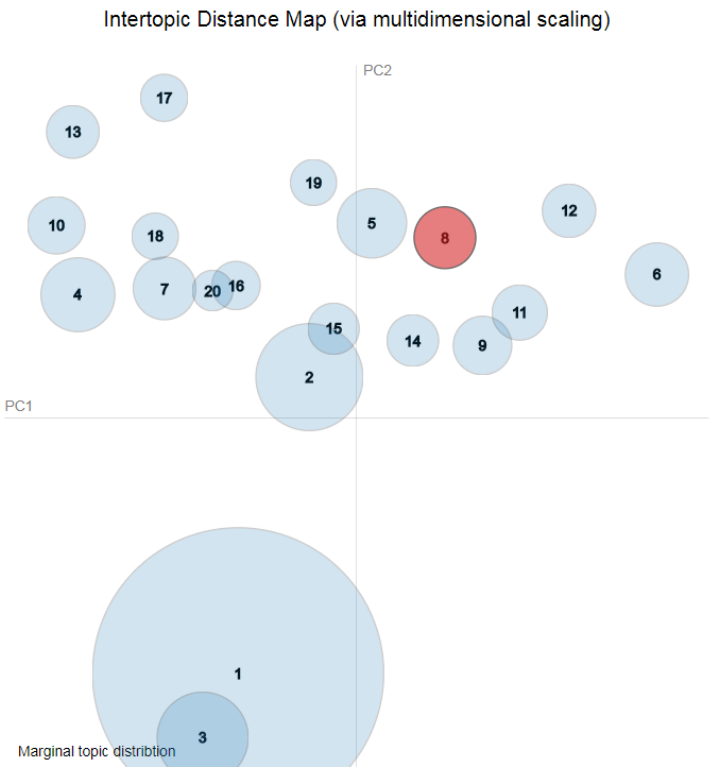


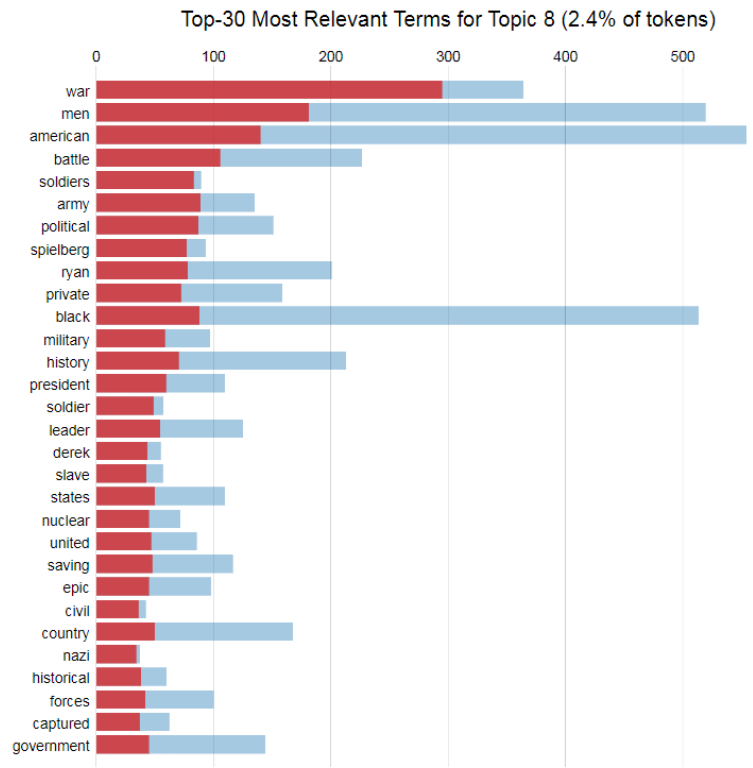
https://github.com/ajbc/tmv

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

Система LDAvis

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



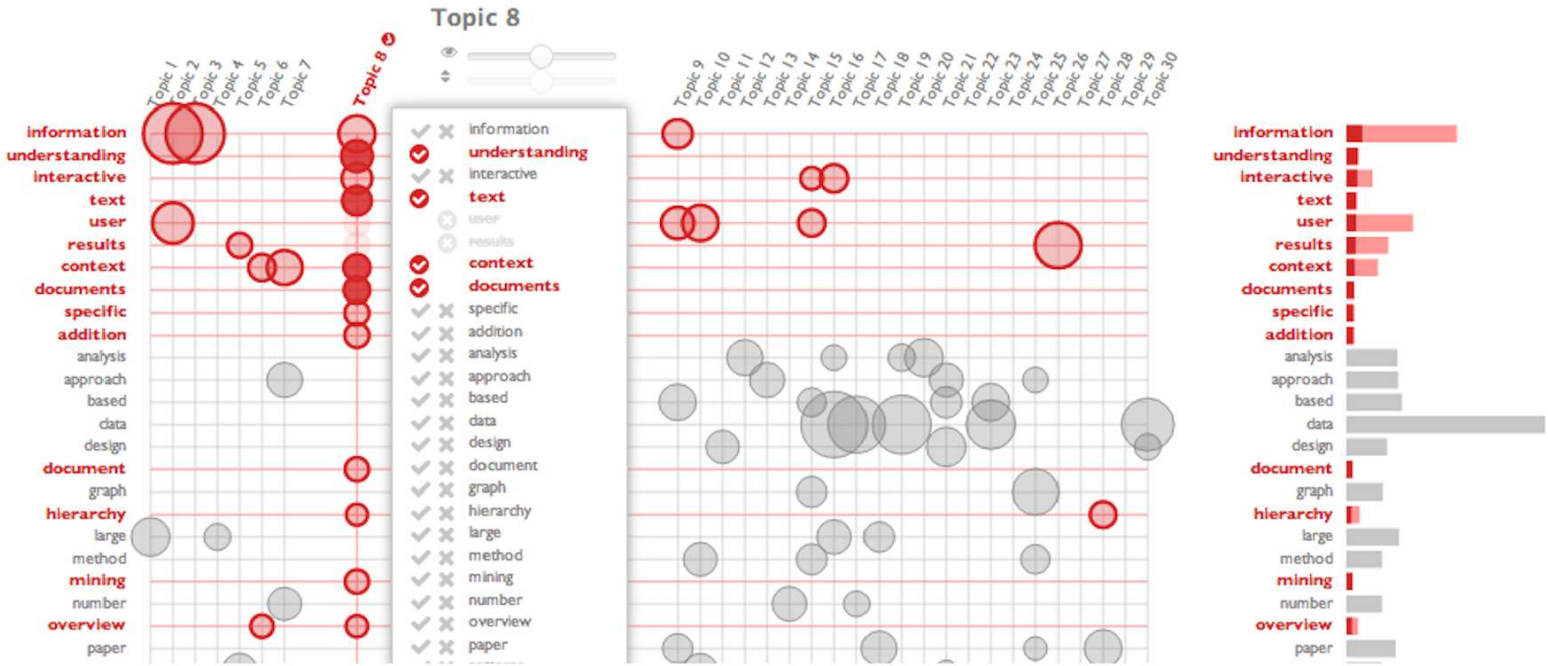


https://github.com/cpsievert/LDAvis

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

Система Termite

Интерактивная визуализация матрицы Φ и сравнение тем:

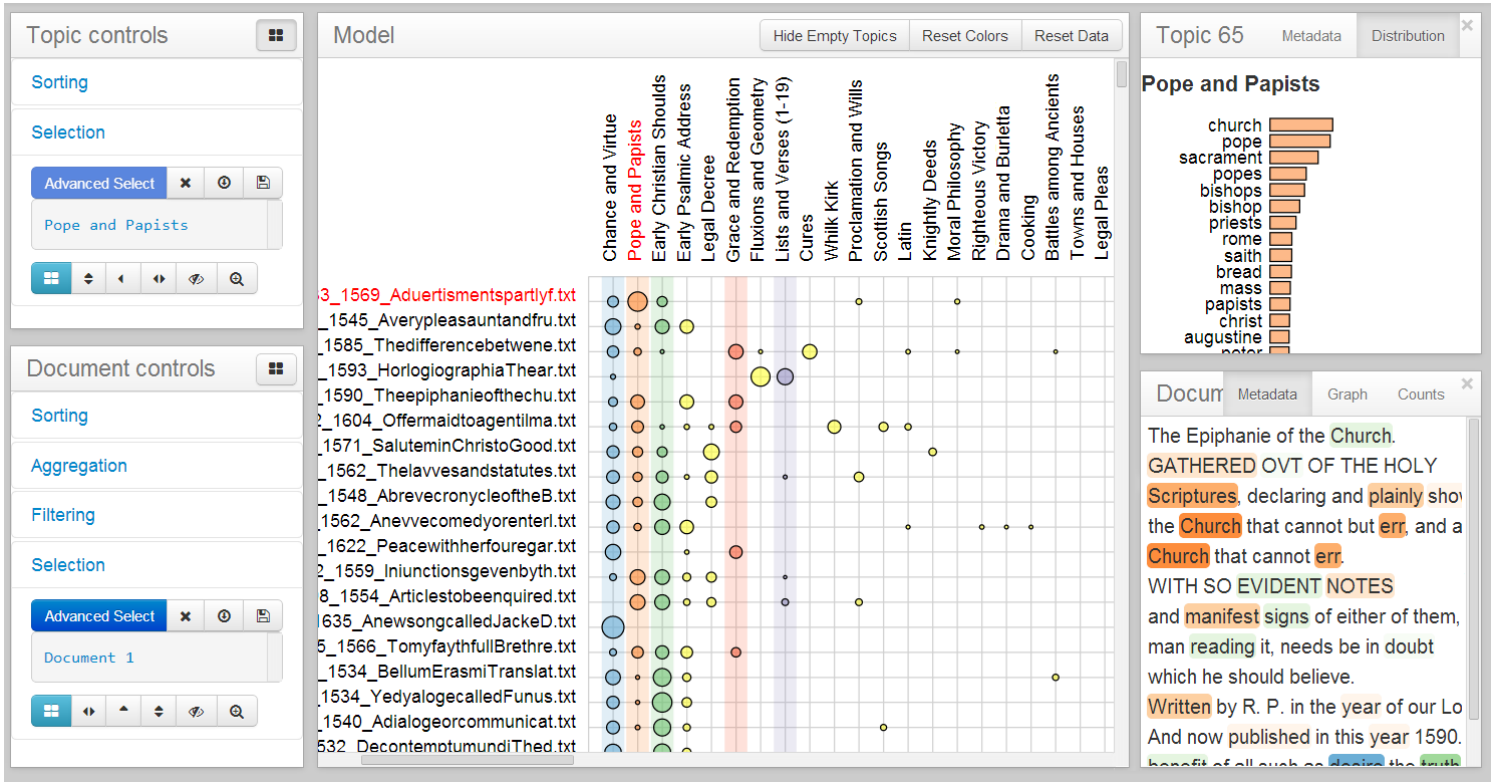


https://github.com/uwdata/termite-visualizations

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

Система Serendip

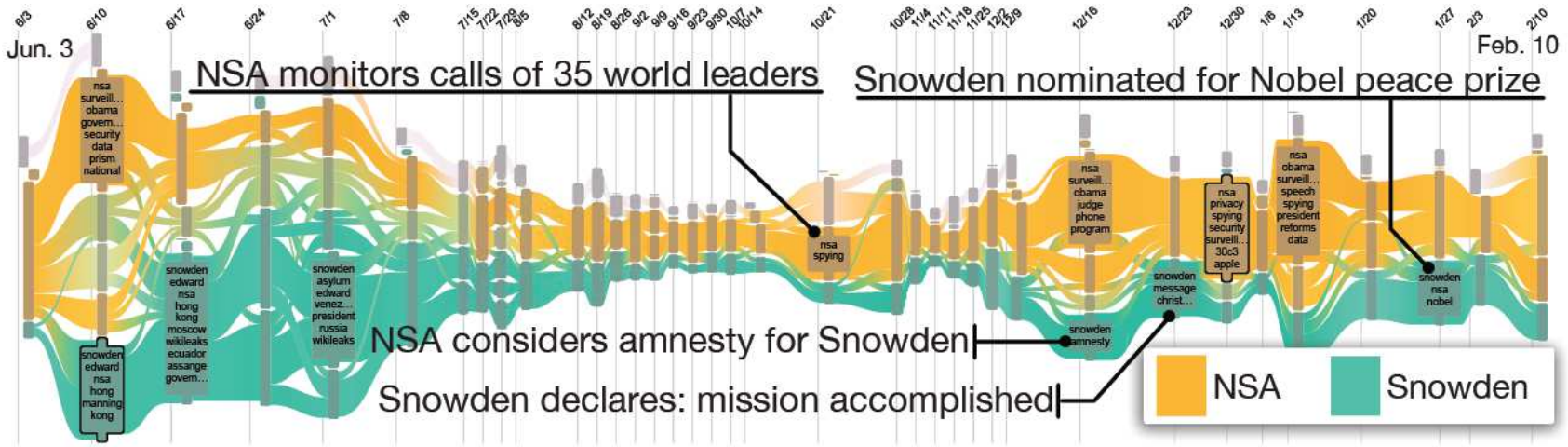
Визуализация матриц Φ, Θ и тематики слов в текстах:



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



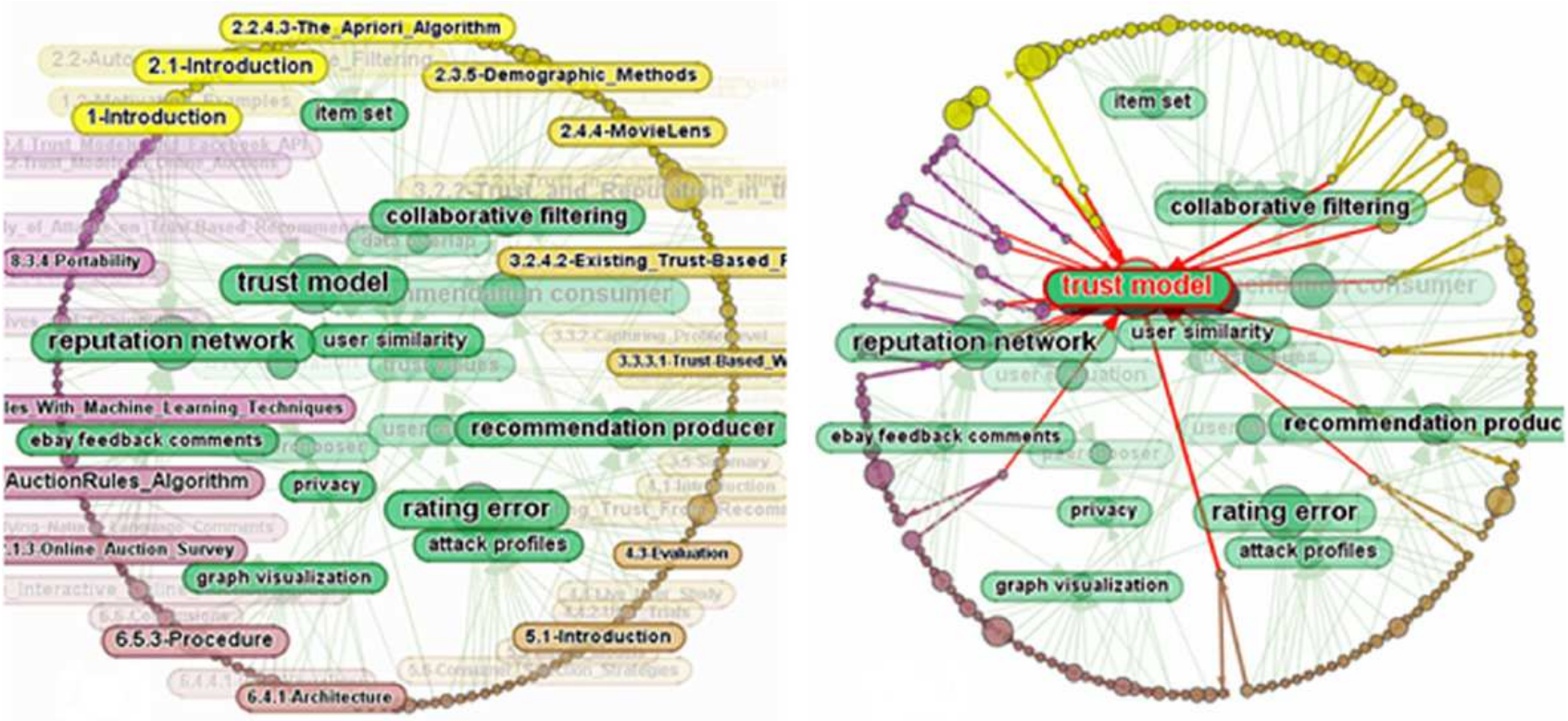
Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

Стратегия визуализации в системах TextFlow и RoseRiver:

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

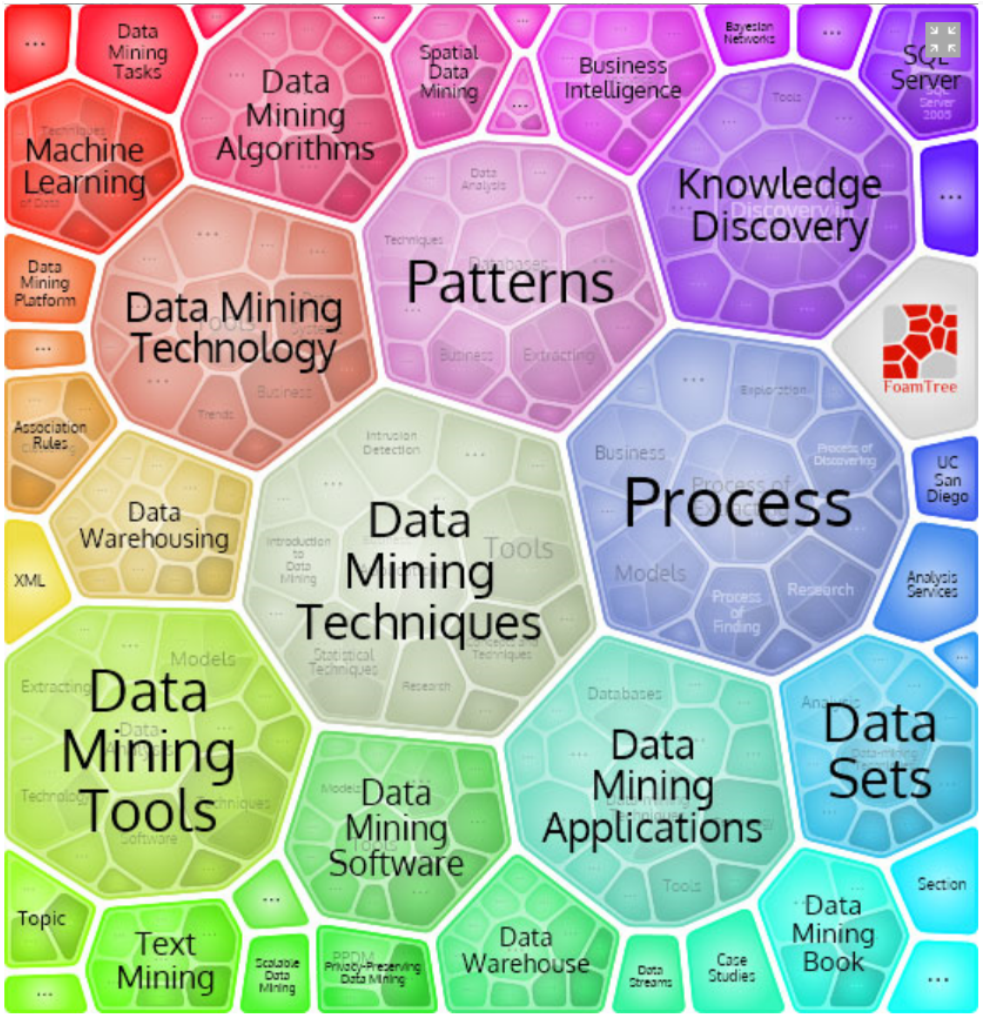
Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. 2014.

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*



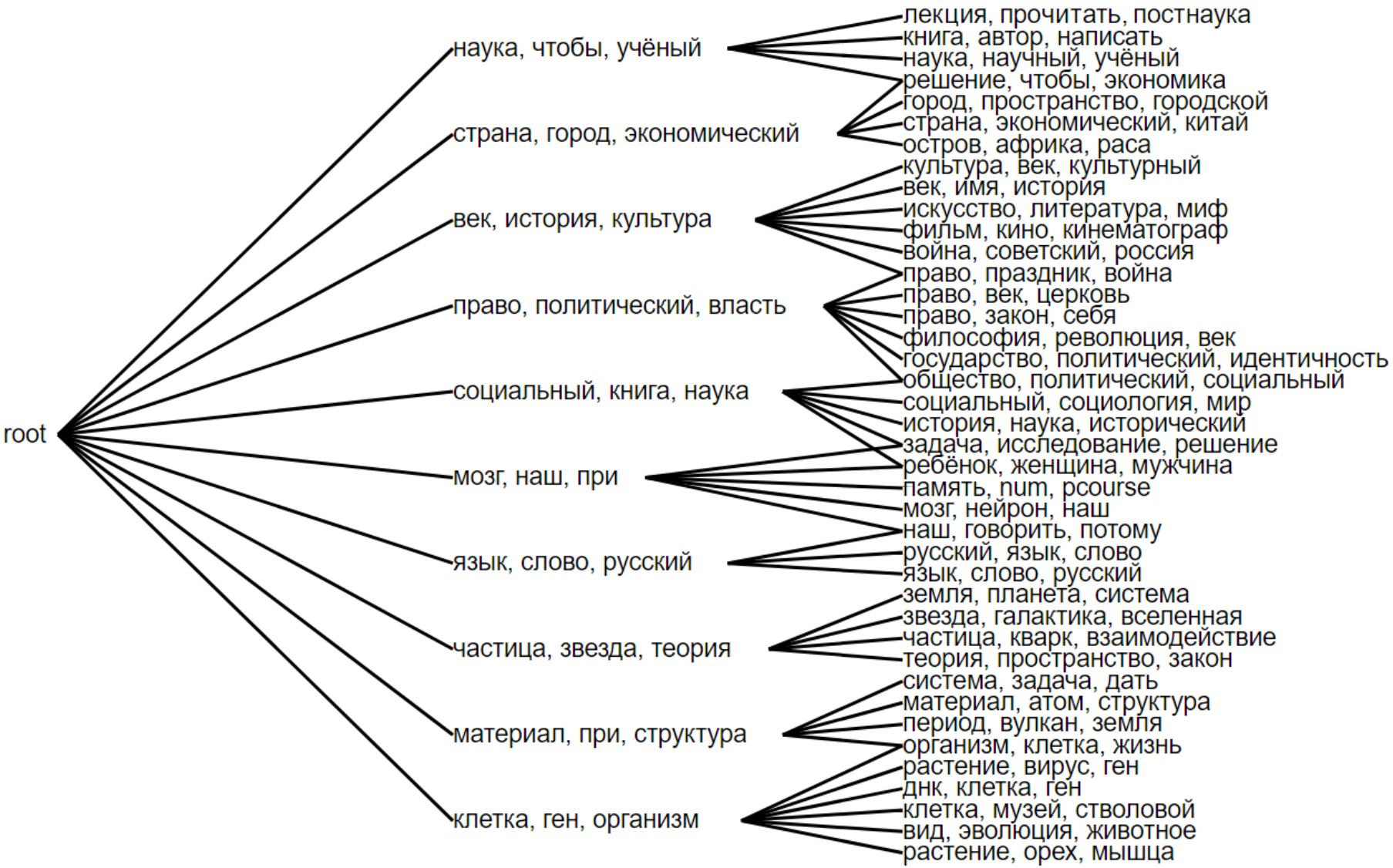
FoamTree: https://carrotsearch.com/foamtree

Библиотека VisARTM для BigARTM (уже не поддерживается)

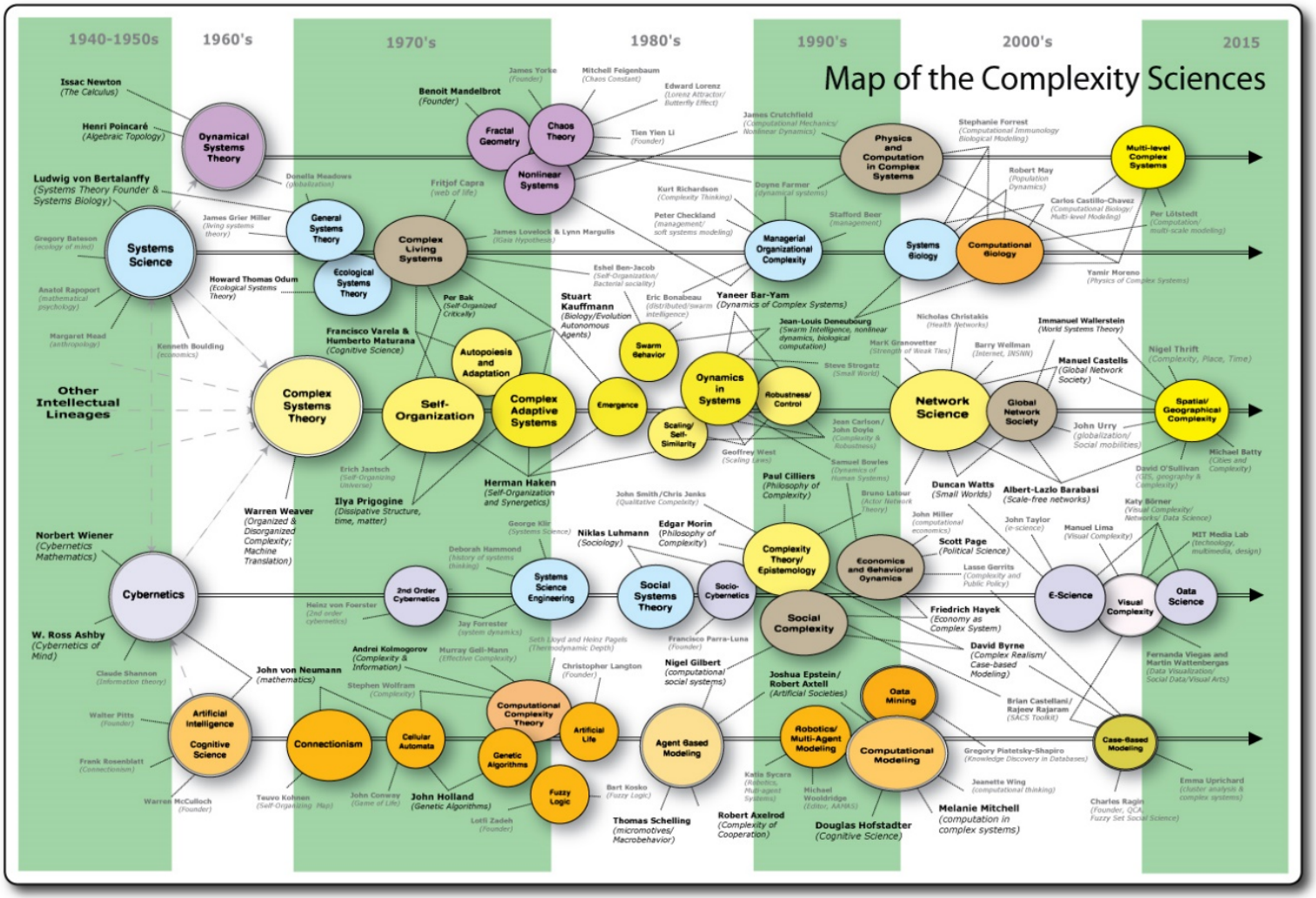
- Web-приложение для визуализации ARTM моделей
- Открытый код: https://github.com/bigartm/visartm
- Автоматическое перестроение моделей через BigARTM
- Текстовые интерактивные визуализации документов, тем, термов, модальностей
- Графическая визуализация иерархических моделей
- Графическая визуализация темпоральных моделей
- Тематические спектры
- Сбор асессорских оценок тем

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

Иерархический спектр тем (коллекция postnauka.ru)



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.

Поиск этно-релевантных тем в социальных сетях

- **Дано:**
  1) данные социальных медиа (ВК и др.)
  2) словарь этнонимов (около 300)
- **Найти:** как можно больше тем
  1) про отдельные этничности
  2) про сочетания этничностей (отношения, конфликты)
- **Критерий:**
  1) интерпретируемость всех тем
  2) точность и полнота поиска этно-релевантных тем

Используемые регуляризаторы:

- сглаживание этно-релевантных тем по словарю этнонимов
- декоррелирование этно-релевантных тем
- модальность этнонимов

Примеры этнонимов (всего около 300)

османский	русич
восточноевропейский	сингапурец
эвенк	перуанский
швейцарская	словенский
аланский	вепсский
саамский	ниггер
латыш	адыги
литовец	сомалиец
цыганка	абхаз
ханты-мансийский	темнокожий
карачаевский	нигериец
кубинка	лягушатник
гагаузский	камбоджиец

Примеры этно-релевантных тем

(русские): русский, князь, россия, татарин, великий, царить, царь, иван, император, империя, грозить, государь, век, московская, екатерина, москва,
(русские): акция, организация, митинг, движение, активный, мероприятие, совет, русский, участник, москва, оппозиция, россия, пикет, протест, проведение, националист, поддержка, общественный, проводить, участие,
(славяне, византийцы): славянский, святослав, жрец, древние, письменность, рюрик, летопись, византия, мефодий, хазарский, русский, азбука,
(сирийцы): сирийский, асад, боевик, район, террорист, уничтожать, группировка, дамаск, оружие, алесио, оппозиция, операция, селение, сша, нусра, турция,
(турки): турция, турецкий, курдский, эрдоган, стамбул, страна, кавказ, горин, полиция, премьер-министр, регион, курдистан, ататюрк, партия,
(иранцы): иран, иранский, сша, россия, ядерный, президент, тегеран, сирия, оон, израиль, переговоры, обама, санкция, исламский,
(палестинцы): террорист, израиль, терять, палестинский, палестинец, террористический, палестина, взрыв, территория, страна, государство, безопасность, арабский, организация, иерусалим, военный, полиция, газ,
(ливанцы): ливанский, боевик, район, ливан, армия, террорист, али, военный, хизбалла, раненый, уничтожать, сирия, подразделение, квартал, армейский,
(ливийцы): ливан, демократия, страна, ливийский, каддафи, государство, алжир, война, правительство, сша, арабский, али, муаммар, сирия,

Примеры этно-релевантных тем

(евреи): израиль, израильский, страна, война, нетаньяху, тель-авив, время, сша, сирия, египет, случай, самолет, еврейский, военный, ближний,
(американцы): американский, американка, война, россия, военный, страна, вашингтон, америка, армия, конгресс, сирия, союзный, российский, обама, войска, русский, оружие, операция,
(немцы): армия, война, войска, советский, военный, дивизия, немец, фронт, немецкий, генерал, борт, операция, оборона, русский, бог, победа,
(немцы): германий, немец, германский, ссср, немецкий, война, старое, советский, россия, береза, русский, правительство, территория, полный, документ, вопрос, сорт, договор, отношение, франция,
(евреи, немцы): еврей, еврейский, холодный, германий, антисемитизм, гетра, немец, синагога, сша, израиль, малиновского, комиссия, нацбол, документ, война, еврейка, миллион, украина,
(украинцы, немцы): украинский, упс, оун, немец, немецкий, ковальков, хохол, волынский, бандера, организация, россиянин, советский, русский, польский, армия, шухевича, ровенский,
(таджики, узбеки): мигрант, страна, россия, миграция, азия, нелегальный, миграционный, таджикистан, гастарбайтер, гражданка, трудовой, рабочий, фмс, коренево, среднее, узбекистан, таджик, проблема, русский, население,
(канадцы): команда, игра, игрок, канадский, сезон, хоккей, сборная, играть, болельщик, победа, кубок, счет, забирать, хоккейный, выигрывать, хоккеист, чемпионат, шайба,

Примеры этно-релевантных тем

(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,

Результат: модель ARTM находит больше этно-тем

Число этно-релевантных тем, найденных моделью:

модель	этно-тем	фон.тем	++	+−	−+	всего
PLSA	300		9	11	18	38
PLSA	400		12	15	17	44
ARTM-1	200	100	18	33	20	71
ARTM-1	250	150	21	27	20	68
ARTM-2	200	100	28	23	23	74
ARTM-2	250	150	38	42	30	104

Регуляризаторы ARTM-1:

этно темы: разреживание, декоррелирование, сглаживание этнонимов

фоновые темы: сглаживание, разреживание этнонимов

Регуляризаторы ARTM-2:

ARTM-1 + модальность этнонимов

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 с выделением 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 и тематическая классификация 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.

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 :(

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

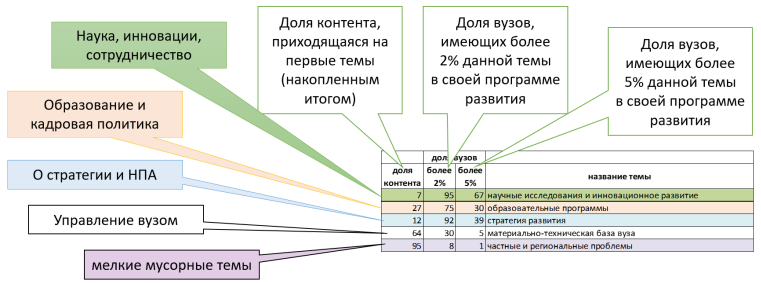
- Для интерпретируемости тем важны биграммы
- Модель построили примерно с 10-й попытки (подбирали число тем, регуляризацию, добивались различности тем)
- Интерпретация 50 тем заняла примерно 20 минут работы
- Иногда выделялись темы исследований и разработок, но для этого нужна более гранулированная модель
- 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 контента	доля вузов более 2%	доля вузов 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 контента	доля вузов более 2%	доля вузов 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

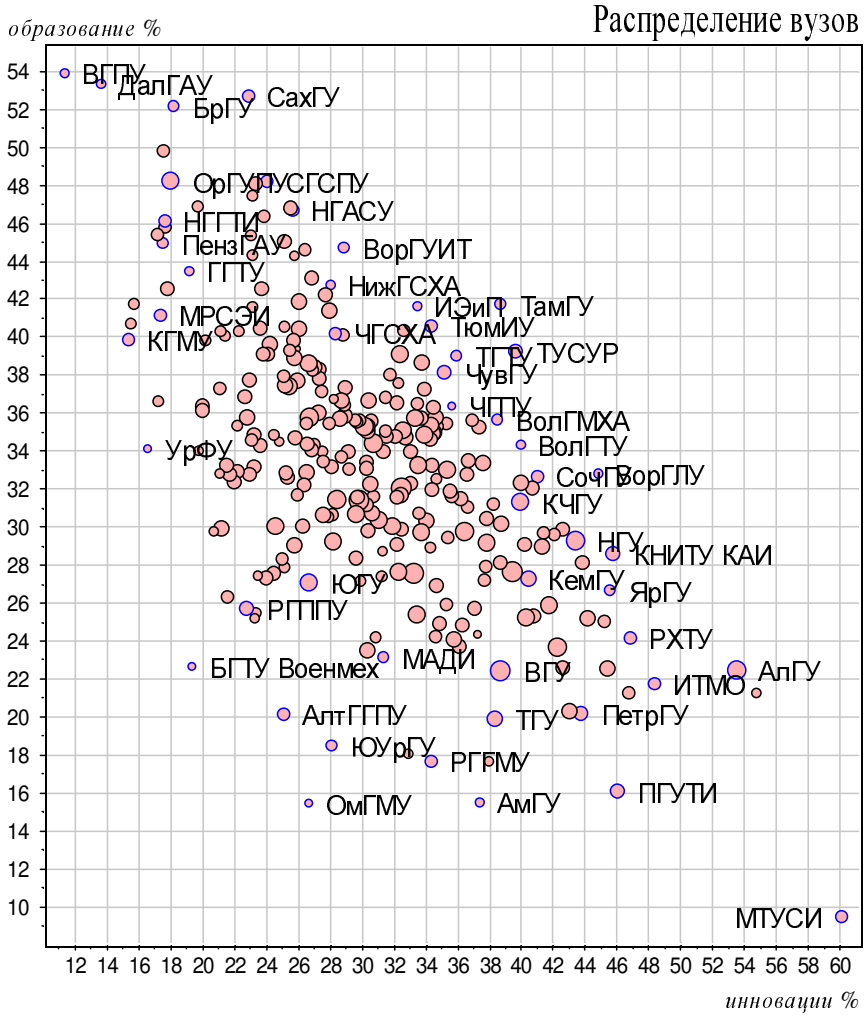
По осям:

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 :)



Исторические исследования: 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.

Исторические исследования: 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.

Исторические исследования: научная и литературная 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 ООН 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.

ТМ 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.

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

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

Функциональные требования (по приоритетности)

1. Визуализация спектра тем и их характеристик
2. Вывод для каждой темы названия, фраз, суммаризации
3. Определение динамики тем во времени
4. Разбиение тем на подтемы иерархически
5. Возможность группировки тем вручную
6. Возможность задавать словари затравок для (групп) тем
7. Отбор и накопление «банка релевантных тем»
8. Тематическая фильтрация коллекции
9. Выявление коротких тем-событий и долгих тем-трендов
10. Выявление связей тем по сочетаемости в документах
11. Тематический поиск по документу или фрагменту
12. Рекомендательный поиск и построение подборок

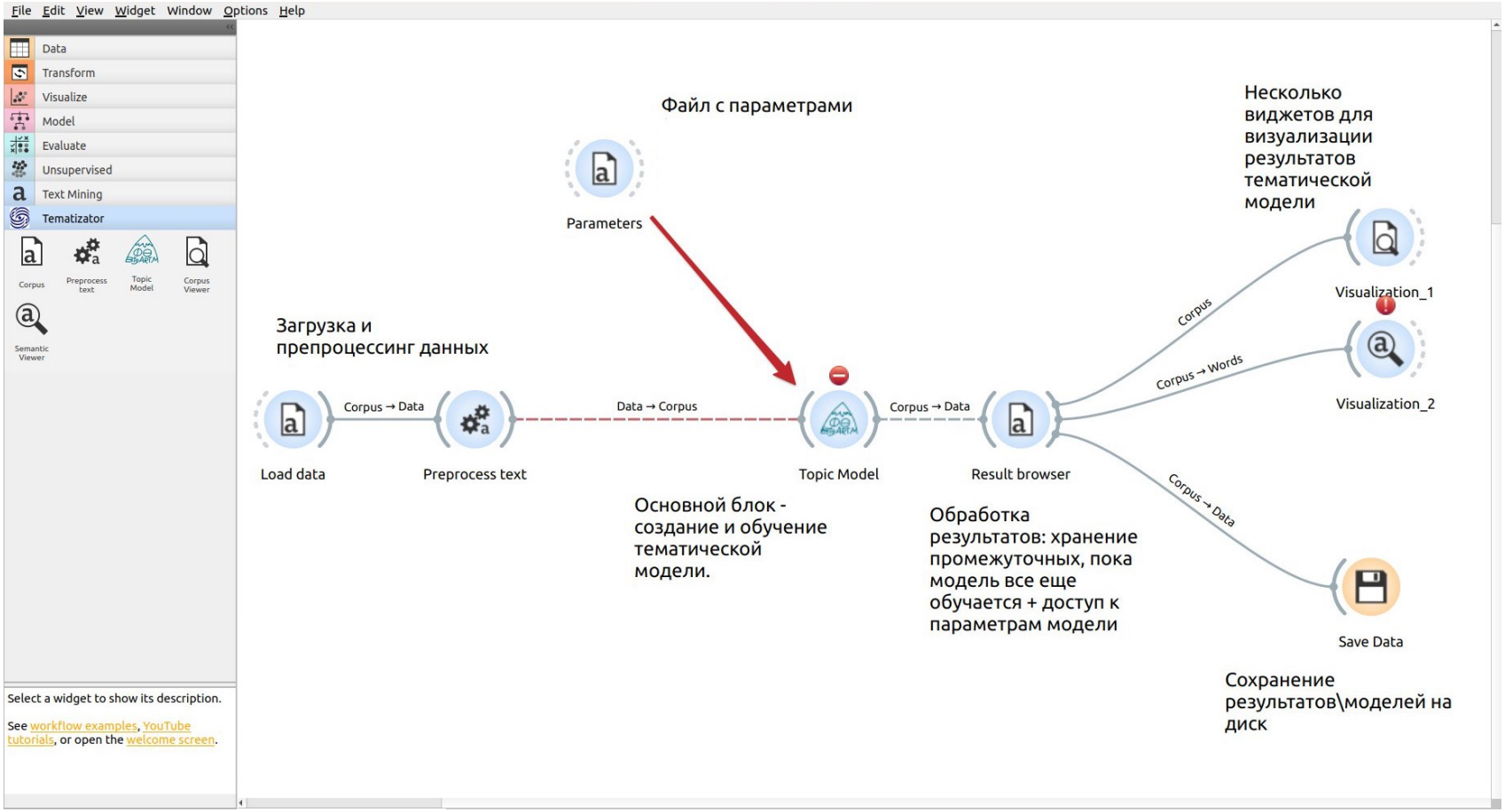
Требования к интерпретируемости (по приоритетности)

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. Решена проблема несбалансированности тем
5. Нет мусорных, дублирующих, плохо интерпретируемых тем
6. Автоматически генерируются суммаризации тем
7. Автоматически выбираются названия тем
8. В иерархии имена дочерних тем уточняют родительские
9. Генерируются суммаризации нетекстовых термов
10. Длинные тексты разбиваются на тематические сегменты
11. Тематика слов согласуется с их локальными контекстами
12. Короткие тексты объяснимо наследуют тематику их слов

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

BigARTM в среде визуального программирования 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

Вместо резюме. Жадная минимизация требований к MVP

Начальная интеграция со средой Orange уже выполнена.

1. **Загрузка**
   - подготовить несколько коллекций для тестирования
   - «мешок слов» в формате Vowpal Wabbit (BigARTM)
   - модальности: языки, время, n-граммы и т.д.
2. **Моделирование**
   - отображение статуса обработки пакетов и текущих метрик
   - возможность прервать обучение и перейти к анализу
   - несколько полезных регуляризаторов — встроены
3. **Визуализация**
   - навигация по темам в духе TMVE
   - спектр тем по семантической близости и релевантности
4. **Коррекция**
   - разметка тем на релевантные, нерелевантные, мусорные
   - перестроение модели с сохранением релевантных тем

Задания по курсу

Задача-минимум: научиться решать задачи анализа текстов с использованием тематического моделирования

Задача-максимум: получить новый научный результат

виды деятельности	оценка
теоретическая задача	X
теоретическая задача*	$2X$
теоретическая задача**	$3X$
решение прикладной задачи	$10X$
обзор по последним PTM/NTM	$10X$
участие в проекте	$20X$
работа над открытой проблемой	$25X$

где X — оценка за вид деятельности по 5-балльной шкале.
score — суммарная оценка по всем видам деятельности.

Итоговая оценка: $\min\left(5, \lfloor \text{score}/20 \rfloor\right)$ по 5-балльной шкале.

Задания к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Биграммная модель коллекции: $p(w|v) = \xi_{wv}$,
где v — слово, идущее в тексте перед w.
Найти параметры модели ξ_{wv}.

2. Биграммная модель документов: $p(w|v,d) = \xi_{dwv}$.
Найти параметры модели ξ_{dwv}.

Подсказка: применить условия ККТ или основную лемму.

3*. Творческое задание (возможны разные решения).
Предложите модель, разделяющую роли слов в текстах:
— тематические слова
— специфичные слова документа (шум)
— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т}, \text{ш}, \text{ф}\}$.
Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.
Подсказка 3: можно использовать документную частоту слов.

Задания к лекции 2

4. Пользуясь основной леммой, докажите, что регуляризатор битермов эквивалентен добавлению псевдодокументов d_u в исходную коллекцию (см. слайд 13)

Прикладная исследовательская задача:
автоматическое выделение научных терминов (ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  метод ATE на основе комбинирования ARTM и TopMine;
  обоснование, что синтаксический анализ не нужен;
  зависимость качества ATE от объёма коллекции
- Критерий:
  качество ATE (Prec, Rec, F1) на размеченных данных

Теоретическое задание к лекции 3

Выведете EM-алгоритм для тематической языковой модели:

5. $p(w|d) = \sum_t \phi_{wt}\theta_{td}$, используя в качестве исходных данных последовательность $(d_i, w_i)_{i=1}^n$ вместо счётчиков n_{dw}.
Докажите эквивалентность обычному EM-алгоритму ARTM.

6. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$, где $p(t)$ фиксировано,
$\phi_{tw} = p(t|w)$, $\theta_{td} = p(t|d)$ — параметры модели.

7. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$, где $p(t)$ фиксировано,
$\phi_{tw} = p(t|w)$ — параметры модели, $\theta_{td} = \sum_w \frac{n_{dw}}{n_d}\phi_{tw}$.

8*. Фиксация $p(t)$ как внешнего параметра упрощает выкладки, но может нарушать условия целостности модели:

$$p(t) = \sum_t \phi_{tw}p(w), \qquad p(t) = \sum_t \theta_{td}p(d).$$

Как обеспечить выполнение этих условий в EM-алгоритме?

Теоретическое задание к лекции 4

9. Докажите, что необходимым условием максимума

$$\sum_{i=1}^{n} \ln \sum_{t \in T} p(w_i, t|i, \Omega) \;\to\; \max_{\Omega}$$

для языковой модели со скрытыми переменными $t \in T$ (не обязательно темами) и параметрами $\Omega = (\omega_{kj})$ — набором неотрицательных нормированных векторов, является система

$$\text{E-шаг:} \quad p(t|w_i, i) = \operatorname*{norm}_{t \in T} p(w_i, t|i, \Omega)$$

$$\text{M-шаг:} \quad \omega_{kj} = \operatorname*{norm}_{k} \left(\sum_{i=1}^{n} \sum_{t \in T} p(t|w_i, i)\, \omega_{kj} \frac{\partial}{\partial \omega_{kj}} \ln p(w_i, t|i, \Omega) \right)$$

10. Выведете отсюда EM-алгоритм для частных случаев:

1) $p(w, t|i, \Omega) = \phi_{wt}\theta_{td_i}$;
2) $p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{w \in d_i} \frac{n_{d_i w}}{n_{d_i}} \phi_{tw}$;
3) $p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c}$.

11. Творческое задание.** Предложите способ ввести обучаемые параметры в тематическую модель внимания.

Исследовательская задача к лекции 4

Реализуйте EM-алгоритм для модели локального контекста (или воспользуйтесь готовой реализацией)

Исследуйте зависимость метрик качества модели

- перплексия: $\mathscr{P} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} p(w|C_i)\right)$
- разреженность, различность, когерентность тем
- дефекты целостности модели:
  $\left\|p(t) - \frac{n_t}{n}\right\|$, $\left\|p(t) - \sum_t \phi_{tw} p(w)\right\|$, $\left\|p(t) - \sum_t \theta_{td} p(d)\right\|$

от номера итерации и от параметров модели:

- $|T|$ — число тем
- L — число проходов
- τ — вес N_{tw} в формуле M-шага, особый случай $\tau = 0$
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, β — баланс левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, секций
- опция «$i \in C_i$ или $i \notin C_i$»

Теоретическое задание к лекции 5

12. Найдите дискретное распределение $P = (p_i)_{i=1}^n$ в задаче $\sum_i n_i \mu(p_i) \to \max$ с гладкой монотонно возрастающей $\mu(p)$.
Отдельно рассмотрите случаи $\mu(p) = p^s$, $s = 1$, $s \to 0$.

13. Выведете EM-алгоритм в случае, когда ln заменён гладкой монотонно возрастающей функцией μ:

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

Подумайте, какие замены логарифма полезны, и почему.

14. Простейшая идея разреживания — обнуление малых вероятностей. Чтобы обосновать эту эвристику, найдите, какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \operatorname*{norm}_w \left(n_{wt} [n_{wt} > \gamma n_t] \right)$$

Подсказка: с учётом подстановки несмещённой оценки ϕ^*_{wt}

Теоретическое задание к лекции 5

Проект «Тематизатор». Аналитик построил модель $\Phi^0\Theta^0$ и отметил среди столбцов матрицы Φ^0 темы двух типов: удачные $T_+ \subset T$ и неудачные $T_- \subset T$.

Теперь он хочет построить модель ещё раз так, чтобы

- удачные темы остались в матрице Φ;
- остальные темы построились по-другому и были не похожи на каждую из неудачных тем $t \in T_-$.

15. Предложите регуляризаторы для этого.

16. Не получится ли так, что новые темы будут отдаляться от суммы неудачных тем $\sum_{t \in T_-} \phi^0_{wt}$ вместо того, чтобы отдаляться от каждой из неудачных тем по отдельности? Почему это плохо и как этого избежать?

17. Предложите способ инициализации Φ для новой модели.

Исследовательская задача к лекции 5

Продолжение исследования по автоматическому выделению научных терминов (Automatic Term Extraction, ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  — оптимальную стратегию регуляризации на основе декоррелирования и сглаживания фоновых тем
  — рекомендации по управлению относительными коэффициентами регуляризации
  — критерий тематичности терминов по расстоянию между распределениями $p(t|w)$ и $p_0(t) = \frac{1}{|T|}$, позволяющий наиболее чётко отличать термины от фоновой лексики
- Критерий:
  — максимум доли терминов в предметных темах
  — минимум доли терминов в фоновых темах

Исследовательская задача к лекции 5

Продолжение исследования модели локального контекста (можно воспользоваться готовой реализацией EM-алгоритма)

Исследуйте устойчивость модели в сравнении с ARTM

- без регуляризации
- с регуляризатором декоррелирования, при различных значениях относительного коэффициента регуляризации

Как на устойчивость модели влияют её параметры:

- $|T|$ — число тем
- L — число проходов
- τ — вес N_{tw} в формуле M-шага, особый случай $\tau = 0$
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, β — баланс левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, секций
- опция «$i \in C_i$ или $i \notin C_i$»

Теоретическое задание к лекции 6

18. Для иерархической тематической модели с рег. $R(\Phi, \Psi)$ предложите способ разреживания матрицы связей $\Psi = \big(p(s|t)\big)$, гарантирующий, что
1) у каждой родительской темы будет хотя бы одна дочерняя;
2) у каждой дочерней темы будет хотя бы одна родительская.

Подсказка: можно придумывать критерий регуляризации, а можно — формулу М-шага для матрицы Ψ.

19. Предложите способ гарантировать, что если родительская тема t получает только одну дочернюю s, то она переходит в неё целиком и как распределение: $p(w|s) = p(w|t)$, то есть тема t на данном уровне не расщепляется на подтемы.

20. Предложите способ согласования вероятностных смесей $p(w|t) \approx \sum_{s \in S} p(w|s)p(s|t)$ и $p(t|d) \approx \sum_{s \in S} p(t|s)p(s|d)$ с учётом тождества $p(s|t)p(t) = p(t|s)p(s)$.

Исследовательское задание к лекции 6

Проект «Мастерская знаний». Нужна тематическая модель подборок научных статей и/или поисковой выдачи.

Дано:

- 1000 подборок, в каждой по 1000 аннотаций научных статей, ранжированные по сходству с аннотацией-запросом по эмбедингам модели SciRus (эмбединги тоже даны)

Найти:

- метод согласования тематической модели с эмбедингами
- метод выделения терминов (Automatic Term Extraction)
- метод отбора терминов по тематичности
- метод отсева тематически нерелевантных аннотаций

Критерии:

- согласованность тематической модели с эмбедингами
- интерпретируемость тем
- качество выделения терминов

Теоретическое задание к лекции 7

21. Выведите EM-алгоритм с регуляризатором семантической однородности, предполагая, что n_{tdw} и n_t — константы (внешние параметры, не зависящие от Φ, Θ).
Докажите, что подстановка этого регуляризатора в M-шаг эквивалентна введению мультипликативной поправки $(1 + \tau\beta_{dw})$ в критерий log-правдоподобия.

22.** Выведите EM-алгоритм с регуляризатором семантической однородности, предполагая, что n_{tdw} и n_t выражаются через параметры модели Φ, Θ.

23*. Предложите формулу средневзешенных статистик S_* для тематической модели локальных контекстов.
Проверьте, что полученная формула совпадает с введённой на лекции, если контекстом является весь документ.

Исследовательская задача к лекции 7

Исследование EM-алгоритма для модели локального контекста

- Оценивание внутритекстовой когерентности
  - реализуйте вычисление средневзвешенной когерентности
  - подберите наилучшее сочетание эвристик rel и coh в калибровочном эксперименте без экспертной разметки
  - какие эвристики в модели локального контекста улучшают внутритекстовую когерентность?
  - воспроизводимо ли это улучшение на разных коллекциях?
- Оценивание средневзвешенных статистик
  - реализуйте вычисление S_t, S_{wt}
  - как зависит вид распределения $\{S_t\}$ от числа тем?
  - есть ли корреляция между S_t и когерентностью coh_t?
  - предложите способ разделения темы с большим S_t на подтемы и их инициализацию термами с большими S_{wt}
- Оценивание несбалансированности тем
  - реализуйте генератор коллекций с заданным дисбалансом тем
  - как дисбаланс влияет на число разделённых и слитых тем?
  - модели локального контекста лишены этой проблемы?
  - уменьшает ли регуляризатор семантической однородности число разделённых и слитых тем?

Примеры датасетов для практических заданий по курсу

1. Открытые датасеты (английский): 20NG, NIPS, KOS
2. Ранжированные результаты поиска научных статей (по данным eLibrary, arXiv, PubMed)
3. Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
4. Техноблоги: Хабр (русский), TechCrunch (английский)
5. Данные социальных сетей: VK, Twitter, Telegram,...
6. Статьи по Complexity Sciences (для хронокарты науки)

- Википедия
- Викиновости (1.5M статей, проект закрыт 30/03/2026)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

---

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

Проекты

- «Тематизатор» для социо-гуманитарных исследований:
  — пользователь задаёт грубый фильтр текстового потока;
  — задача: «классифицировать иголки в стоге сена»,
  — разделив темы на информативные и мусорные,
  — выделив аспекты и тональности в каждой теме;
  — конечная цель: кол./кач. анализ предметной области,
  — реализация данного сценария как модуля в среде Orange

- «Мастерская знаний» для научного поиска:
  — пользователь строит тематические подборки статей,
  — поисковая выдача формируется моделью SciRus;
  — задача: показать пользователю тематику подборки;
  — понадобится: автоматическое выделение терминов,
  — выделение тематических фраз из документов,
  — автоматическое именование и суммаризация тем;
  — конечная цель: помочь в понимании предметной области

Открытые проблемы тематического моделирования

1. Тематические модели внимания последовательного текста
2. Проблема несбалансированности тем в коллекции
3. Измерение интерпретируемости тем (когерентность)
4. Обеспечение 100%-й интерпретируемости тем
5. Автоматическое именование и суммаризация тем
6. Калибровка моделей тематической фильтрации
7. Согласование тем с предобученными эбмедингами LLM
8. Статистические оценки состоятельности тем
9. Обнаружение новых тем или трендов в потоке текстов
10. Обеспечение устойчивости и полноты множества тем
11. Автоматический подбор гиперпараметров, AutoML
12. Гиперграфовые тематические модели для RecSys