Тематический анализ текстов

Константин Вячеславович Воронцов

ФИЦ ИУ РАН • МФТИ • ВШЭ • МГУ • Яндекс • Форексис • Айтея

Москва • НИУ ВШЭ • 26 июля 2017

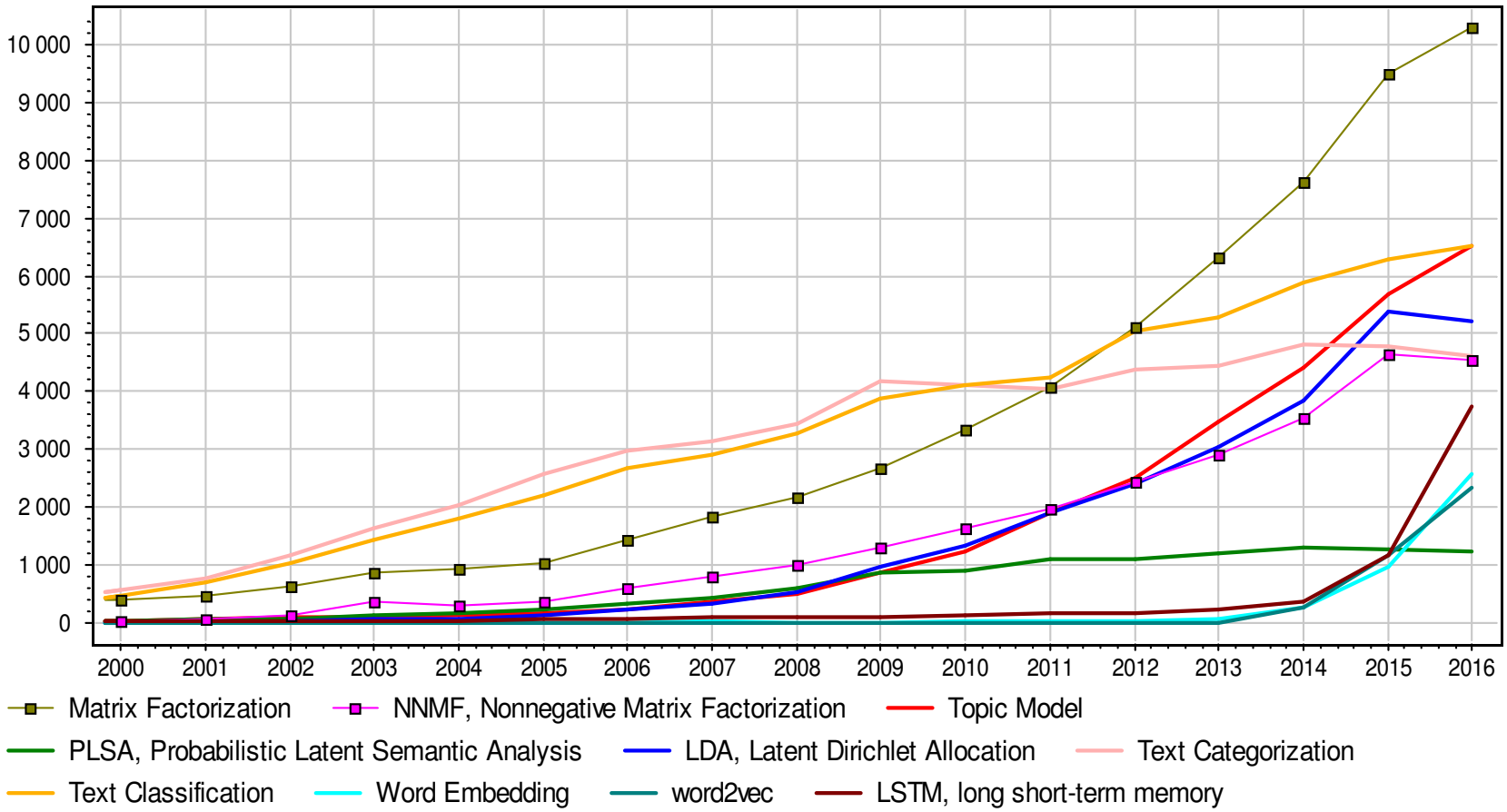
Содержание

Что такое «тематическое моделирование» (Topic Modeling)

- Одно из направлений обработки естественного языка
- Разновидность статистического анализа текстов
- Технология поиска информации не по словам, а по смыслу
- Выявление скрытых интересов по наблюдаемым данным
- «Мягкая кластеризация» текстовых документов
- Би-кластеризация слов и документов по кластерам-темам
- Модель машинного обучения без учителя
  (но есть и тематические модели, обучаемые с учителем)
- Модель языка, основанная на гипотезе «мешка слов»
  (но есть и модели, преодолевающие это ограничение)
- Сотни моделей, тысячи публикаций, тысячи приложений

Тематическое моделирование и смежные области исследований

Динамика цитирования, по данным Google Scholar:



Что такое «тема» в коллекции текстовых документов?

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

Более формально,

- *тема* — условное распределение на множестве терминов, $p(w|t)$ — вероятность (частота) термина w в теме t;
- *тематика* документа — условное распределение $p(t|d)$ — вероятность (частота) темы t в документе d.

Когда автор писал термин w в документе d, он думал о теме t, и мы хотели бы выявить, о какой именно.

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

Пример 1. Мультиязычная модель Википедии

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

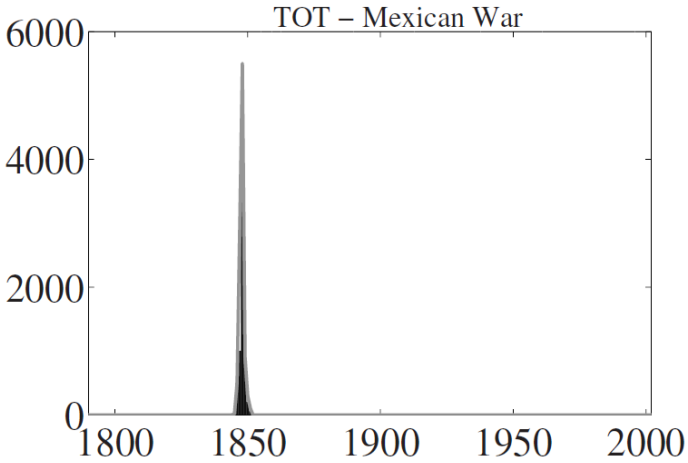
Коллекция 1000 статей конференций ММРО, ИОИ на русском

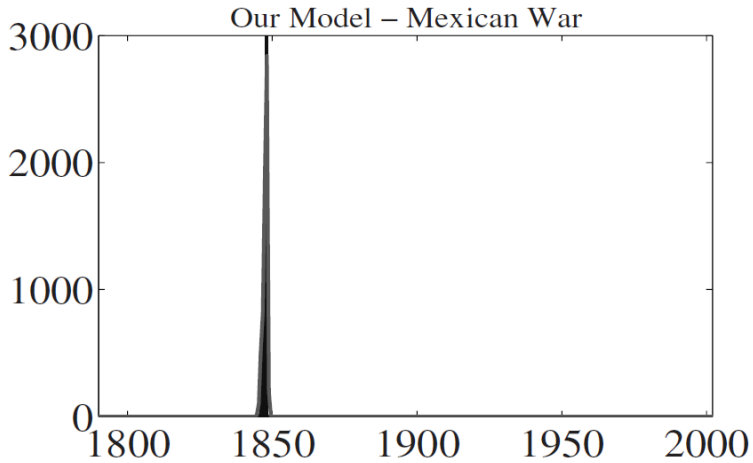
распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	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 динамической и n-граммной модели

Эксперименты на коллекции выступлений президентов США





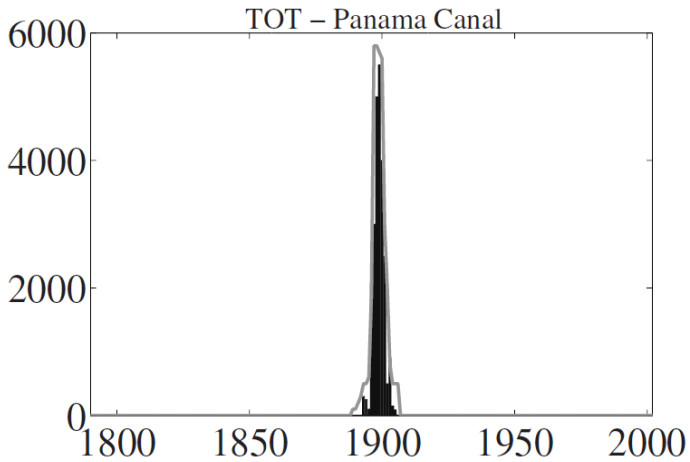
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

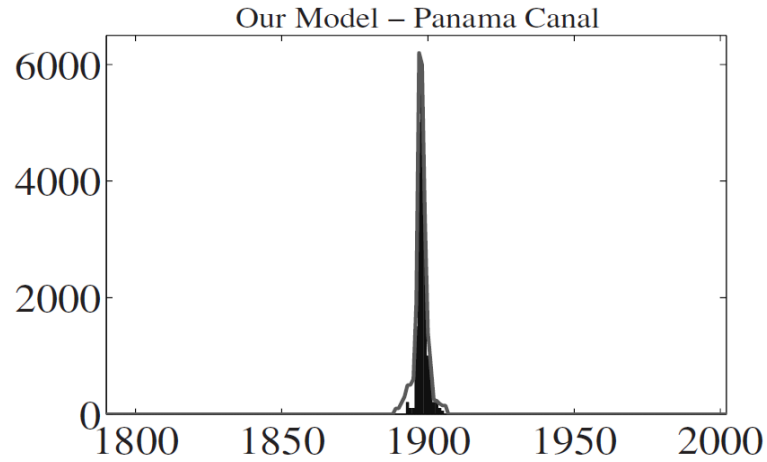
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

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

Пример 3. Совмещение динамической и n-граммной модели

Эксперименты на коллекции выступлений президентов США





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

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

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

Пример 4. Поиск этно-релевантных тем в социальных сетях

Основные задачи проекта:

- Разведочный поиск этничных тем в социальных медиа
- Мониторинг этих тем во времени и по регионам
- Выявление очагов враждебности, конфликности
- Поддержка социологических исследований

В чём сложность задачи:

- «Классифицировать иголки в стоге сена»:
  подробно тематизировать малую долю (менее 1%) контента
- Найти как можно больше этно-релевантных тем
- Выявить событийные и региональные темы
- Выявить тематические сообщества

Пример 4. Поиск этно-релевантных тем в социальных сетях

Фрагмент словаря этнонимов (слова-затравки, seed words)

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

Пример 4. Этно-релевантные темы в социальных сетях

(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,

Пример 4. Этно-релевантные темы в социальных сетях

(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,

Пример 4. Этно-релевантные темы в социальных сетях

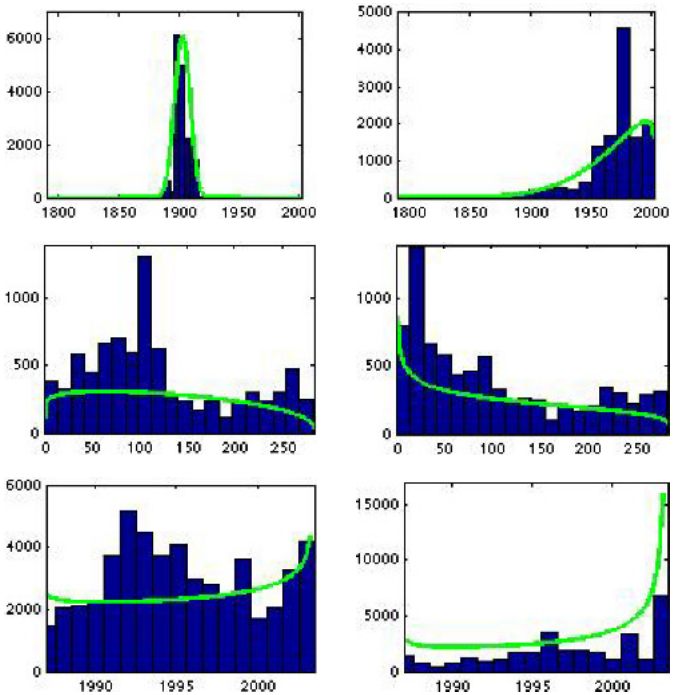
(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, сообщать, океан, станция, хатико, район, правительство, атомный,
(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, 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,

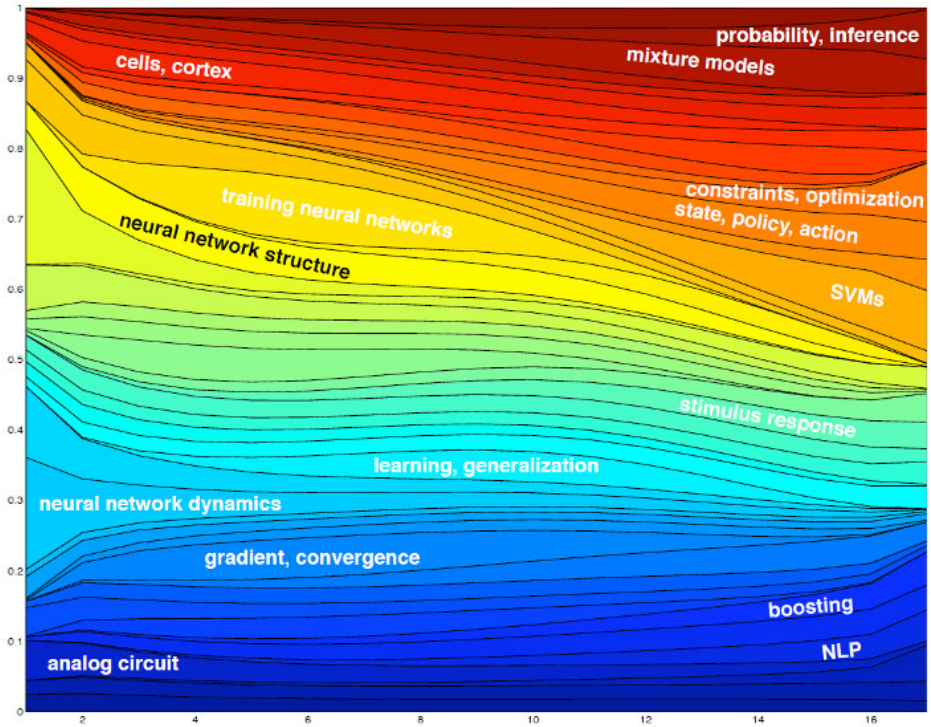
Разновидности тематических моделей (основные вехи)

- PLSA (1999) вероятностный латентный семантический анализ
- LDA (2003) латентное размещение Дирихле
- ATM (2004) авторы документов
- TOT (2006) метки времени документов
- HDP (2006) определение числа тем
- TNG (2007) группирование слов в мультиграммы
- CTM (2007) корреляции между темами
- NetPLSA (2008) граф связей между документами
- ML-LDA (2009) многоязычные параллельные тексты
- ssLDA (2012) частичное обучение
- Dependency-LDA (2012) классификация документов
- BitermTM (2013) битермы в коротких документах
- mLDA (2013) метаданные с тремя и более модальностями
- WNTM (2014) локальные контексты слов

Темпоральная модель TOT (Topics over Time)

1. Каждая тема имеет непрерывное β-распределение во времени
2. Каждое слово имеет метку времени





Xuerui Wang, *Andrew McCallum*. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends // ACM SIGKDD-2006

Темпоральные тематические модели

Неадекватность TOT очевидна даже по картинкам из статьи!

Наши предположения:

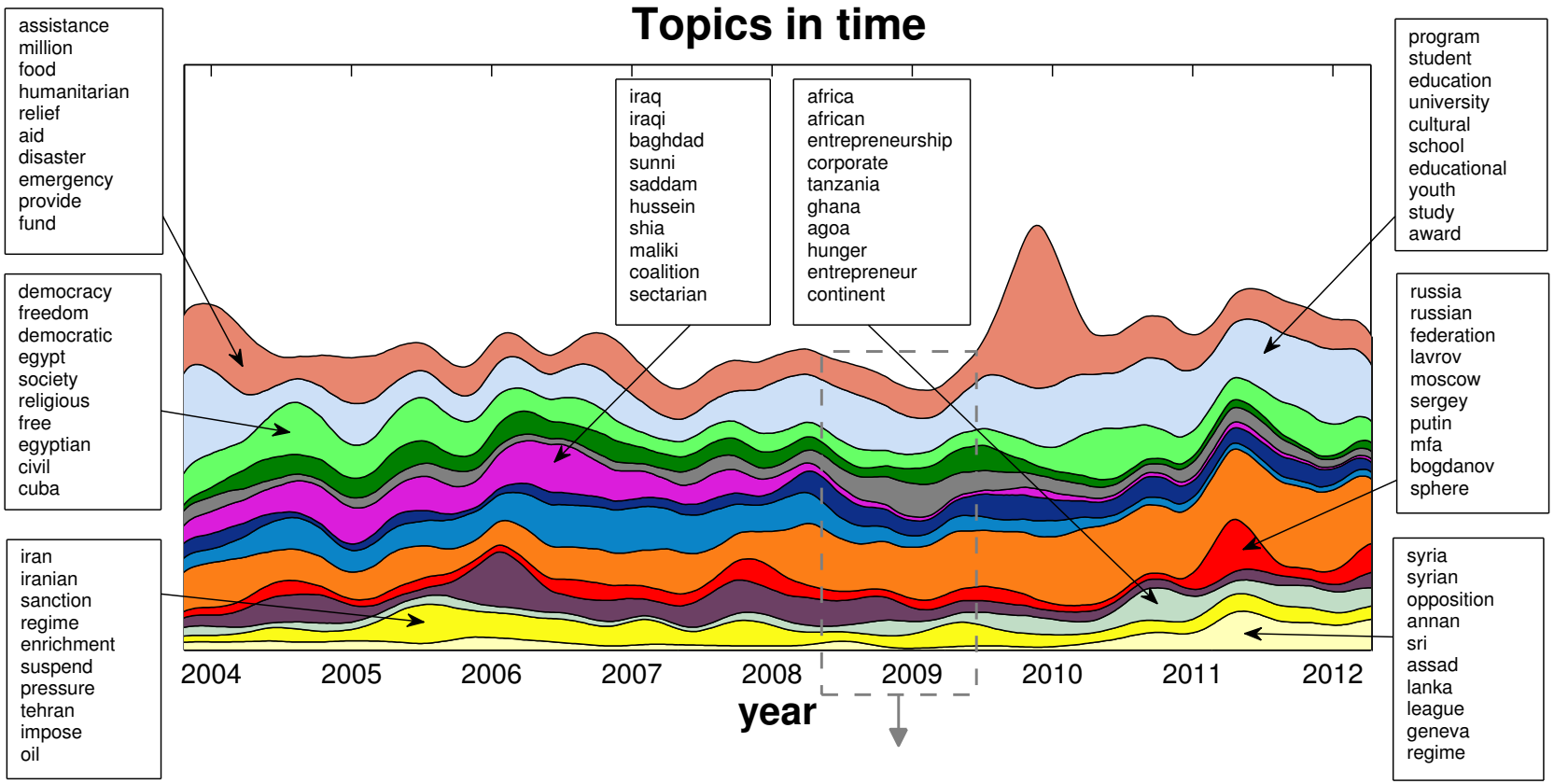
- Время дискретно, $i \in I$ — интервалы времени
- Как и в TOT, темы $p(w|t)$ не меняются во времени
- *Перманентные* темы имеют медленно меняющиеся $p(i|t)$
- *Событийные* темы имеют $p(i|t) = 0$ почти всё время
- Метки времени приписываются документам, а не словам
- Параметрические модели не используются

Цели моделирования:

- какие темы общие, какие специфичны для источников?
- какие темы событийные, какие перманентные?
- какие темы и когда коррелируют с заданной темой?

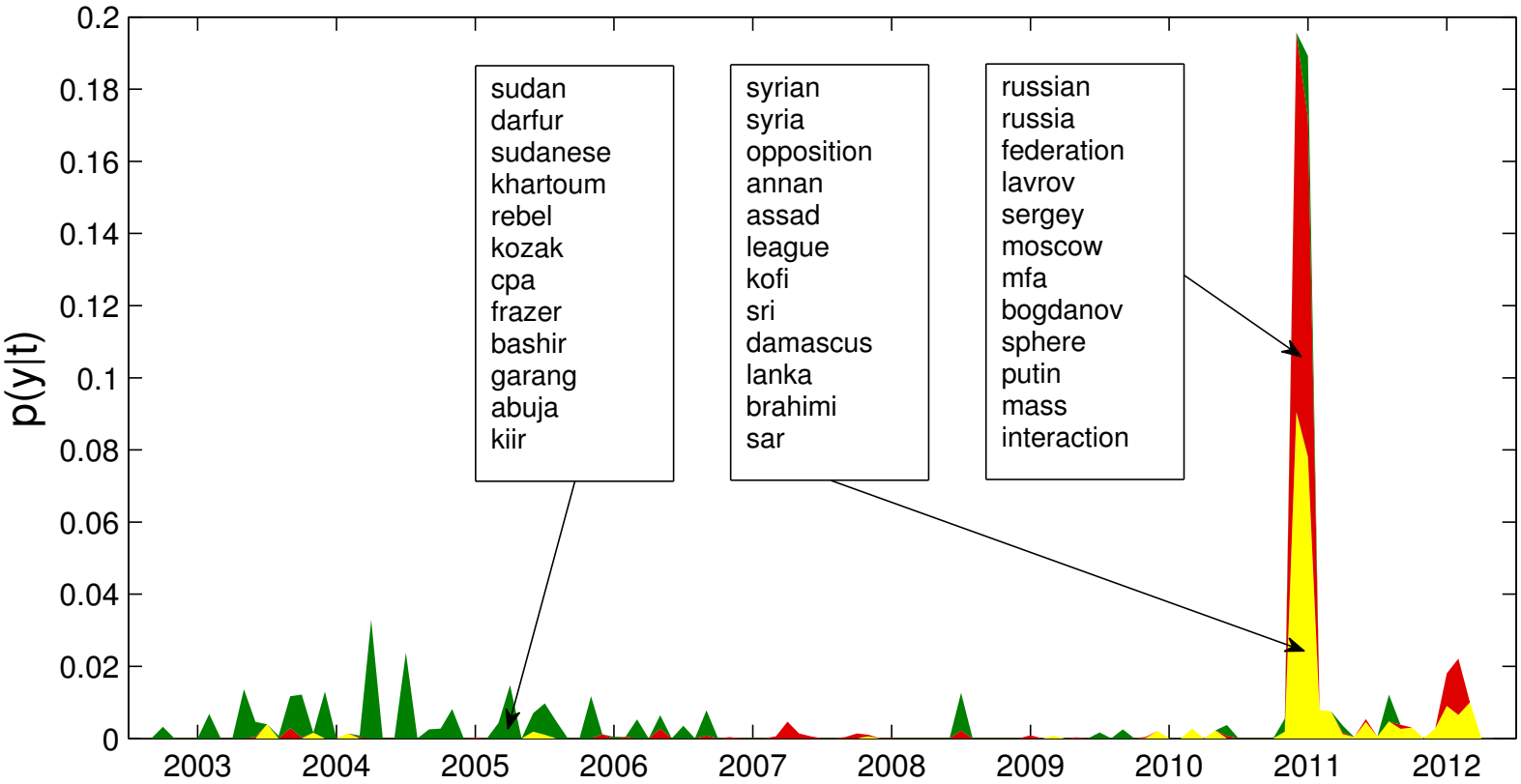
Пример 5. Динамика тем во времени

Коллекция официальных пресс-релизов внешнеполитических ведомств ряда стран на английском языке.
Более 20 тыс. сообщений за 10 лет, 180Мб текста.



Пример 5. Динамика тем во времени

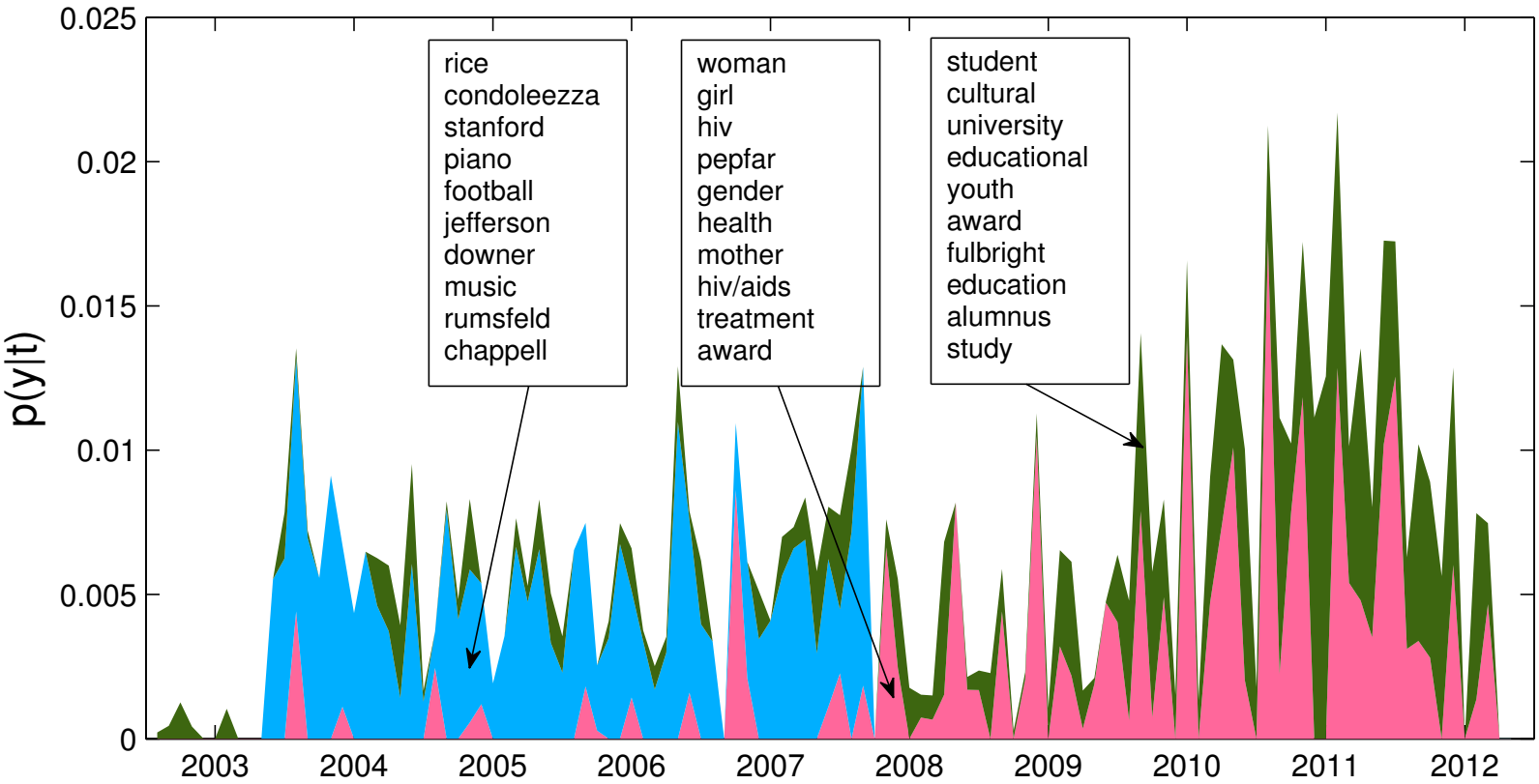
Пример: событийные темы и момент их совместного всплеска



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // Бакалаврская диссертация, ВМК МГУ. 2015.

Пример 5. Динамика тем во времени

Примеры перманентных тем (сглаживание отключено)



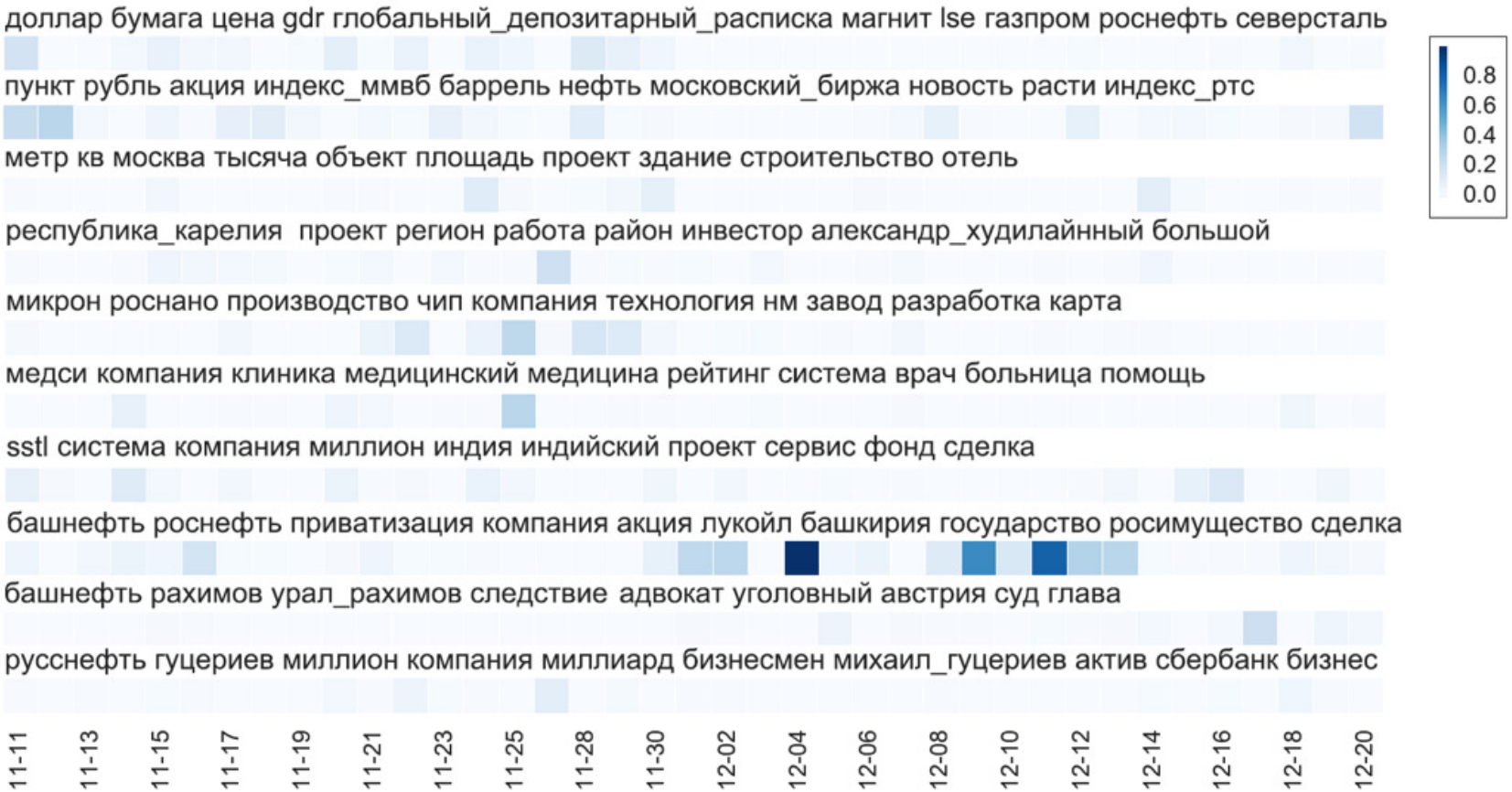
Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Пример 6. Мониторинг сообщений СМИ по компаниям

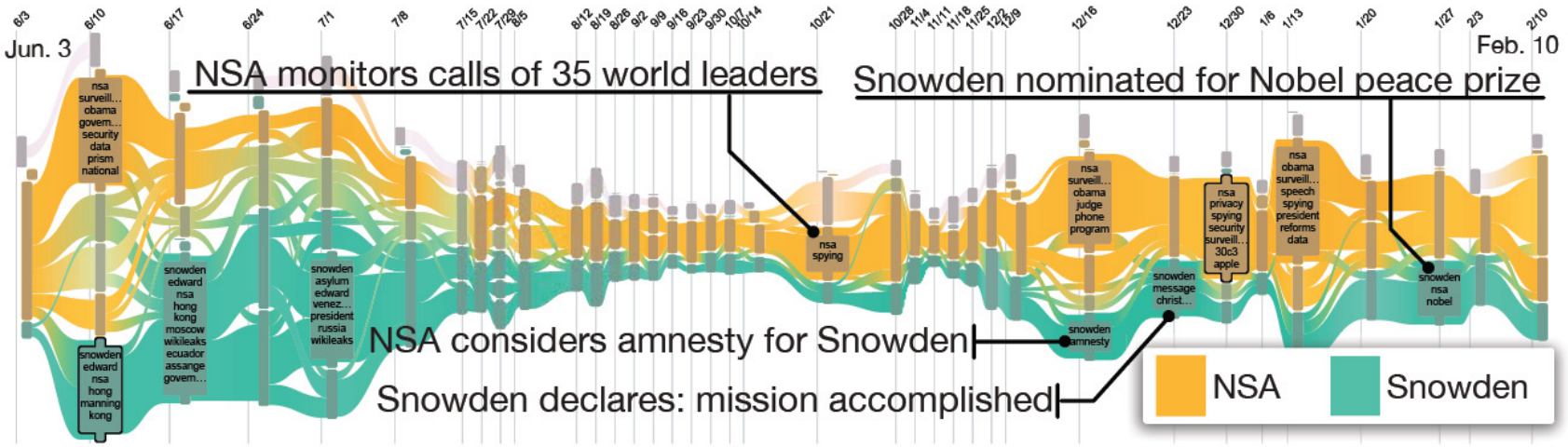
Темы проблемных ситуаций (АФК Система, ноя-дек 2016)

Пример 6. Мониторинг сообщений СМИ по компаниям

Новые выявленные темы (АФК Система, ноя-дек 2016)



Динамика тем: эволюция предметной области



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

Стратегия визуализации в системах TextFlow и RoseRiver:

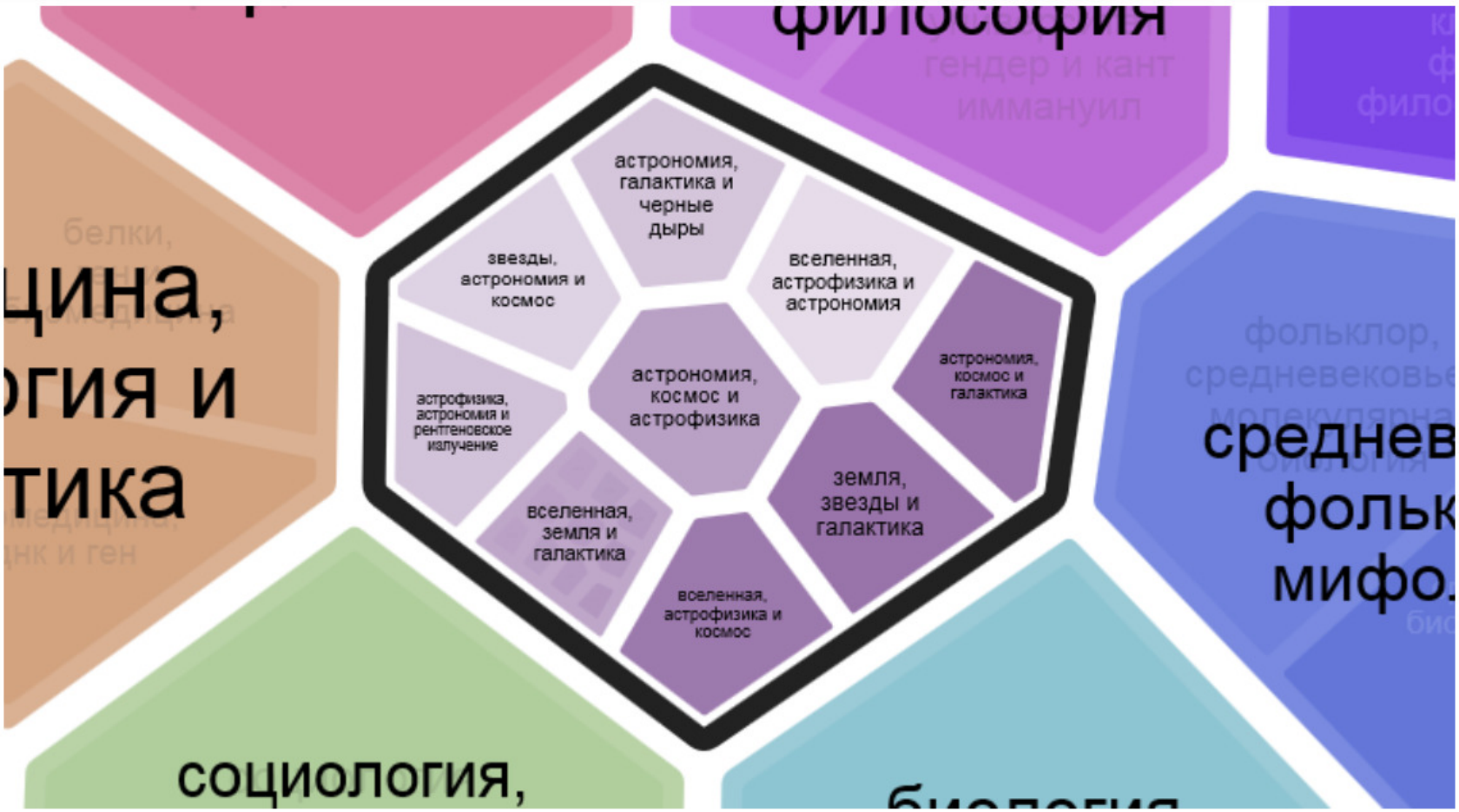
- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- получает сгенерированный отчёт с инфографикой.

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

Иерархические тематические модели

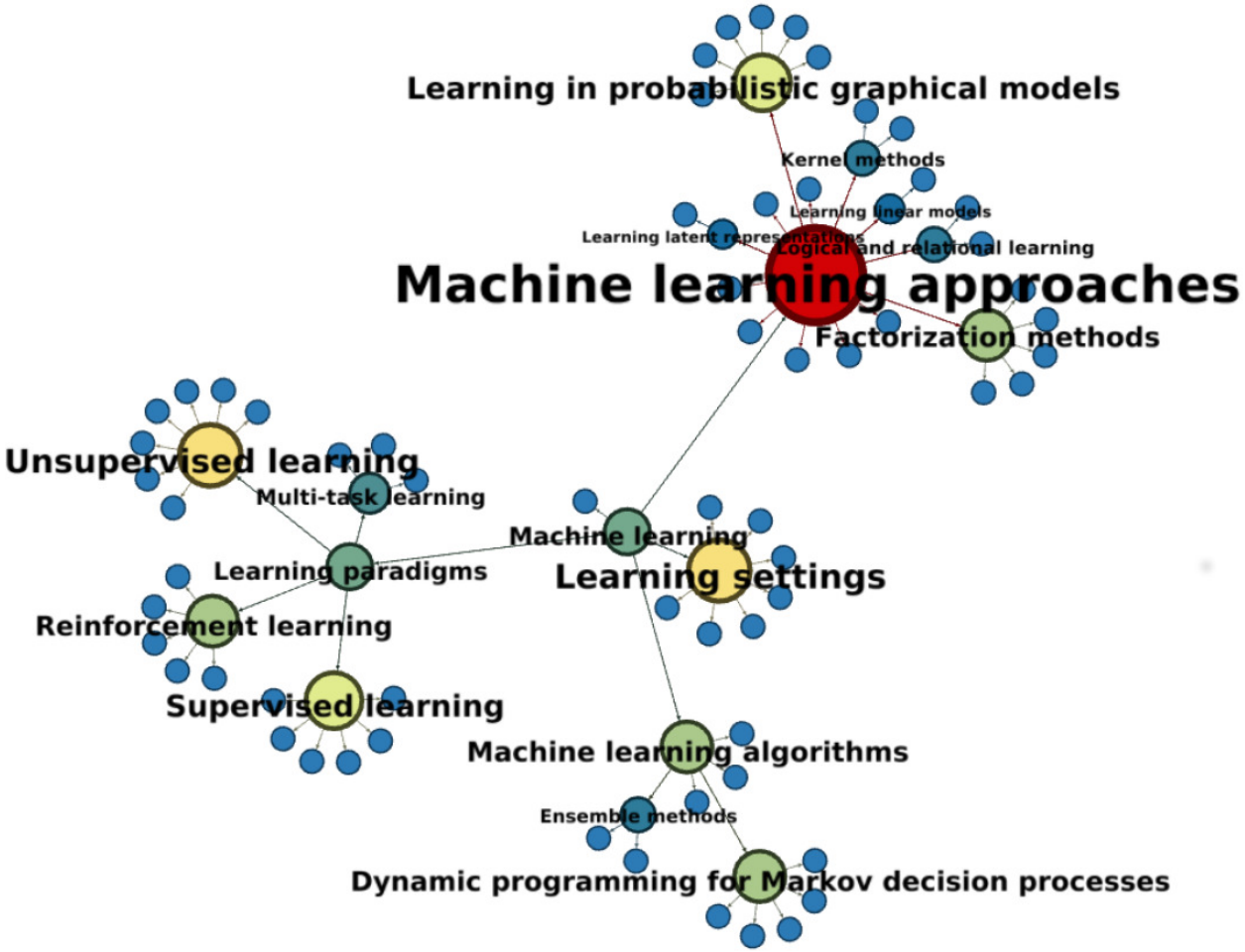
https://carrotsearch.com/foamtree-overview

Иерархические тематические модели



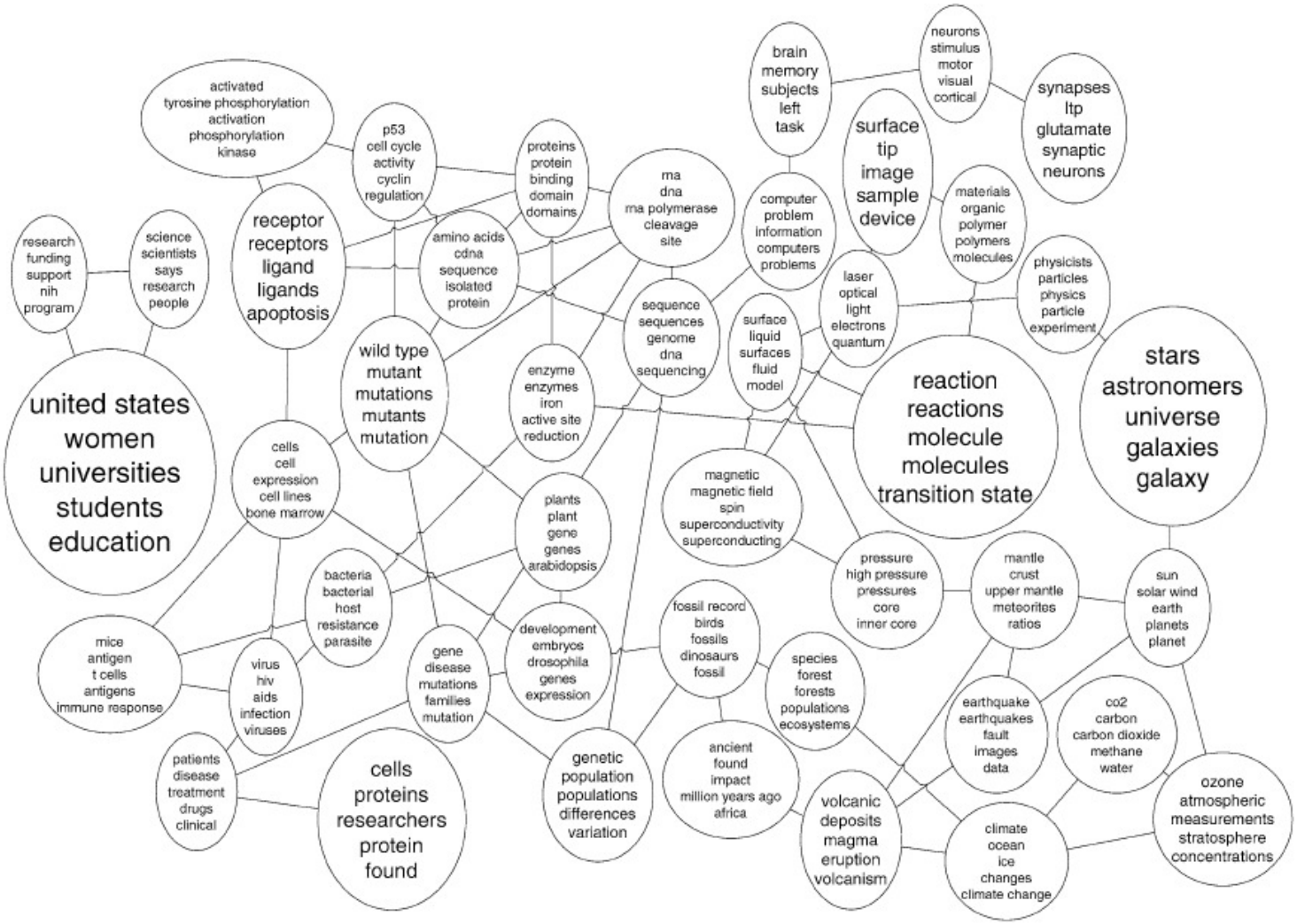
https://carrotsearch.com/foamtree-overview

Иерархические тематические модели



Georgeta Bordea. Domain adaptive extraction of topical hierarchies for Expertise Mining. 2013.

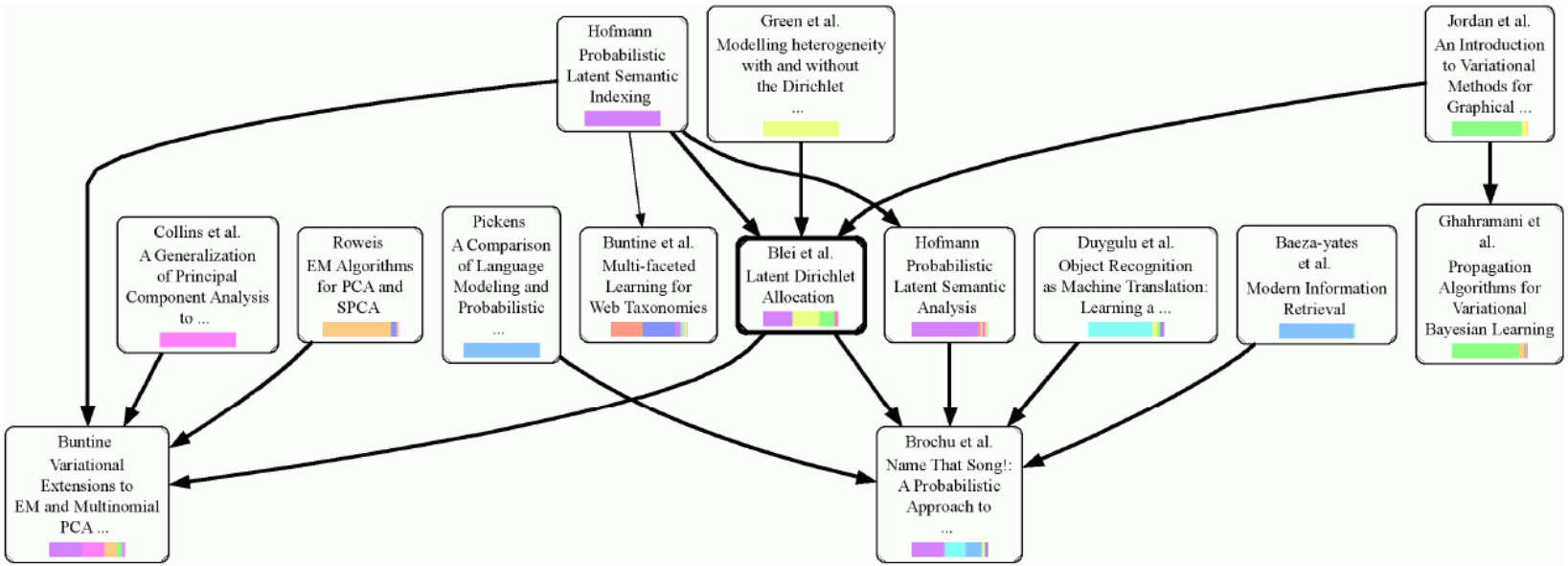
Выявление взаимосвязей между темами



D. Blei, J. Lafferty. A correlated topic model of Science. 2007.

Тематические модели цитирования и гиперссылок

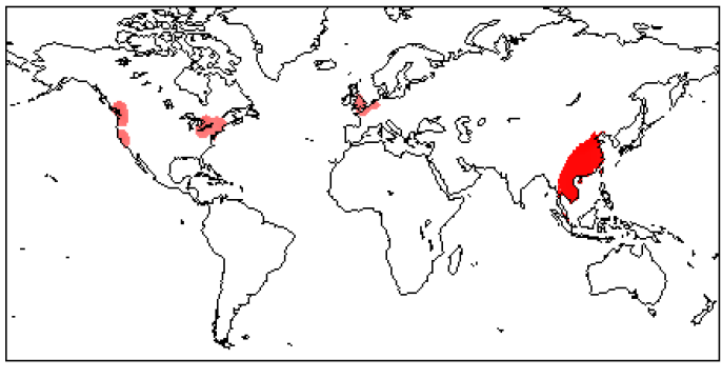
- Учёт ссылок уточняет тематическую модель
- Тематическая модель выявляет влиятельные ссылки



Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences // ICML-2007, Pp. 233–240.

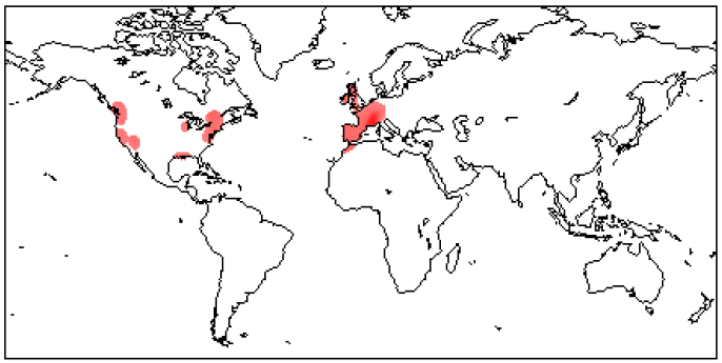
Гео-пространственные тематические модели

Пример: Food dataset. Где и что едят пользователи Flickr?

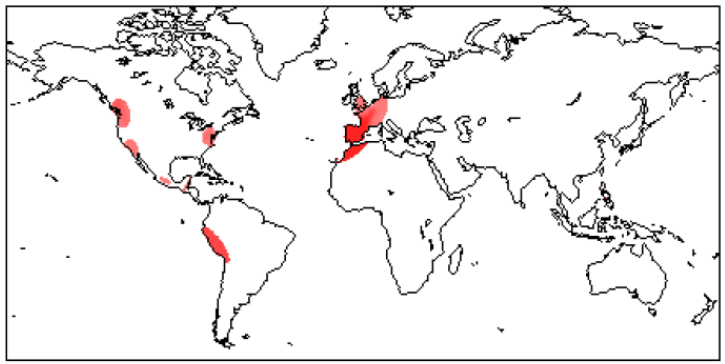


Chinese Food

Japanese Food



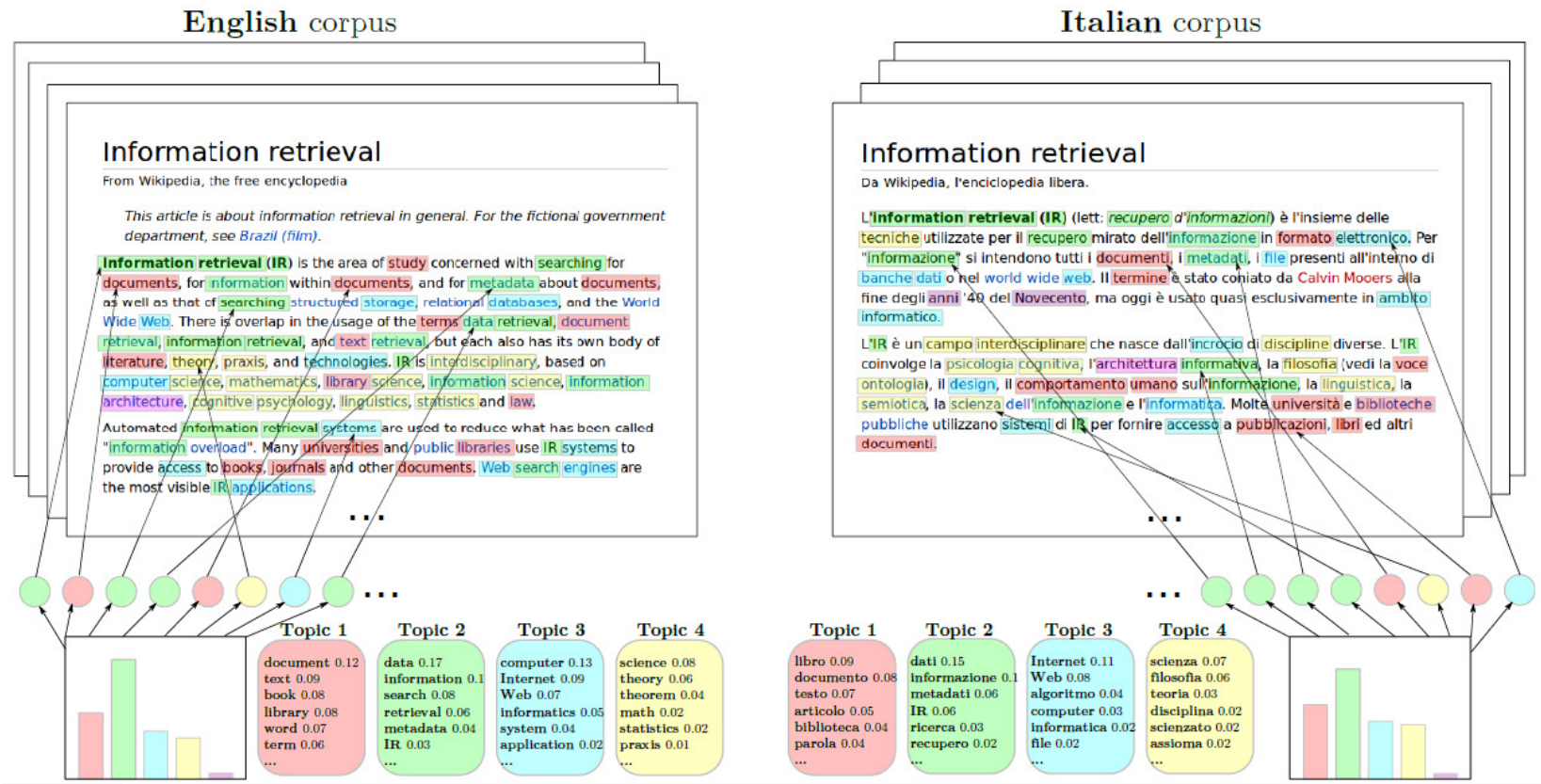
French Food



Spanish Food

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, Thomas Huang. Geographical Topic Discovery and Comparison // WWW 2011.

Многоязычные модели параллельных коллекций



Неожиданное открытие: двуязычные словари не нужны!

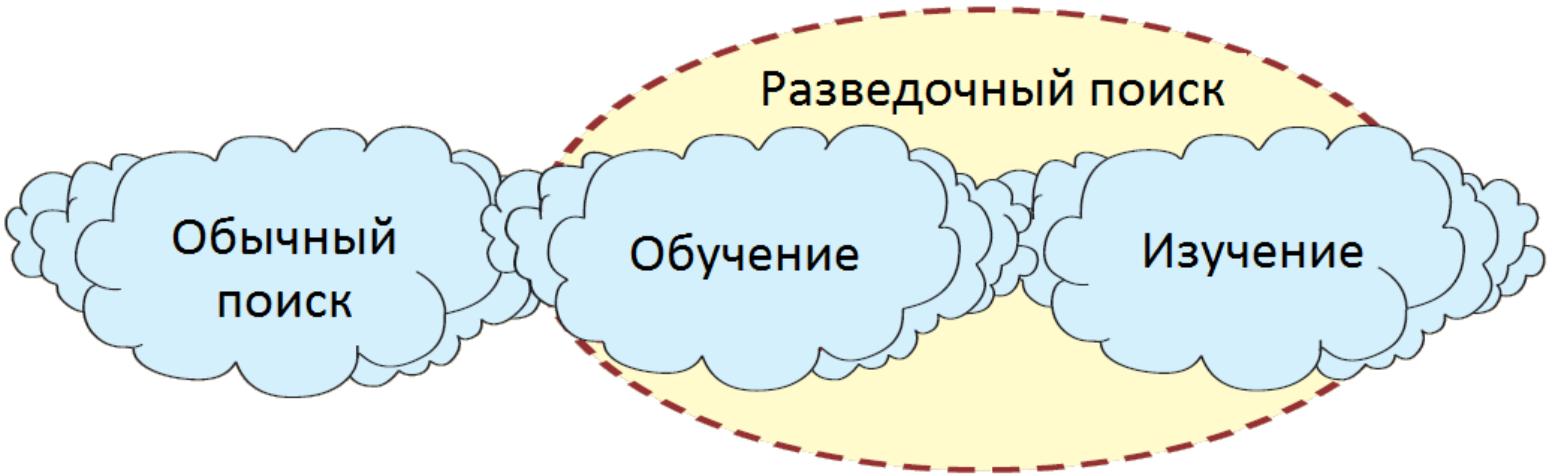
I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications. 2012.

Приложения тематического моделирования

- Разведочный информационный поиск (exploratory search)
- Поиск трендов, фронта исследований (research front)
- Поиск экспертов, рецензентов, подрядчиков (expert search)
- Анализ, агрегирование, фильтрация новостных потоков
- Рекомендательные системы
- Категоризация и классификация текстовых документов
- Аннотирование и суммаризация текстовых документов
- Тематическая сегментация текстовых документов
- Аннотирование изображений, видео, музыки
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов

Концепция разведочного поиска (exploratory search)

- пользователь может не знать ключевых терминов,
- запросом может быть текст произвольной длины,
- информационной потребностью — систематизация знаний



навигация в сети, поиск фактов, упоминаний, конкретных ответов	самообразование, тематический поиск систематизация знаний	исследование, экспертиза, реферирование, мониторинг тем

Gary Marchionini. Exploratory Search: from finding to understanding. 2006.

Возможный сценарий разведочного поиска

Поисковый запрос:

- документ любой длины или даже коллекция документов

Цели поиска:

- к каким темам относится мой запрос?
- что ещё известно по этим темам?
- какова тематическая структура этой предметной области?
- какие области являются смежными?
- что ещё есть понятного, обзорного, важного, свежего?

Сценарий поиска:

1. имея любой текст под рукой, в любом приложении,
2. получаем картину содержащихся в нём тем-подтем
3. и «дорожную карту» предметной области в целом

От ближнего чтения (close reading) к дальнему (distant reading)

Концепция дальнего чтения Франко Моретти

Дальнее чтение — это специальная форма представления знаний: меньше элементов, грубее смысл их взаимосвязей, важны лишь общие очертания и структуры.

Мантра Шнейдермана

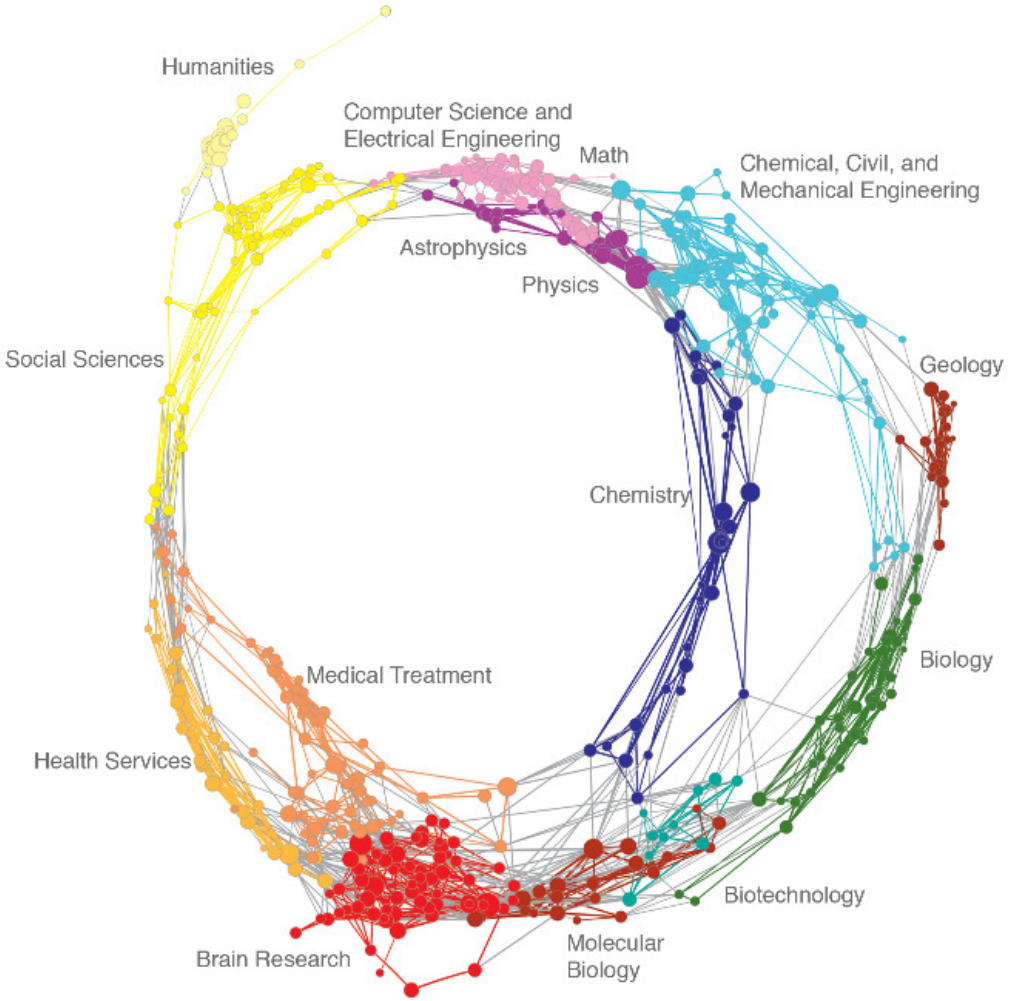
«Сначала крупный план, затем масшабирование и фильтрация, детали по требованию»

B.Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Visual Languages, 1996.

F.Moretti. Graphs, Maps, Trees: Abstract Models for a Literary History. 2005.

S.Janicke, G.Franzini, M.F.Cheema, G.Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis, 2015.

Пример карты науки



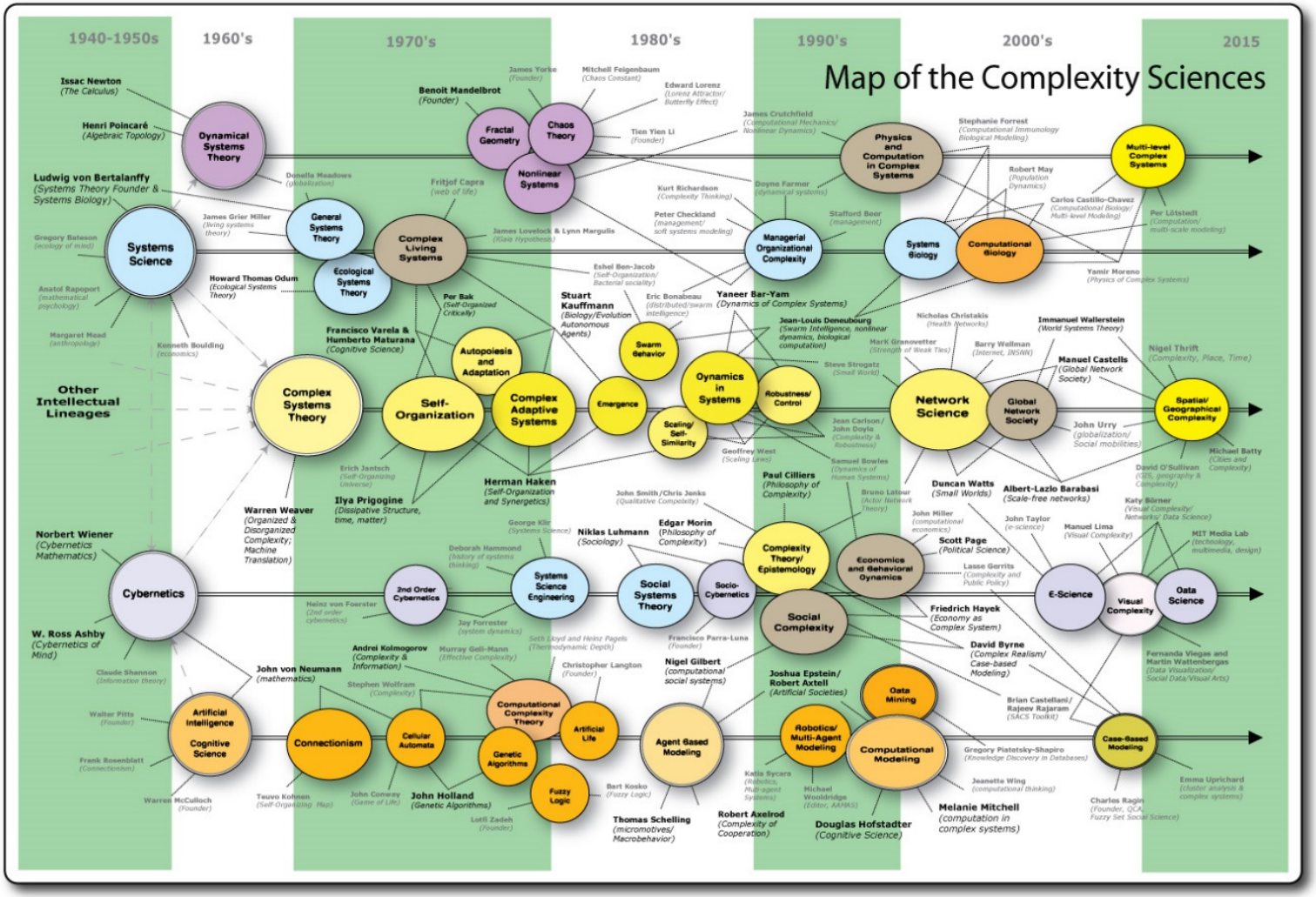
http://scimaps.org

Важное наблюдение:
области знания самопроизвольно располагаются по кругу, значит,
их можно располагать и вдоль прямой линии.

Недостатки:

- оси не имеют интерпретации
- искажение сходства при двумерном проецировании

Пример карты предметной области, построенной вручную



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

Визуализация тематического разведочного поиска (концепт)

- Оси интерпретируемы: время–темы или сложность–темы
- Спектр тем: гуманитарные $\rightarrow$ естественные $\rightarrow$ точные
- Иерархичность: темы делятся на подтемы
- Интерактивность: реализация мантры Шнейдермана
- Фрактальность: при любом масштабе текста на карте много

Комбинирование требований к тематическим моделям

- **Динамическая**: выявление истории развития тем
- **Иерархическая**: выявление иерархических связей тем
- **Интерпретируемая**: каждая тема понятна людям
- **Мультиграммная**: термины-словосочетания неразрывны
- **Мультимодальная**: авторы, связи, теги, пользователи,...
- **Мультиязычная**: для кросс- и много-языкового поиска
- **Сегментирующая**: выделение тем внутри документа
- **Обучаемая** по оценкам асессоров и логам пользователей
- **Определяющая число тем** автоматически
- **Создающая и именующая новые темы** автоматически
- **Онлайновая**: обрабатывающая коллекцию за 1 проход
- **Параллельная, распределённая** для больших коллекций

Постановка задачи тематического моделирования

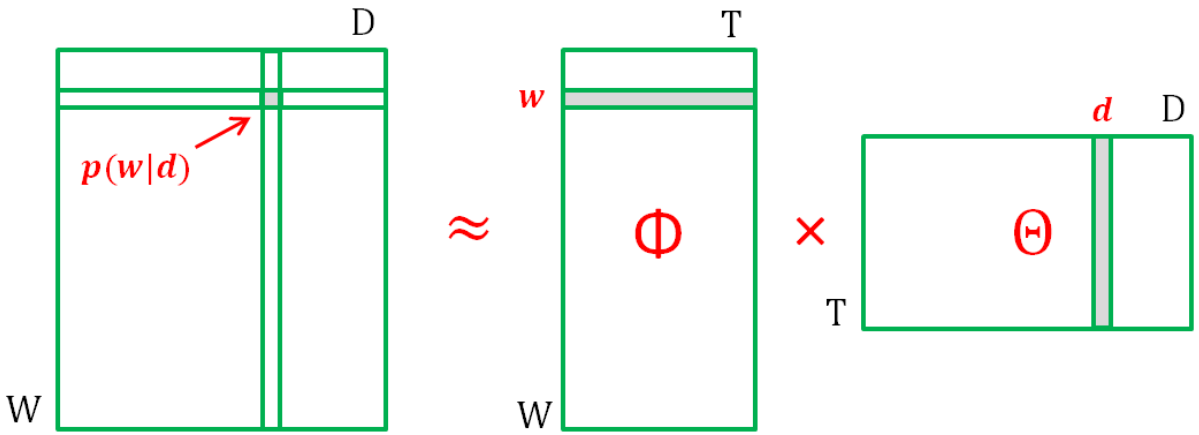
Дано: коллекция текстовых документов

- n_{dw} — частота термина w в документе d, $p(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической модели $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$

- $\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:



Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

$$\prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}$$

Максимизация логарифма правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) p(d) \to \max_{\Phi, \Theta}$$

эквивалентна максимизации функционала

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

при ограничениях неотрицательности и нормировки

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

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар
(1865–1963)

Задача стохастического матричного разложения *некорректно поставлена*, так как имеется бесконечное множество решений:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

для невырожденных $S_{T\times T}$ таких, что Φ', Θ' — стохастические.

Регуляризация — стандартный приём, введение новых ограничений или критериев, доопределяющих решение.

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

Онлайновый EM-алгоритм с регуляризацией

Вход: коллекция D, число тем $|T|$, параметры $i_{\max}$, $j_{\max}$, γ;
Выход: матрицы терминов тем Θ и тем документов Φ;

инициализировать $n_{wt} := 0$;
для всех $i = 1, \ldots, i_{\max}$ (итерации по коллекции)
  для всех документов $d \in D$
    для всех $j = 1, \ldots, j_{\max}$ (итерации по документу)
      $p_{tdw} := \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$;
      $\theta_{td} := \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$;
    $n_{wt} := \gamma n_{wt} + n_{dw} p_{tdw}$;
    если пора обновить матрицу Φ **то**
      $\phi_{wt} := \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$;

Сравнение оффлайнового и онлайнового алгоритмов

Оффлайн EM-алгоритм:

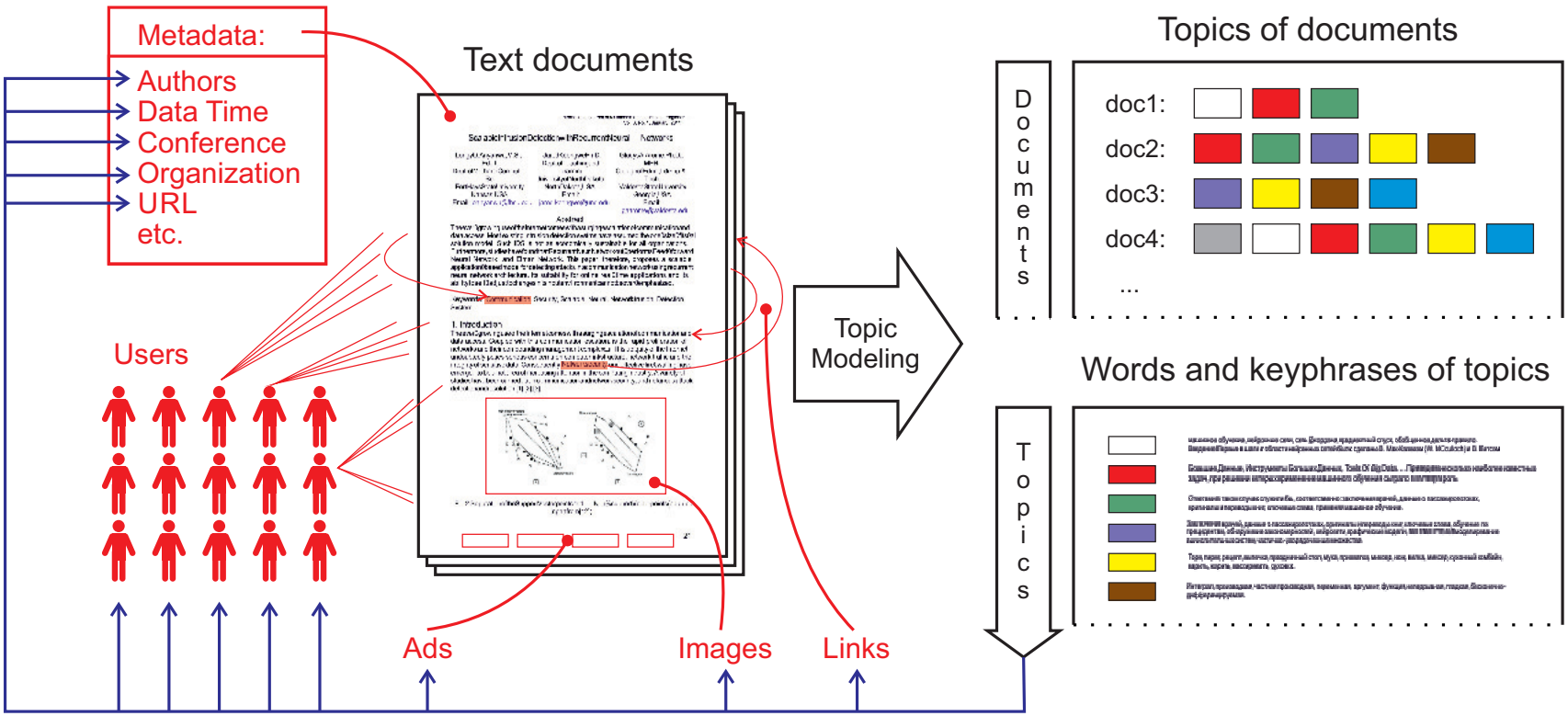
1. многократное итерирование по коллекции
2. однократный проход по документу
3. хранение матрицы Θ
4. обновление Φ в конце каждого прохода по коллекции
5. применяется при обработке небольших коллекций

Онлайн EM-алгоритм:

1. однократный проход по коллекции
2. многократное итерирование по каждому документу
3. нет необходимости хранить матрицу Θ
4. обновление Φ через заданное число пакетов
5. применяется при потоковой обработке больших коллекций

Задачи мультимодального тематического моделирования

Темы определяют распределения не только терминов $p(w|t)$, но и других *модальностей*: $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{элемент\_изображения}|t)$, $p(\text{пользователь}|\text{t}), \ldots$



Мультимодальная ARTM

W^m — словарь токенов m-й модальности, $m \in M$

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Обзор полезных регуляризаторов

- Разреживание матриц Φ и Θ
- Сглаживание фоновых тем для выделения общей лексики
- Декоррелирование (усиление различности) тем
- Отбор незначимых, дублирующих и зависимых тем
- Связывание тем с родительским уровнем иерархии
- Сглаживание $p(i|t)$ в темпоральных моделях
- Разреживание $p(t|i)$ для выделения событийных тем
- Обучение с учителем для классификации и регрессии
- Модель дистрибутивной семантики WNTM, аналог word2vec
- Модель битермов Biterm-TM для коротких текстов (twitter)

Воронцов К. В. Обзор вероятностных тематических моделей, 2017.
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

Примеры регуляризаторов

1. разреживание и сглаживание:

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

2. декоррелирование тем как столбцов Φ:

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

3. учёт *битермов* — пар слов, близко стоящих n_{uw} раз:

$$R(\Phi) = \tau \sum_{u,w} n_{uw} \ln \sum_t n_t \phi_{ut}\phi_{wt} \to \max$$

4. удаление неинформативных тем:

$$R(\Theta) = -\tau \sum_t \ln p(t) \to \max, \qquad p(t) = \sum_d \theta_{td} p(d)$$

Критерии качества тематических моделей

Внешние критерии:

- Полнота и точность тематического поиска
- Качество ранжирования при тематическом поиске
- Качество тематических рекомендаций
- Качество категоризации документов
- Экспертные оценки качества тем

Внутренние критерии:

- Перплексия
- Средняя когерентность (согласованность) тем
- Разреженность матриц Φ и Θ

Интерпретируемости и когерентность

Тема интерпретируемая, если по топовым словам темы эксперт может определить, о чём эта тема, и дать ей название.

- Экспертные оценки:
  — интерпретируемость темы по балльной шкале;
  — каждую тему оценивают несколько экспертов.
- Метод интрузий (intrusion):
  — в список топовых слов внедряется лишнее слово;
  — измеряется доля ошибок экспертов его при определении

Нужна автоматически вычисляемая мера интерпретируемости, коррелирующая с экспертными оценками.

Ею оказалась *когерентность* (согласованность, coherence).

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Эксперимент. Связь когерентности и интерпретируемости

Измерялась ранговая корреляция Спирмена между 15 метрикам и экспертными оценками интерпретируемости.

PMI — лучшая метрика.

Gold-standard — средняя корреляция Спирмена между оценками разных экспертов.

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	−0.20	0.19
	LCH	−0.31	−0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	−0.08	0.27
	WUP	0.41	0.26
Wikipedia	RACO	0.62	0.69
	MIW	0.68	0.70
	DOCSIM	0.59	0.60
	PMI	**0.74**	**0.77**
Google	TITLES	0.51	
	LOGHITS	−0.19	
Gold-standard	IAA	0.82	0.78

Вывод: когерентность близка к «золотому стандарту».

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Когерентность как внутренняя мера интерпретируемости

Когерентность (согласованность) темы t по k топовым словам:

$$\mathrm{PMI}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^{k} \mathrm{PMI}(w_i, w_j)$$

где w_i — i-й термин в порядке убывания ϕ_{wt}.

$\mathrm{PMI}(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация (pointwise mutual information),

N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов),

N_u — число документов, в которых u встретился хотя бы 1 раз.

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Предварительная обработка текстовой коллекции

- Удаление форматирования, переносов, не-текста
- Исправление опечаток
- Слияние слишком коротких текстов
- Лемматизация (русский) или стемминг (английский)
- Выделение терминов (term extraction)
- Выделение именованных сущностей (named entities)
- Удаление стоп-слов и слишком редких слов
- Подготовка батчей для пакетной обработки

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый параллельный мультимодальный ARTM
- Большие данные: коллекция не хранится в памяти
- 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:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

BigARTM упрощает разработку тематических моделей

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

Этапы моделирования	Bayesian TM	ARTM	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная порождающая модель данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Общий регуляризованный EM-алгоритм для любых моделей	
Реализация:	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)	
Оценивание:	Исследовательские метрики, исследовательский код	Стандартные метрики	Свои метрики
	Внедрение	Внедрение	

-- *нестандартизуемые этапы, уникальная разработка для каждой задачи*

-- *стандартизуемые этапы*

Тесты производительности

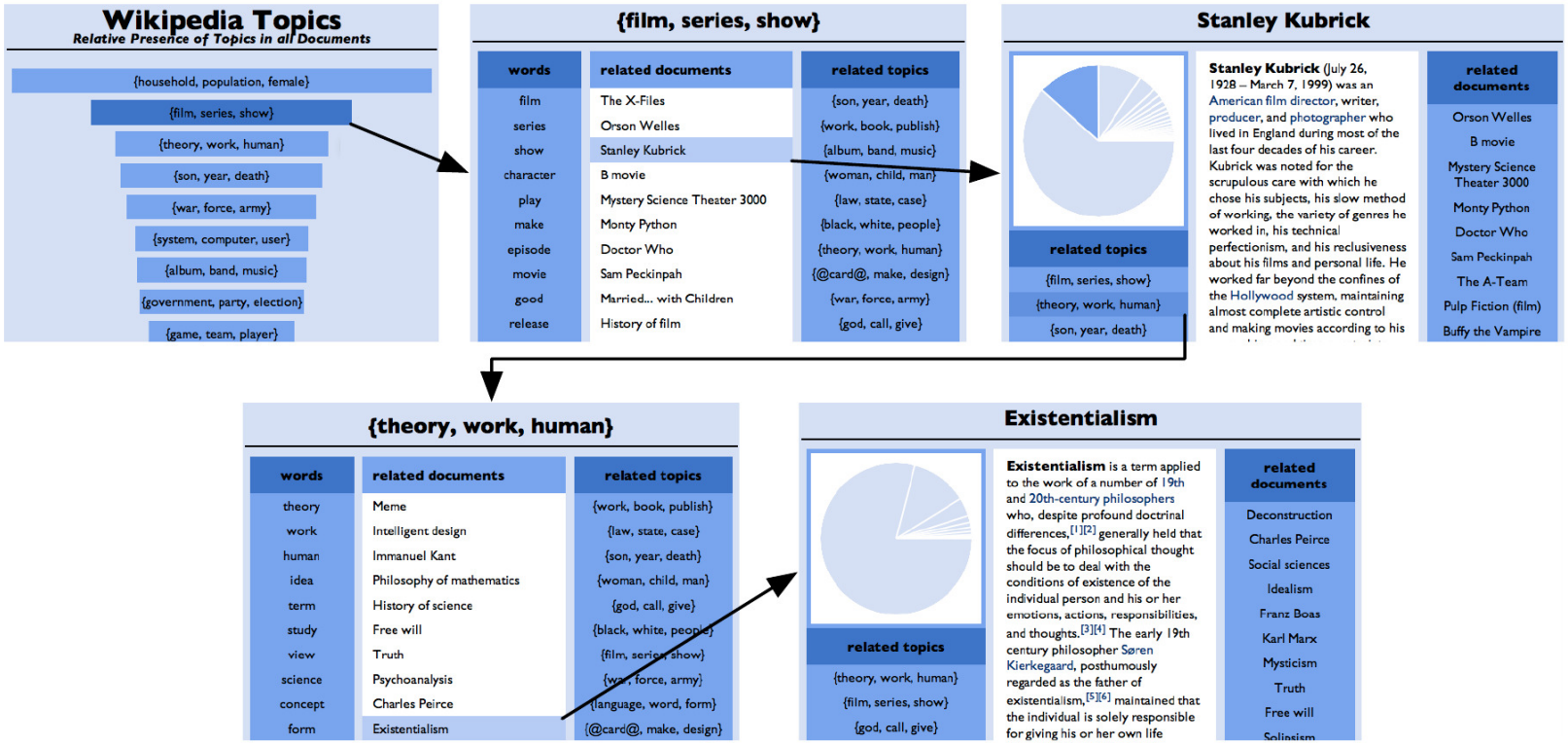
- 3.7M статей английской Вики, 100K уникальных слов

	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

- *procs* = число параллельных потоков
- *inference* = время тематизации 100K тестовых документов
- *perplexity* = перплексия на тестовой выборке документов

Система TMVE — Topic Model Visualization Engine

Тематический навигатор с веб-интерфейсом:

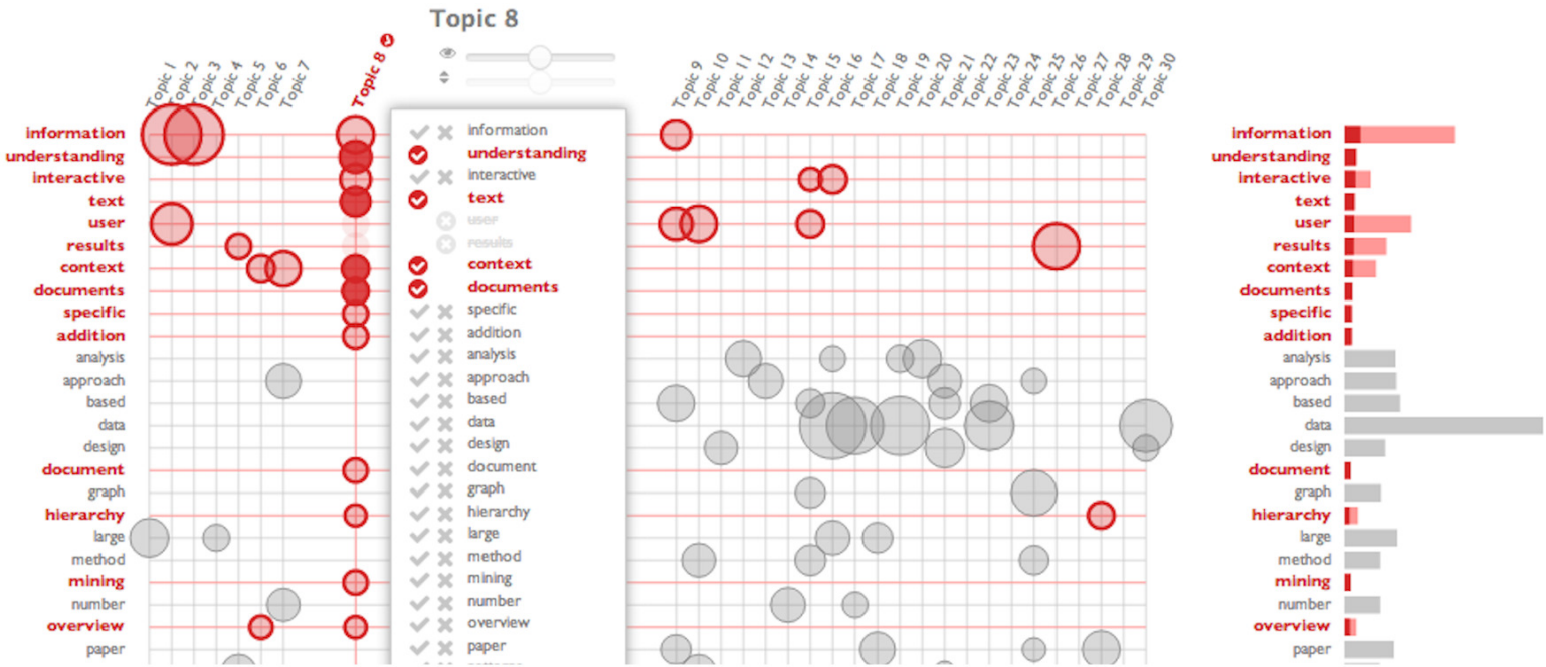


https://github.com/ajbc/tmv

Chaney A., Blei D. Visualizing Topic Models // Frontiers of computer science in China, 2012. — 55(4), pp. 77–84.

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.

http://textvis.lnu.se

Интерактивный обзор 380 средств визуализации текстов

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

Резюме

- Тематическое моделирование — инструмент для поиска и систематизации текстовой информации
- Когда нужна тематическая модель, обычно пользуются пост-обработкой тем, найденных LDA
- BigARTM позволяет оптимизировать темы под задачу, строить модели с заданными свойствами, комбинировать различные критерии и источники данных
- VisARTM — визуализация в BigARTM (в разработке)

http://bigartm.org

Литература

- *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
- *K.Vorontsov, A.Potapenko.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. AIST 2014.
- *K.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.
- *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Yanina.* Non-bayesian additive regularization for multimodal topic modeling of large collections. 2015.
- *K.Vorontsov, A.Potapenko, A.Plavin.* Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.
- *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.
- *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.
- *N.A.Chirkova, K.V.Vorontsov.* Additive Regularization for Hierarchical Multimodal Topic Modeling. JMLDA 2016.
- *А.О.Янина, К.В.Воронцов.* Мультимодальные тематические модели для разведочного поиска в коллективном блоге. JMLDA 2016.