Тематический информационный поиск в цифровых гуманитарных исследованиях

Воронцов Константин Вячеславович
д.ф.-м.н., профессор РАН
Институт ИИ МГУ • ВМК МГУ • МФТИ • ФИЦ ИУ РАН

Искусственный интеллект в исторических исследованиях
Научный семинар РАНХиГС • 11 февраля 2023

Содержание

Задача тематического моделирования

Дано:

- коллекция текстовых документов

Найти:

- T — множество тем, составляющих эту коллекцию
- $\phi_{wt} = p(w|t)$ — вероятности слов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

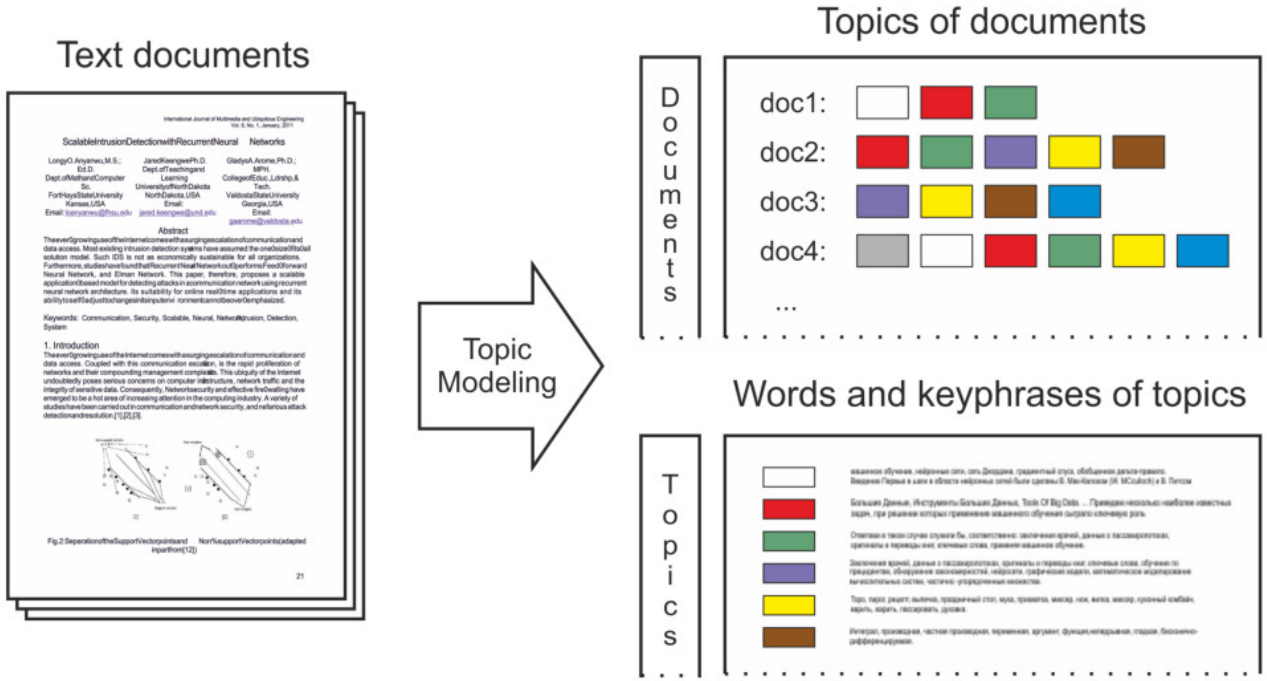
Критерий:

- вероятностная тематическая модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$ должна предсказывать появление слов w в документах d,
- заодно максимизируя сумму регуляризаторов $R_i(\Phi, \Theta)$:

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

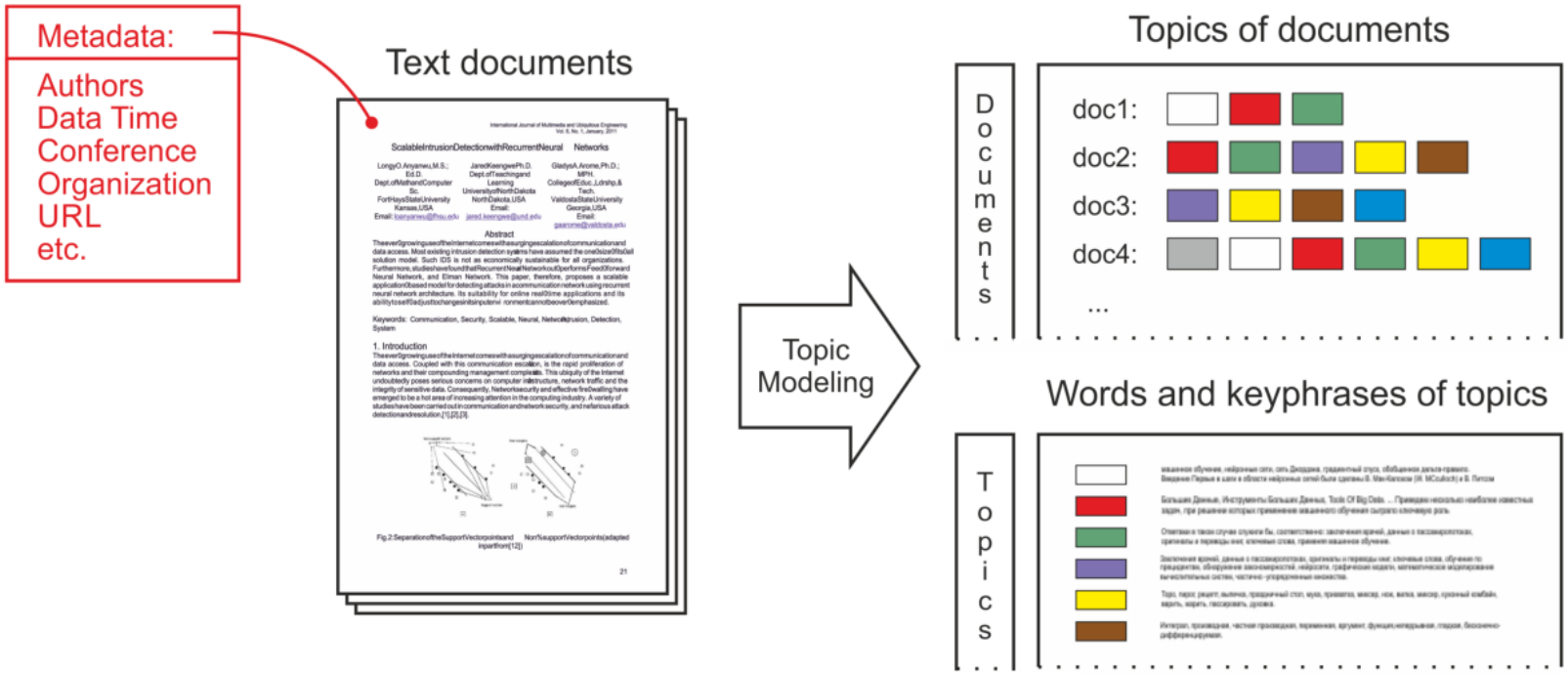
Мультимодальные тематические модели

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



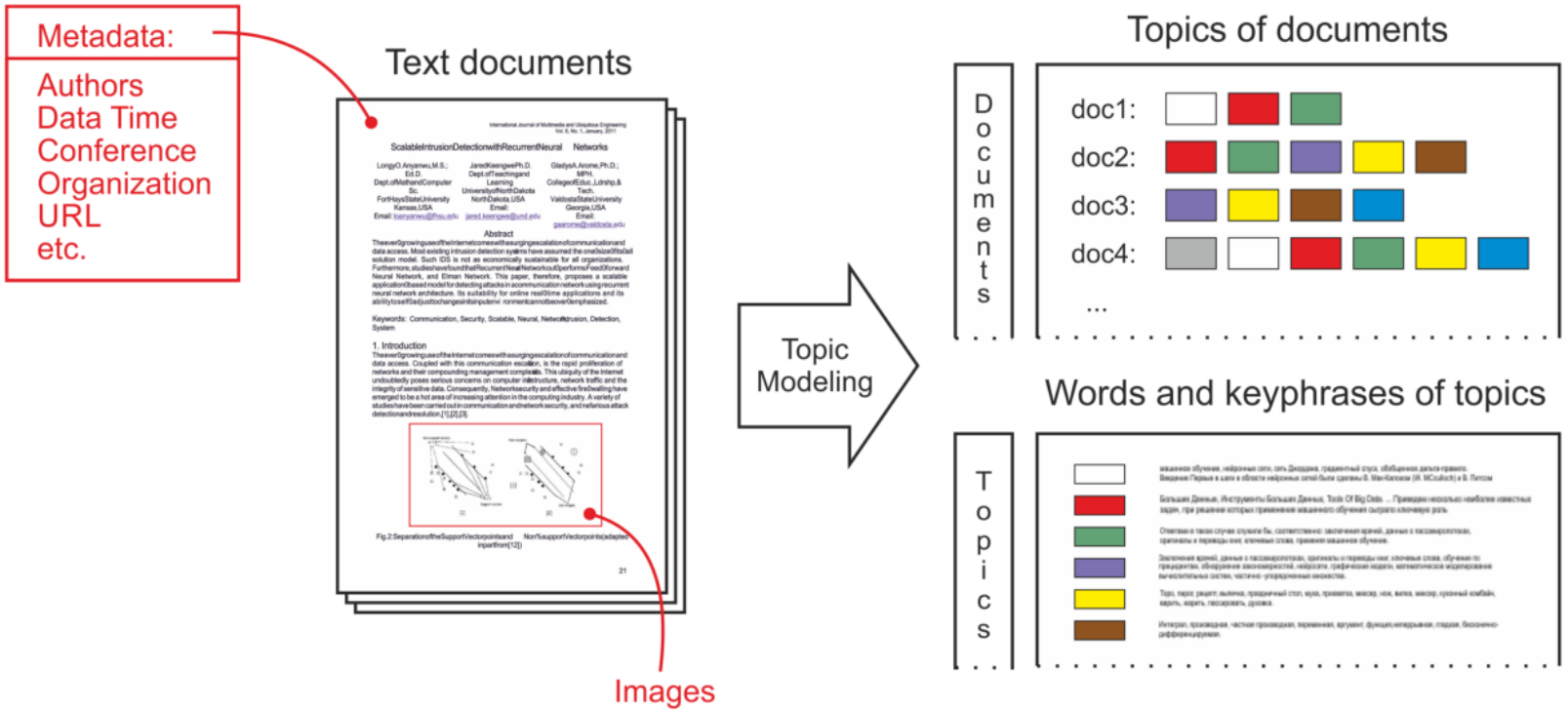
Мультимодальные тематические модели

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



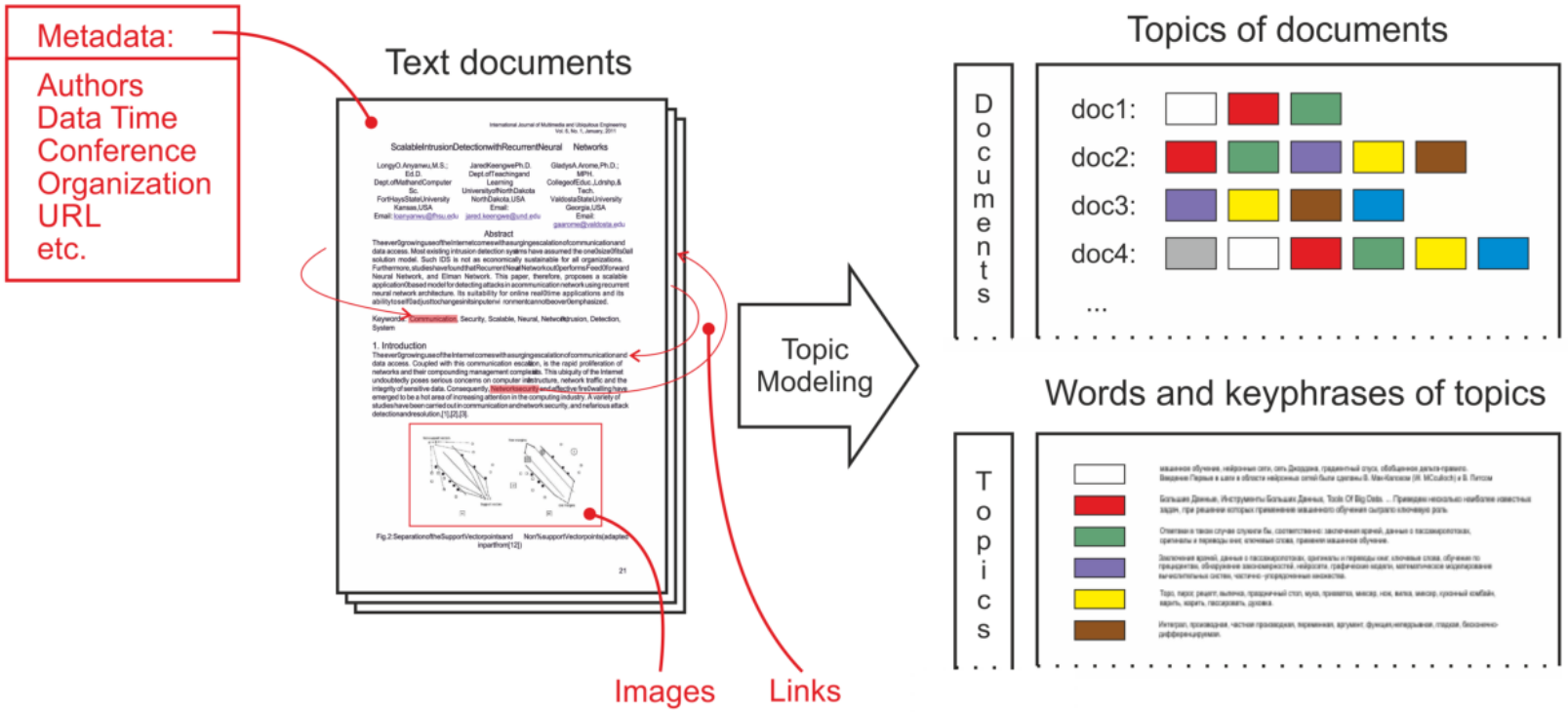
Мультимодальные тематические модели

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



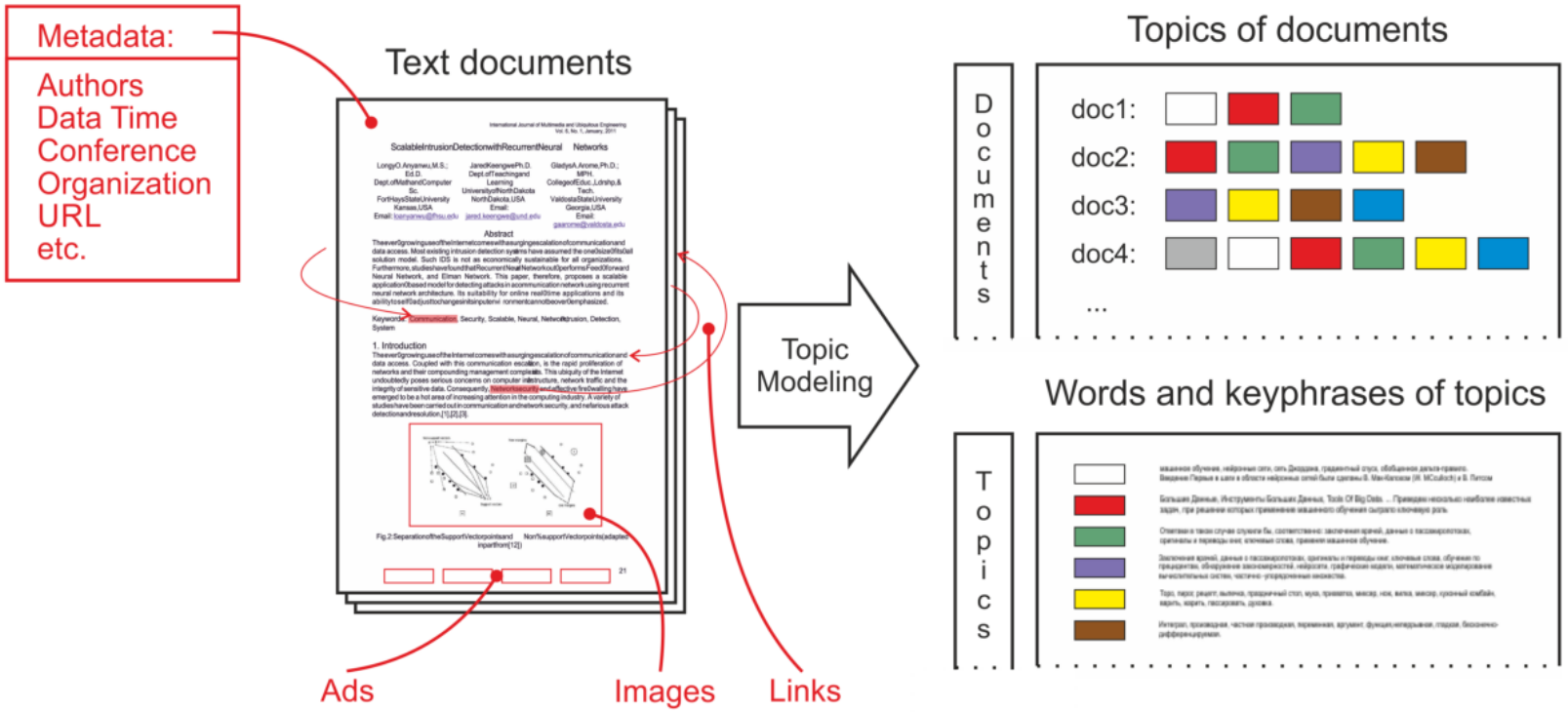
Мультимодальные тематические модели

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



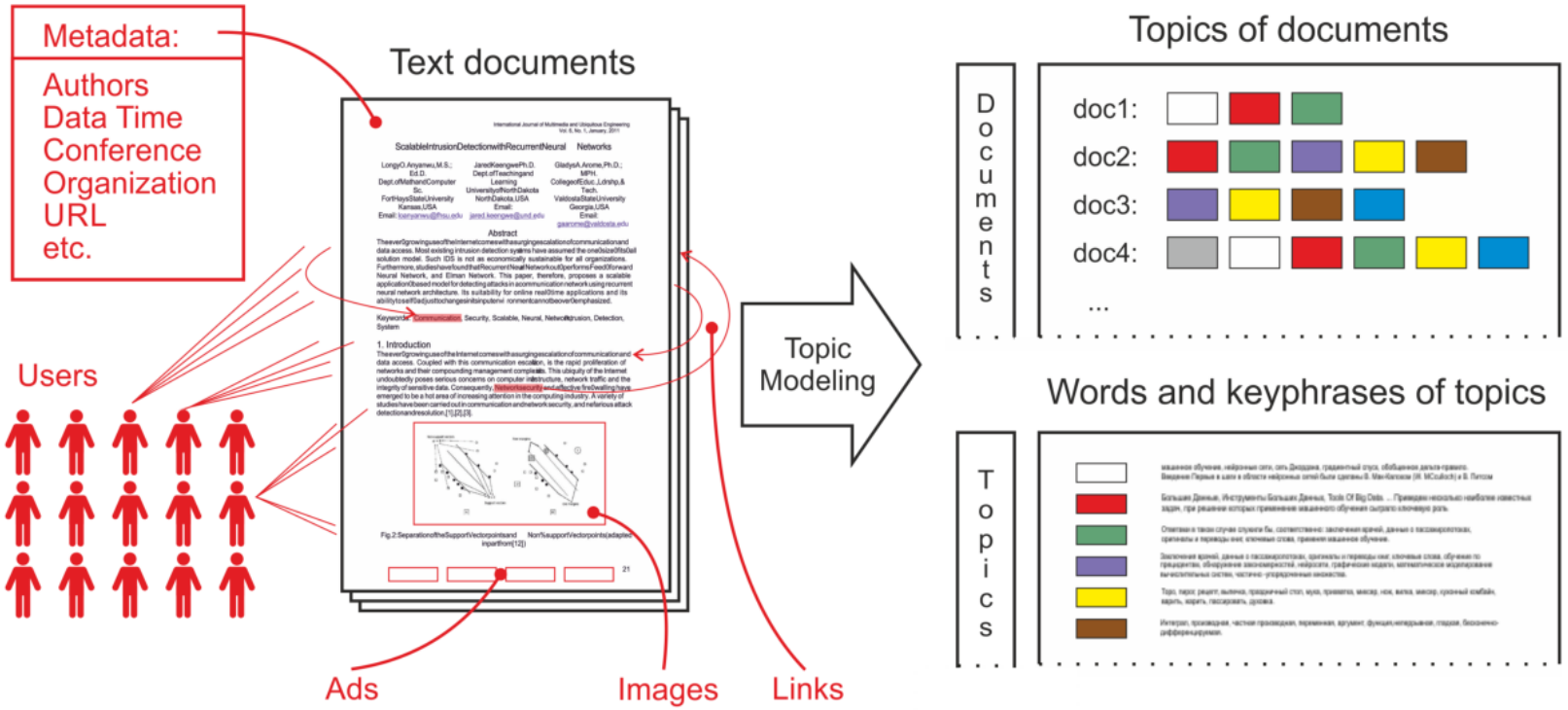
Мультимодальные тематические модели

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



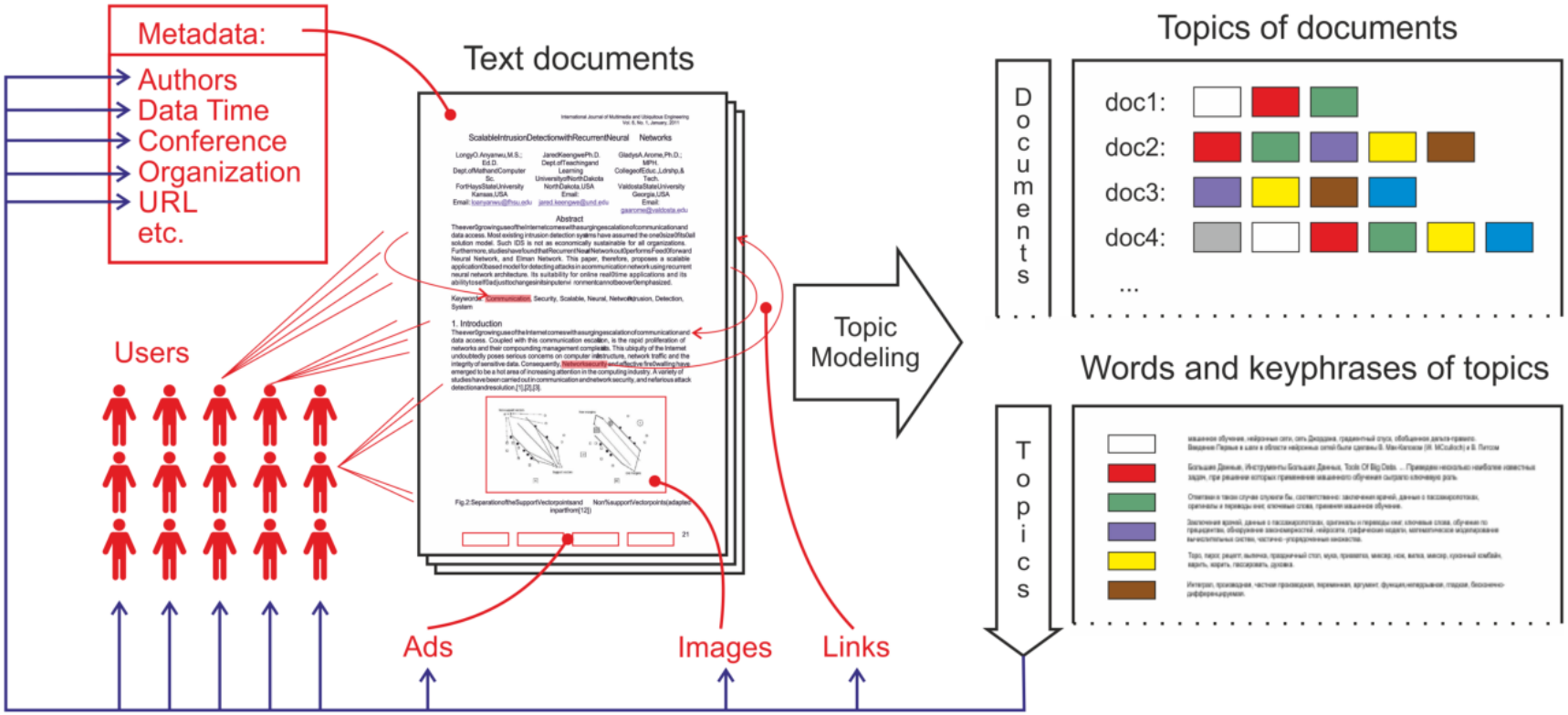
Мультимодальные тематические модели

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



Мультимодальные тематические модели

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



Пример 1. Мультиязычная модель Википедии

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

Тема №68				Тема №79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример 1. Мультиязычная модель Википедии

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

Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример 2. Биграммная модель научных конференций

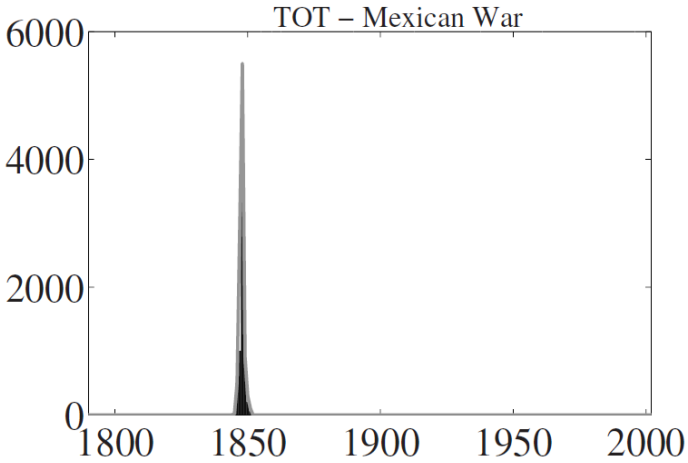
Коллекция 1000 статей конференций ММРО, ИОИ на русском

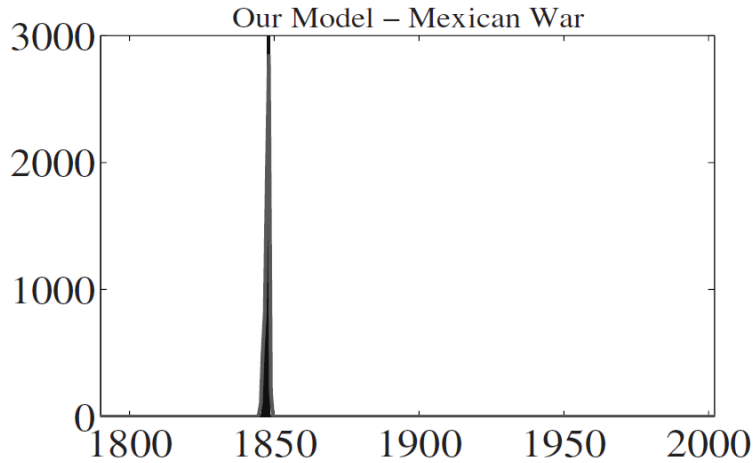
распознавание образов в биоинформатике		теория вычислительной сложности	
униграммы	биграммы	униграммы	биграммы
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

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

Пример 3. Совмещение темпоральной и n-граммной модели

Коллекция еженедельных выступлений президентов США





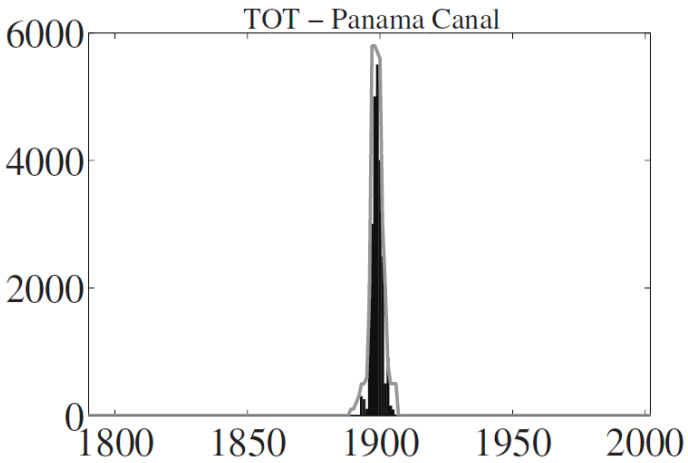
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

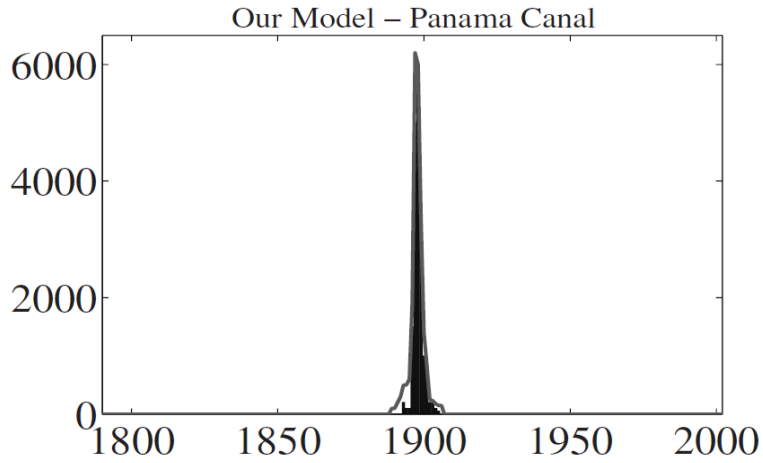
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An N-gram topic model for time-stamped documents. 2013.

Пример 3. Совмещение темпоральной и n-граммной модели

Коллекция еженедельных выступлений президентов США





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

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

Shoaib Jameel, Wai Lam. An N-gram topic model for time-stamped documents. 2013.

Цели и не-цели тематического моделирования

Цели:

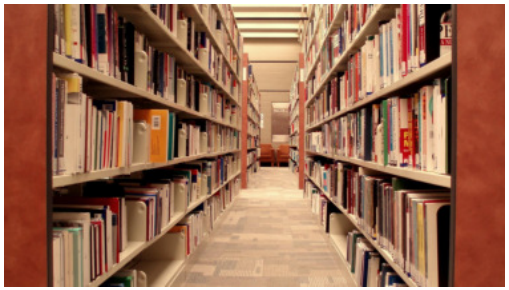
- Выяснить тематическую кластерную структуру текстовой коллекции, сколько в ней тем и какие они
- Получать интерпретируемые тематические векторные представления (эмбединги) документов $p(t|d)$, слов $p(t|w)$, фрагментов и прочих объектов $p(t|x)$
- Решать задачи поиска, категоризации, сегментации, суммаризации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

Не-цели:

- Угадывать следующие слова (ТМ — слабые модели языка)
- Генерировать связный текст
- Понимать смысл текста

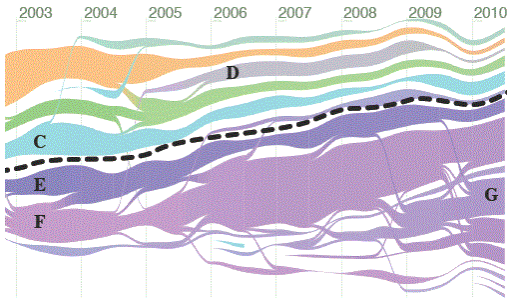
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

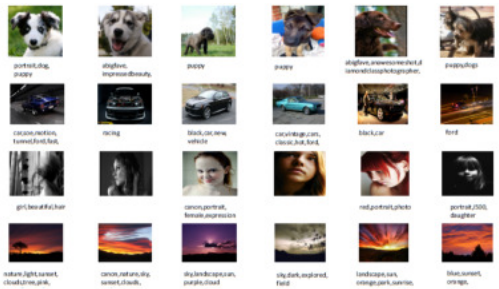


поиск тематического контента в соцсетях

выявление и отслеживание цепочек новостей

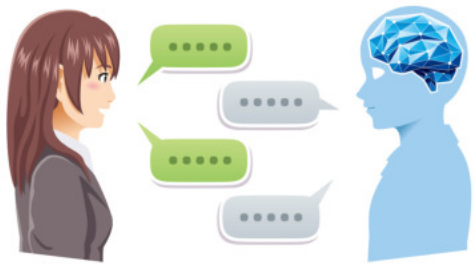


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

управлением диалогом в разговорном интеллекте



J.Boyd-Graber, Yuening Hu, D.Mimno. Applications of Topic Models. 2017.
H.Jelodar et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. 2019.

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

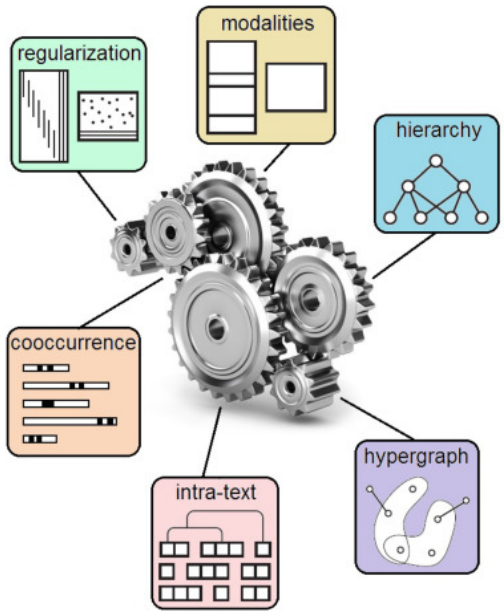
Лицензия и среда разработки:

- 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

Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

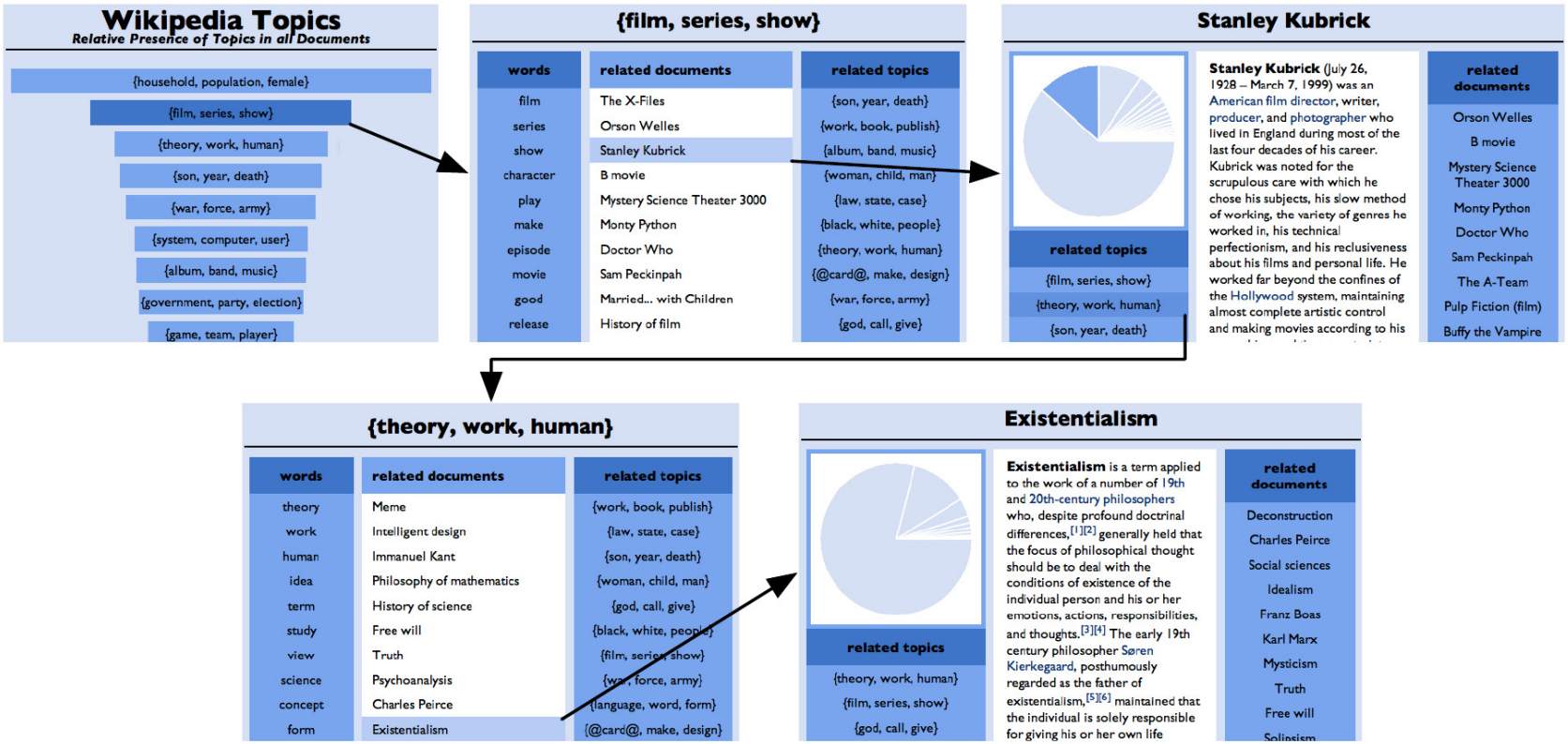
3.7M статей Википедии, 100K слов: время min (перплексия)

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

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

Система TMVE — Topic Model Visualization Engine

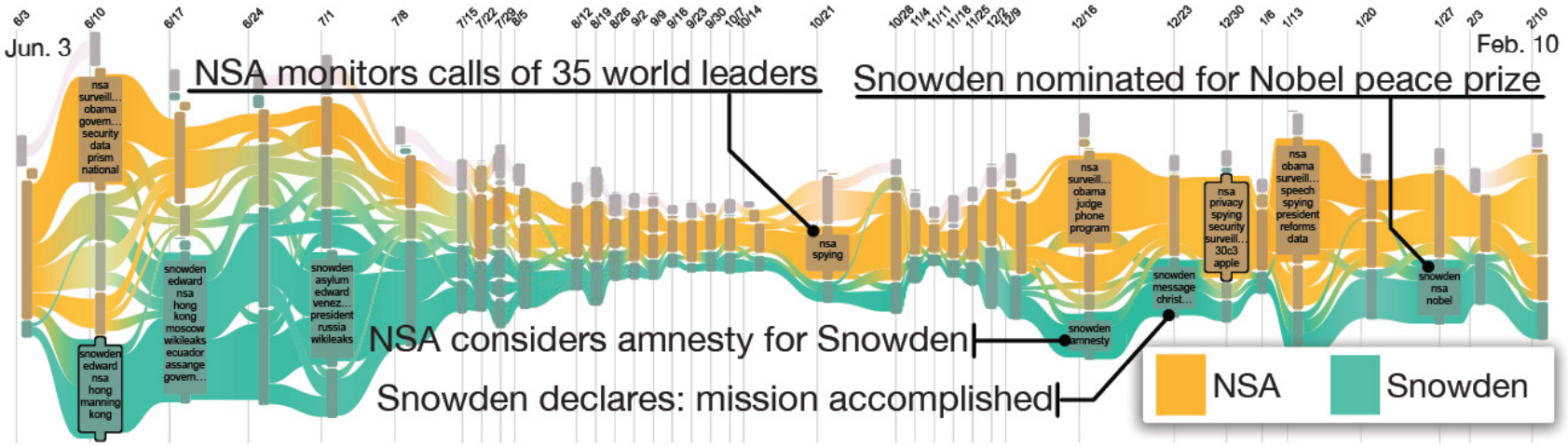
Тематический навигатор с веб-интерфейсом:



https://github.com/ajbc/tmv

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

Динамика тем: эволюция предметной области



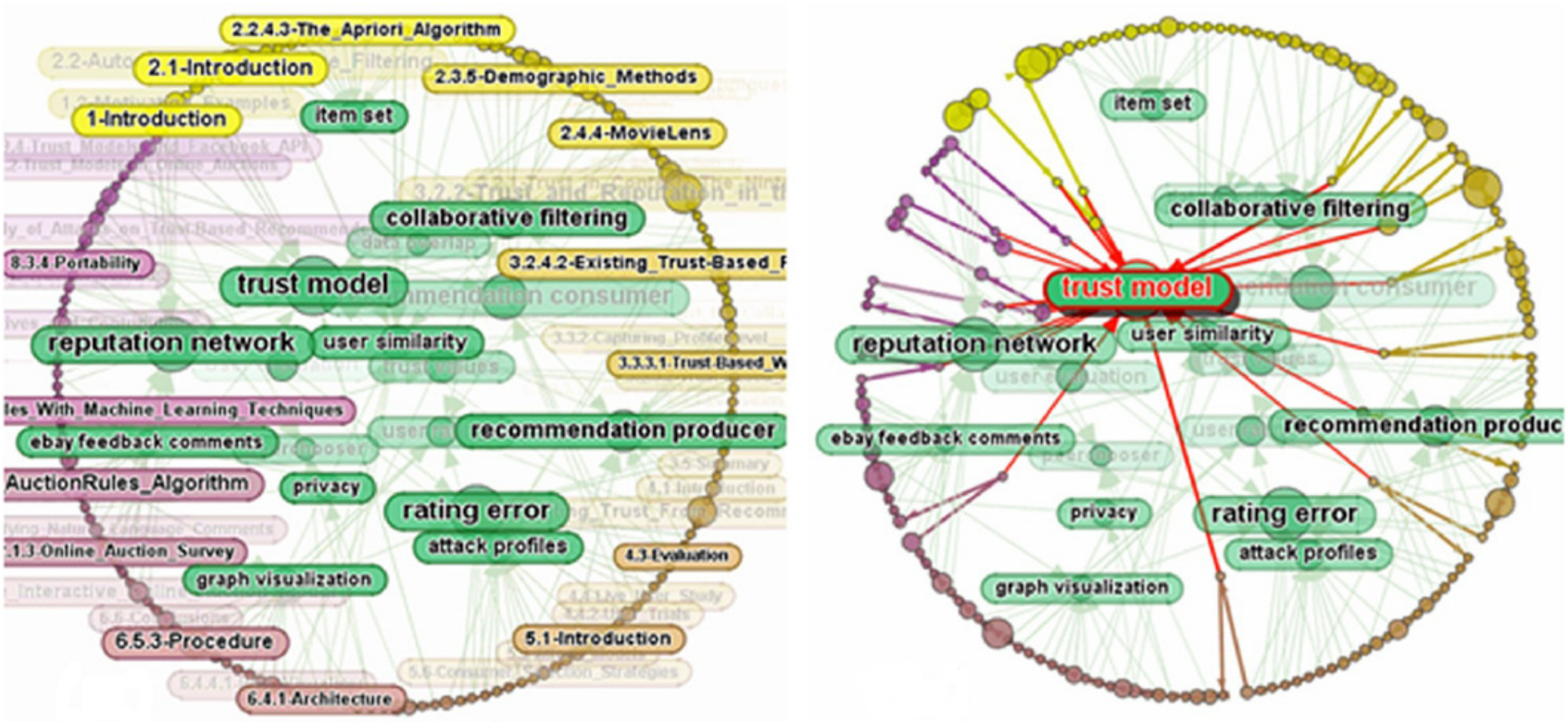
Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

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

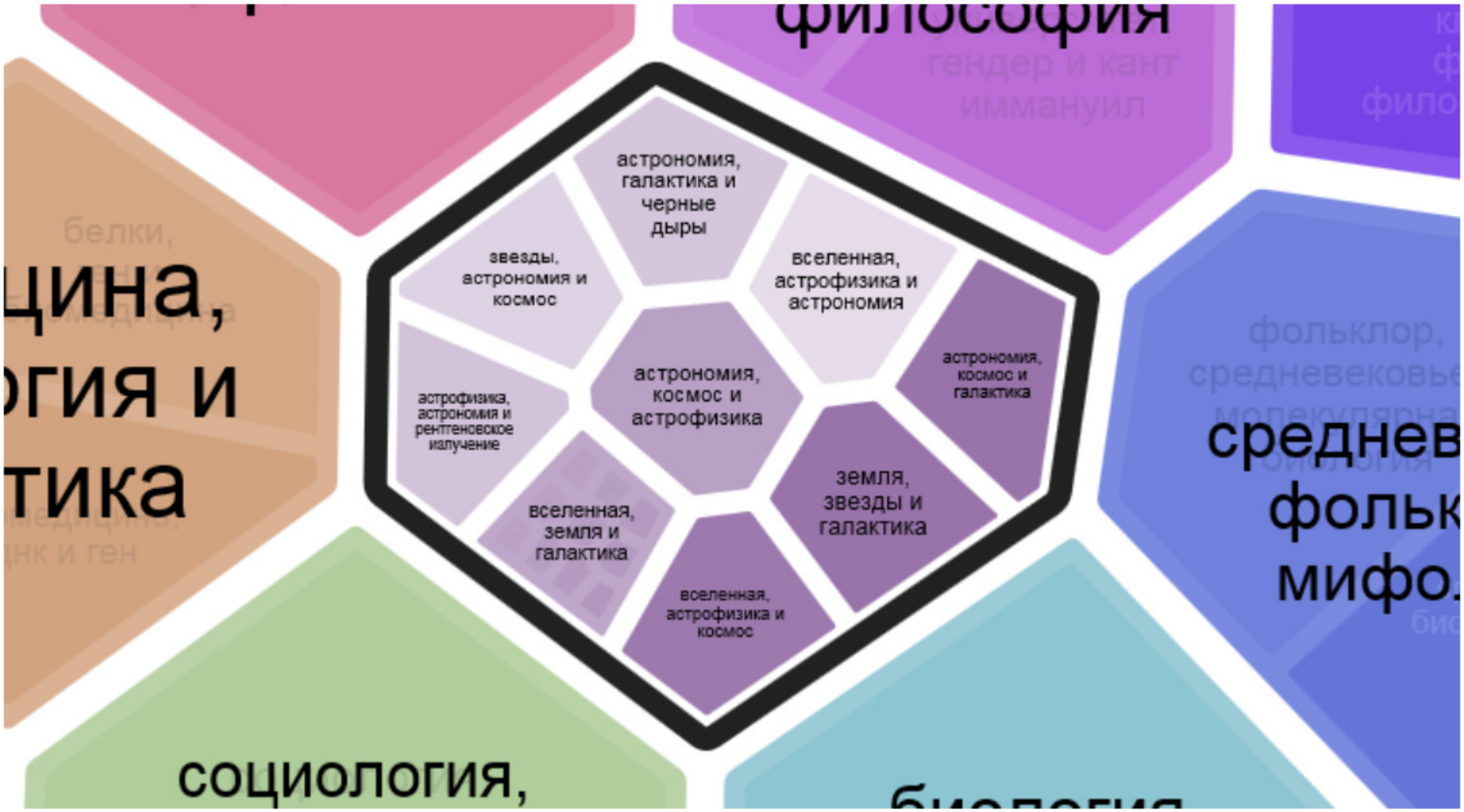


Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P. TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology. 2012.

Визуализация иерархической тематической модели

Chirkova N. A., Vorontsov K. V. Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.

Визуализация иерархической тематической модели

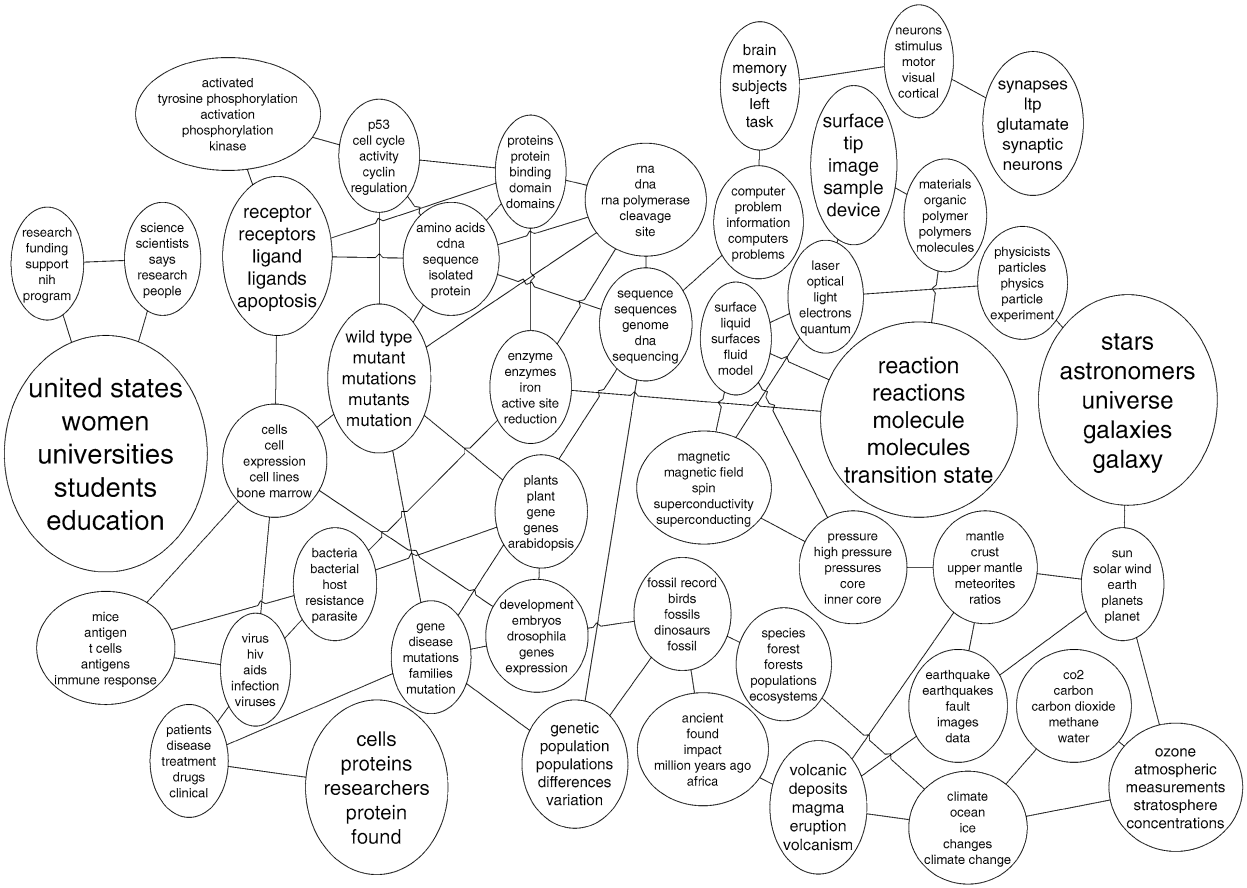


Belyy A.V., *Seleznova M.S.*, *Sholokhov A.K.*, *Vorontsov K.V.* Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue 2018.

Модель коррелированных тем CTM

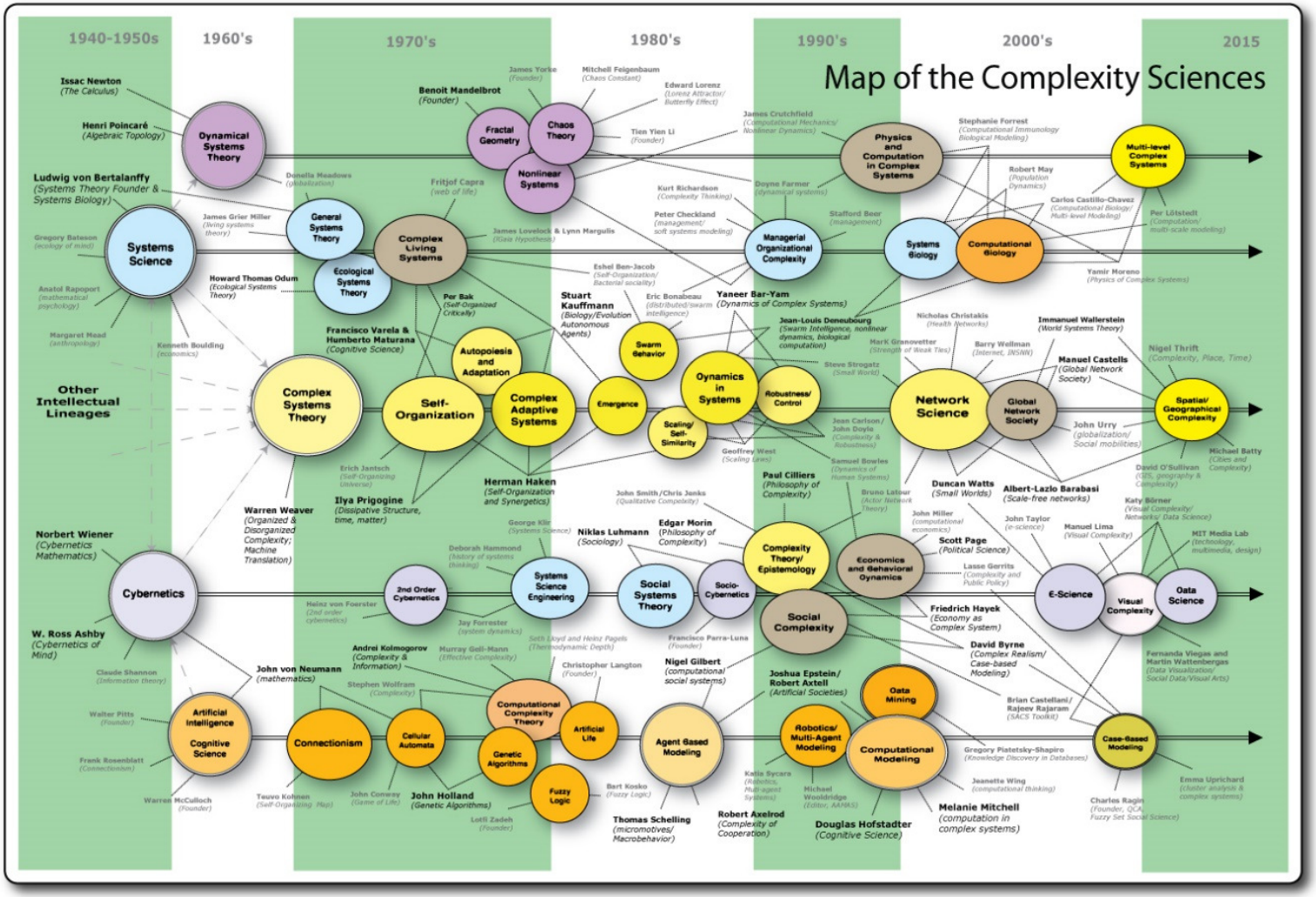
Мотивация: выявлять междисциплинарные связи.

Статьи по археологии чаще связаны с историей и геологией, чем с генетикой



David Blei, John Lafferty. A Correlated Topic Model of SCIENCE, 2007.

Пример карты предметной области, построенной вручную



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

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

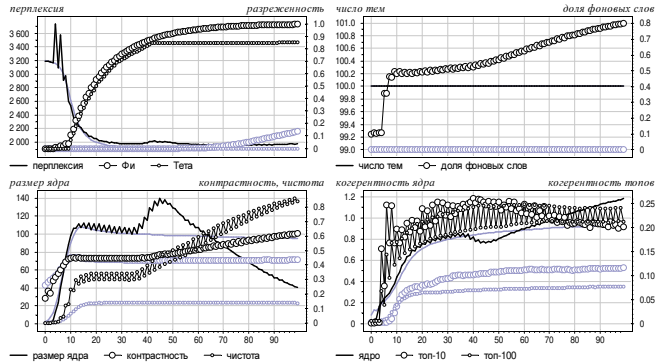
Интерактивный обзор 440 средств визуализации текстов

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.

Декоррелирование, сглаживание, разреживание

Цель: найти комбинацию регуляризаторов, улучшающую интерпретируемость тем по совокупности критериев.



Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{decorrelated}\right) + R\left(\text{sparse}\right) + R\left(\text{background}\right) \to \max$$

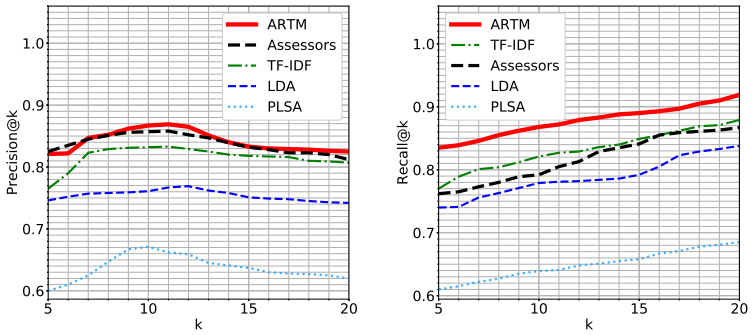
Результаты:

- разреженность $0 \to 95\%$, когерентность $0.25 \to 0.96$, чистота $0.14 \to 0.89$, контрастность $0.43 \to 0.52$,
- без заметного ущерба для перплексии: $1920 \to 2020$
- выработаны рекомендации по стратегии регуляризации

K.Vorontsov, A.Potapenko. Additive regularization of topic models. Mach. Learn. 2015.

Разведочный поиск в технологических блогах — 1

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



Регуляризаторы:

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

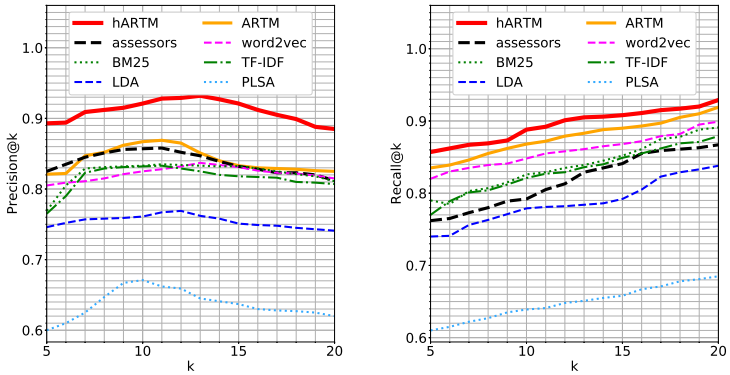
Результаты:

- Точность и полнота 88%, превосходит асессоров и другие методы (tf-idf, word2vec, PLSA, LDA).
- Векторный поиск мгновенный, асессоры тратили 5–65 мин.

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

Разведочный поиск в технологических блогах — 2

Цель: улучшение качества поиска с помощью иерархической тематической модели hARTM и отсечения нерелевантных тем.



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.

Две коллекции новостей про технологии

Habrahabr.ru

175 143 статей на русском
10 552 слов (униграмм)
742 000 биграмм
524 авторов статей
10 000 авторов комментариев
2546 тегов
123 хаба (категории)

TechCrunch.com

759 324 статей на английском
11 523 слов (униграмм)
1.2 млн. биграмм
605 авторов
184 категорий

Предобработка текстов

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

Методика оценивания качества разведочного поиска

Поисковый запрос

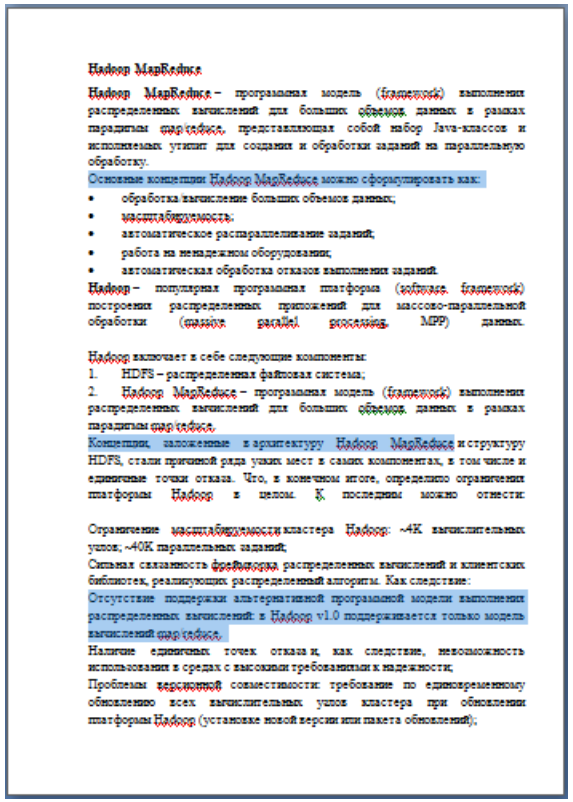
набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

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

Два задания асессорам

1. найти как можно больше статей, пользуясь любыми средствами поиска (и засечь время)
2. оценить релевантность поисковой выдачи на том же запросе



Hadoop MapReduce

Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce, представляющая собой набор Java-классов и исполняемых утилит для создания и обработки заданий на параллельную обработку.

Основные концепции Hadoop MapReduce можно сформулировать как:

- обработка/вычисление больших объемов данных;
- масштабируемость;
- автоматическое распараллеливание заданий;
- работа на ненадежном оборудовании;
- автоматическая обработка отказов выполнения заданий.

Hadoop – популярная программная платформа (software framework) построения распределенных приложений для массово-параллельной обработки (massive parallel processing, MPP) данных.

Hadoop включает в себе следующие компоненты:

1. HDFS – распределенная файловая система;
2. Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce.

Концепции, заложенные в архитектуру Hadoop MapReduce и структуру HDFS, стали причиной ряда узких мест в самих компонентах, в том числе и единичные точки отказа. Что, в конечном итоге, определило ограничения платформы Hadoop в целом. К последним можно отнести:

Ограничение масштабируемости кластера Hadoop: ~4K вычислительных узлов; ~40K параллельных заданий;

Сильная связанность фреймворка распределенных вычислений и клиентских библиотек, реализующих распределенный алгоритм. Как следствие:

Отсутствие поддержки альтернативной программной модели выполнения распределенных вычислений: в Hadoop v1.0 поддерживается только модель вычислений map/reduce.

Наличие единичных точек отказа и, как следствие, невозможность использования в средах с высокими требованиями к надежности;

Проблемы версионной совместимости: требование по единовременному обновлению всех вычислительных узлов кластера при обновлении платформы Hadoop (установке новой версии или пакета обновлений);

Пример запроса для разведочного поиска

Пример: фрагмент запроса «Система IBM Watson»

IBM Watson — суперкомпьютер фирмы IBM, оснащённый вопросно-ответной системой искусственного интеллекта, созданный группой исследователей под руководством Дэвида Феруччи. Его создание — часть проекта DeepQA. Основная задача Уотсона — понимать вопросы, сформулированные на естественном языке, и находить на них ответы в базе данных. Назван в честь основателя IBM Томаса Уотсона.

IBM Watson представляет собой когнитивную систему, которая способна понимать, делать выводы и обучаться. Она также позволяет преобразовывать целые отрасли, различные направления науки и техники. Например, предсказывать появление эпидемий или возникновения очагов природных катастроф в различных регионах, вести мониторинг состояния атмосферы больших городов, оптимизировать бизнес-процессы, узнавать, какие товары будут в тренде в ближайшее время.

... ...

Релевантные тексты: примеры сервисов и приложений, основа которых — когнитивная платформа IBM Watson, используемые в IBM Watson технологии, вопрос-ответные системы, сопоставление IBM Watson с Wolfram-Alpha.

Нерелевантные тексты: общие вопросы искусственного интеллекта, другие коммерческие решения на рынке бизнес-аналитики.

Тематика запросов разведочного поиска

Примеры заголовков разведочных запросов к Хабру (объём каждого запроса — около одной страницы A4):

Алгоритмы раскраски графов	Система IBM Watson
Рекомендательная система Netflix	3D-принтеры
Методики быстрого набора текста	CERN-кластер
Космические проекты Илона Маска	AB-тестирование
Технологии Hadoop MapReduce	Облачные сервисы
Беспилотный автомобиль Google car	Контекстная реклама
Криптосистемы с открытым ключом	Марсоход Curiosity
Обзор платформ онлайн-курсов	Видеокарты NVIDIA
Data Science Meetups в Москве	Распознавание образов
Образовательные проекты mail.ru	Сервисы Google scholar
Межпланетная станция New horizons	MIT MediaLab Research
Языковая модель word2vec	Платформа Microsoft Azure

Поиск и рубрикация научных публикаций на 100 языках

Цель: мультиязыковой 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.

П.Потапова, А.Грабовой, О.Бахтеев, Е.Егоров, Н.Зиновкин, Ю.Чехович, К.Воронцов и др. Мультиязыковая автоматическая рубрикация научных документов. 2023.

Поиск и классификация этно-релевантных тем в соцсетях

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

(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, сообщать, океан, станция, хатико, район, правительство, атомный,
(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,
(венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадуро, 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.

Поиск этно-релевантных тем в социальных сетях

- **Дано:**
  1) данные социальных медиа (ВК и др.)
  2) словарь этнонимов (около 300)
- **Найти:**
  1) как можно больше тем про этничности
  2) темы с сочетанием этничностей (возможные 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-релевантных 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.

Аналогичные по структуре исследования

Метафора: «ищем и классифицируем иголки 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.

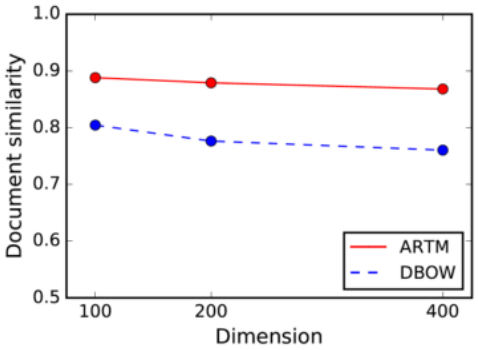
3. *M.A.Basher, A.Rahman, B.C.M.Fung.* Analyzing topics and authors in chat logs for crime investigation. 2014.

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.

Тематические модели коротких текстов

Цель: интерпретируемые разреженные тематические эмбединги на основе дистрибутивной семантики, аналоги word2vec и WNTM.



Регуляризаторы:

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

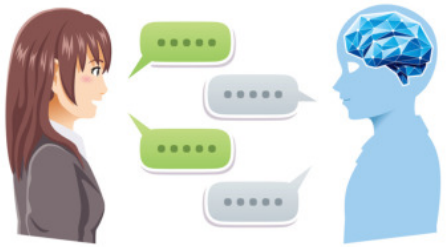
Результаты:

- Точность поиска схожих документов: 0.8 → 0.9
- Когерентность тем: 0.08 → 0.33
- Семантическая близость слов: 0.53 → 0.58, 0.38 → 0.61

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.

Выявление намерений клиентов для построения чат-ботов

Цель: выявить тематику и интенты (намерения) клиентов по коллекции обращений в контактный центр. Построить рубрикатор интентов для последующей разметки диалогов.



Регуляризаторы:

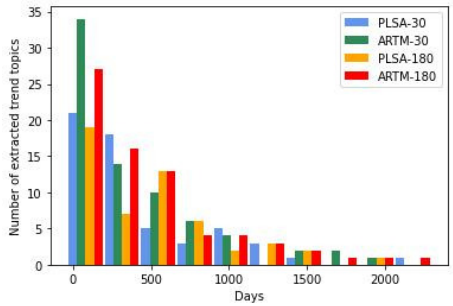
$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{interpretable}\right) + R\left(\text{hierarchy}\right) + R\left(\text{segmentation}\right)$$
$$+ R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) + R\left(\text{syntax}\right) \to \max$$

Результаты: точность классификации интентов 60% → 66%.

A.Popov, V.Bulatov, D.Polyudova, E.Veselova. Unsupervised dialogue intent detection via hierarchical topic model. RANLP, 2019.

Выявление трендов в коллекции научных публикаций

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



Регуляризаторы:

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

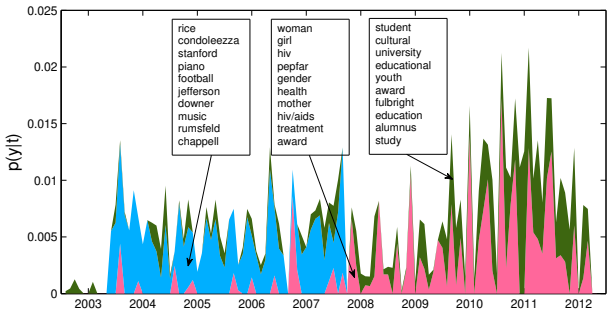
Результаты:

- 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:

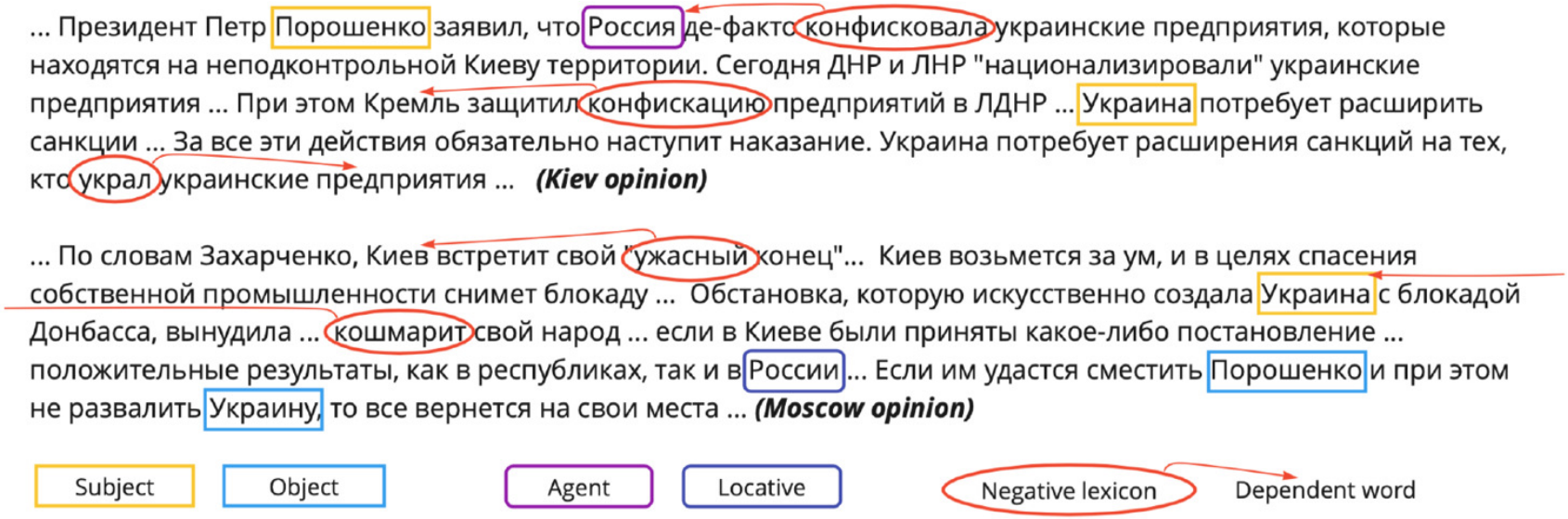
$$\mathscr{L}\left(\underset{\Phi\ \Theta}{\text{PLSA}}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) + R\left(\text{syntax}\right) \to \max$$

Результаты:

- выделение мнений внутри тем: F1-мера = 0.86%
- совместное использование трёх модальностей: факты «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.

Выделение поляризованных мнений в политических новостях



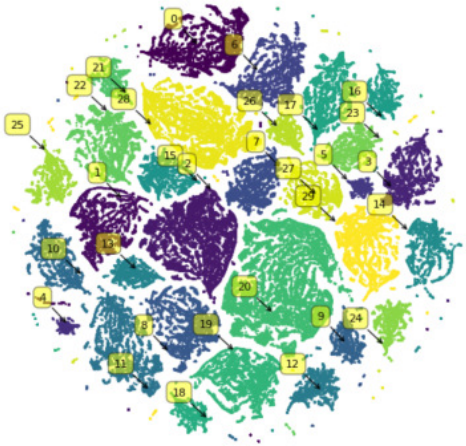
Слова «Порошенко», «Россия», «Украина» встречаются в тексте-1 и тексте-2 одинаково часто, однако:

- «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.

Тематическая модель банковских транзакционных данных

Цель: Выявление паттернов потребительского поведения клиентов банка, причём

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



Регуляризаторы:

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

Результаты:

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

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

Исследования газетных архивов

[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.

Исследования документальной литературы и дневников

[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.

Исследования научной и литературно-художественной периодики

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.

Резюме

- Тематическое моделирование — инструмент для поиска и систематизации больших текстовых коллекций
- Теория ARTM и библиотека BigARTM позволяют строить модели с нужным набором свойств
- Есть наработанные приёмы для
  — улучшения интерпретируемости тем
  — улучшения качества тематического поиска
  — исследования динамики тем во времени
  — выделения тем по большому списку слов-затравок
  — иерархического дробления тем на более мелкие подтемы
  — навигации по темам и их визуального анализа
  — учёта обратной связи с экспертом
- Эти приёмы активно используются для обработки больших массивов исторической информации
- Фактор успеха — качественная предобработка: сканирование, опечатки, токенизация, секционирование