Вероятностные тематические модели: разведочный информационный поиск, теория аддитивной регуляризации, проект BigARTM

К. В. Воронцов
vokov@forecsys.ru

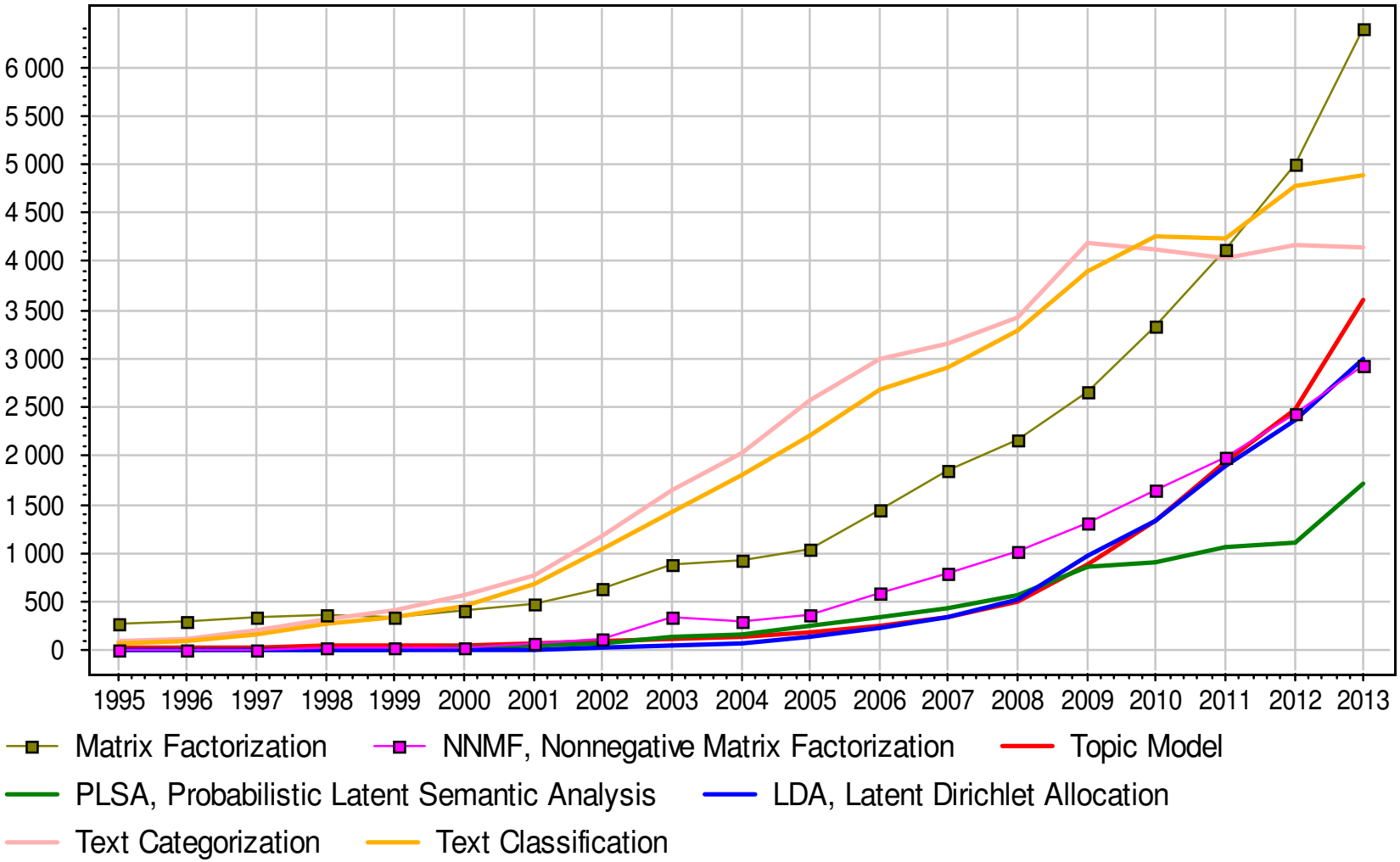
Курс лекций на странице вики-ресурса
http://www.MachineLearning.ru/wiki
Вероятностные тематические модели (курс лекций, К.В.Воронцов)

2016

Содержание

Тематическое моделирование и смежные области исследований

Динамика цитирования, по данным Google Scholar:



Что такое «тема»?

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.

Более формально,

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

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

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

Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей Википедии
Первые 10 слов с их вероятностями $p(w|t)$ в %:

Topic 68				Topic 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

Результат. Независимый асессор оценил 396 тем из $|T| = 400$ как хорошо интерпретируемые.

Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей Википедии
Первые 10 слов с их вероятностями $p(w|t)$ в %:

Topic 88				Topic 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

Результат. Независимый асессор оценил 396 тем из $|T| = 400$ как хорошо интерпретируемые.

Пример 2. Мультиграммная модель коллекции ММРО-ИОИ

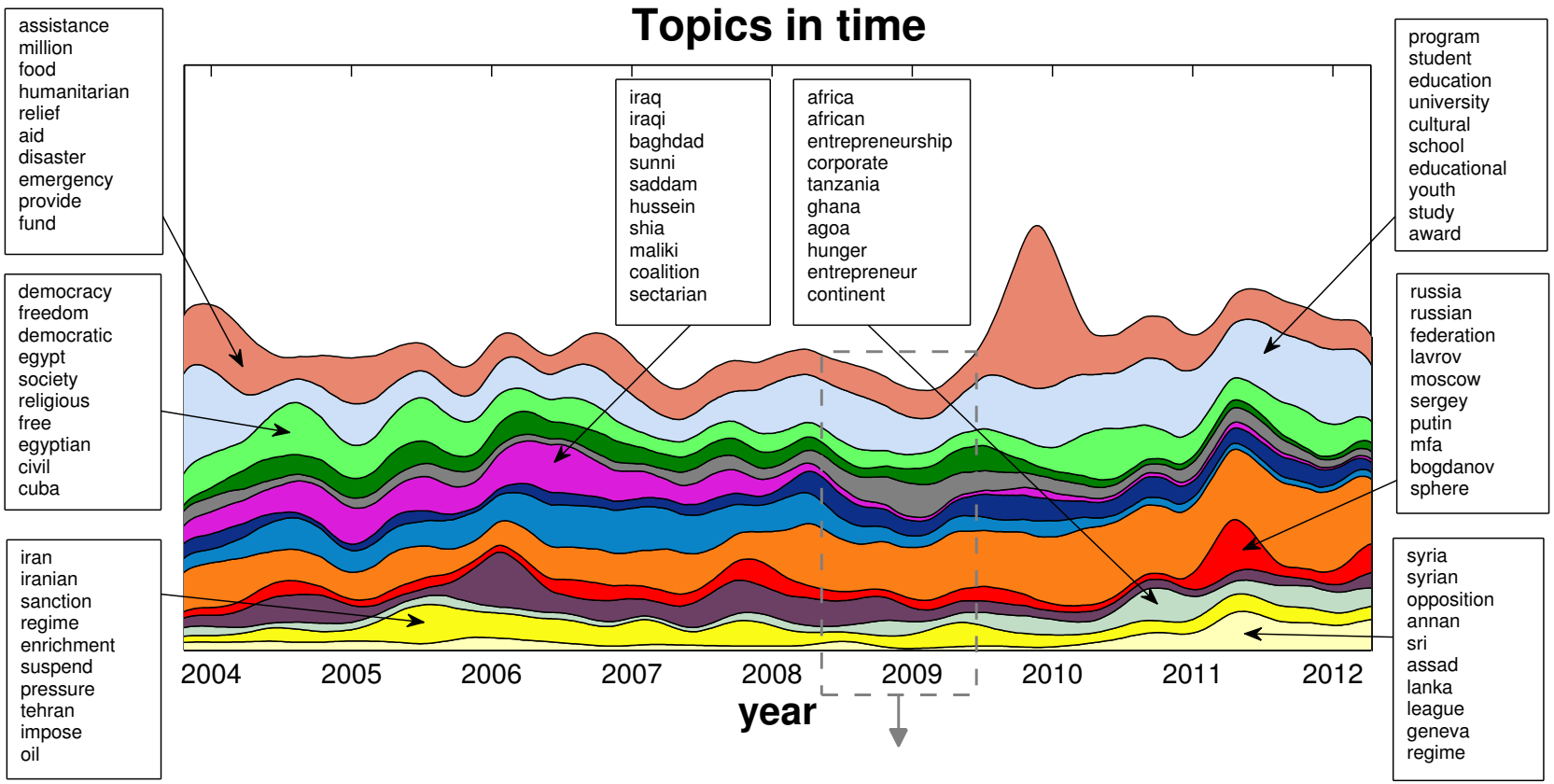
Коллекция 850 статей конференций ММРО, ИОИ на русском

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

Результат. Биграммная модель лучше интерпретируется.

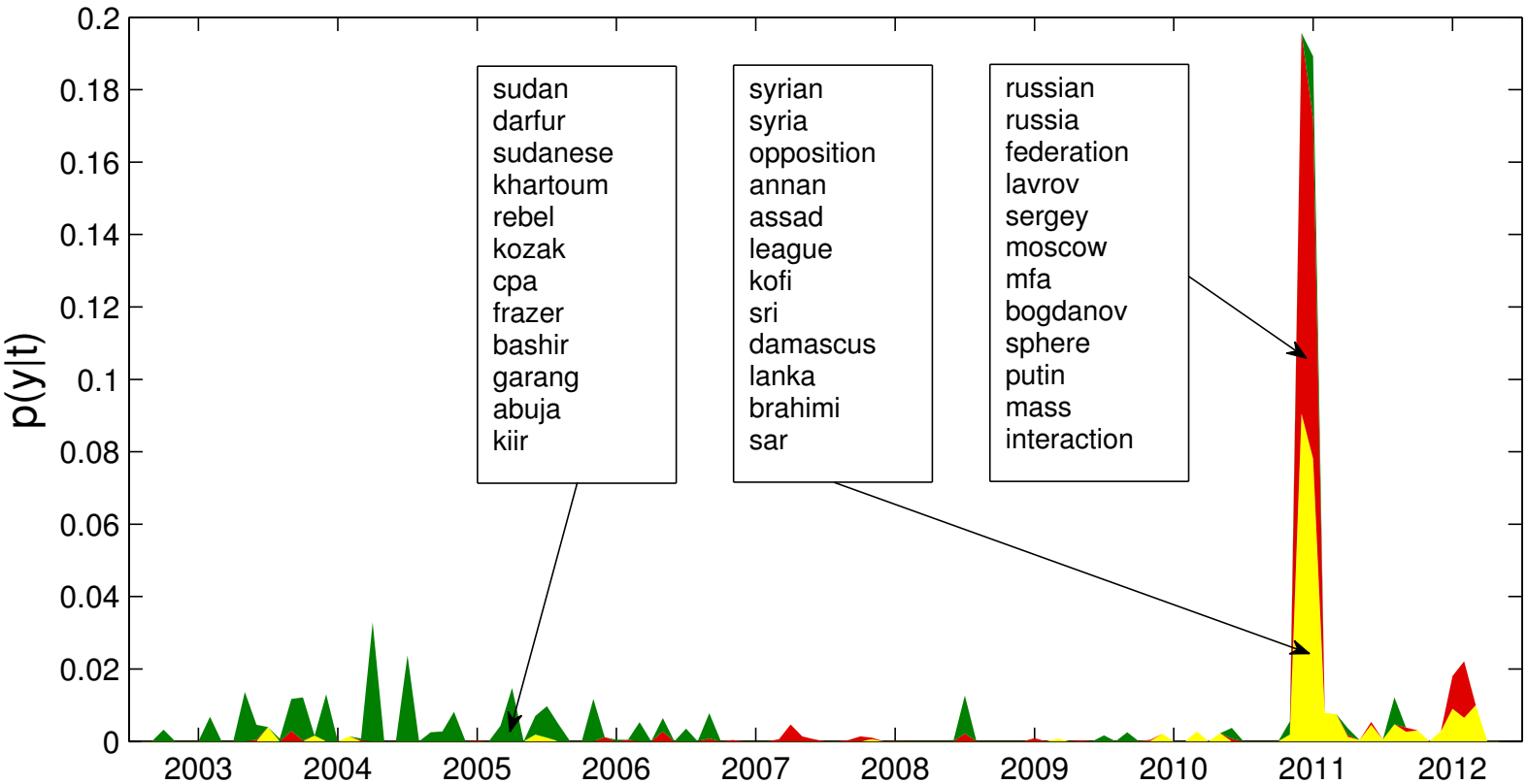
Пример 3. Темпоральная модель коллекции пресс-релизов

Коллекция внешнеполитических пресс-релизов ряда стран:
20 тыс. сообщений, 10 лет, 180Мб текста, английский язык.



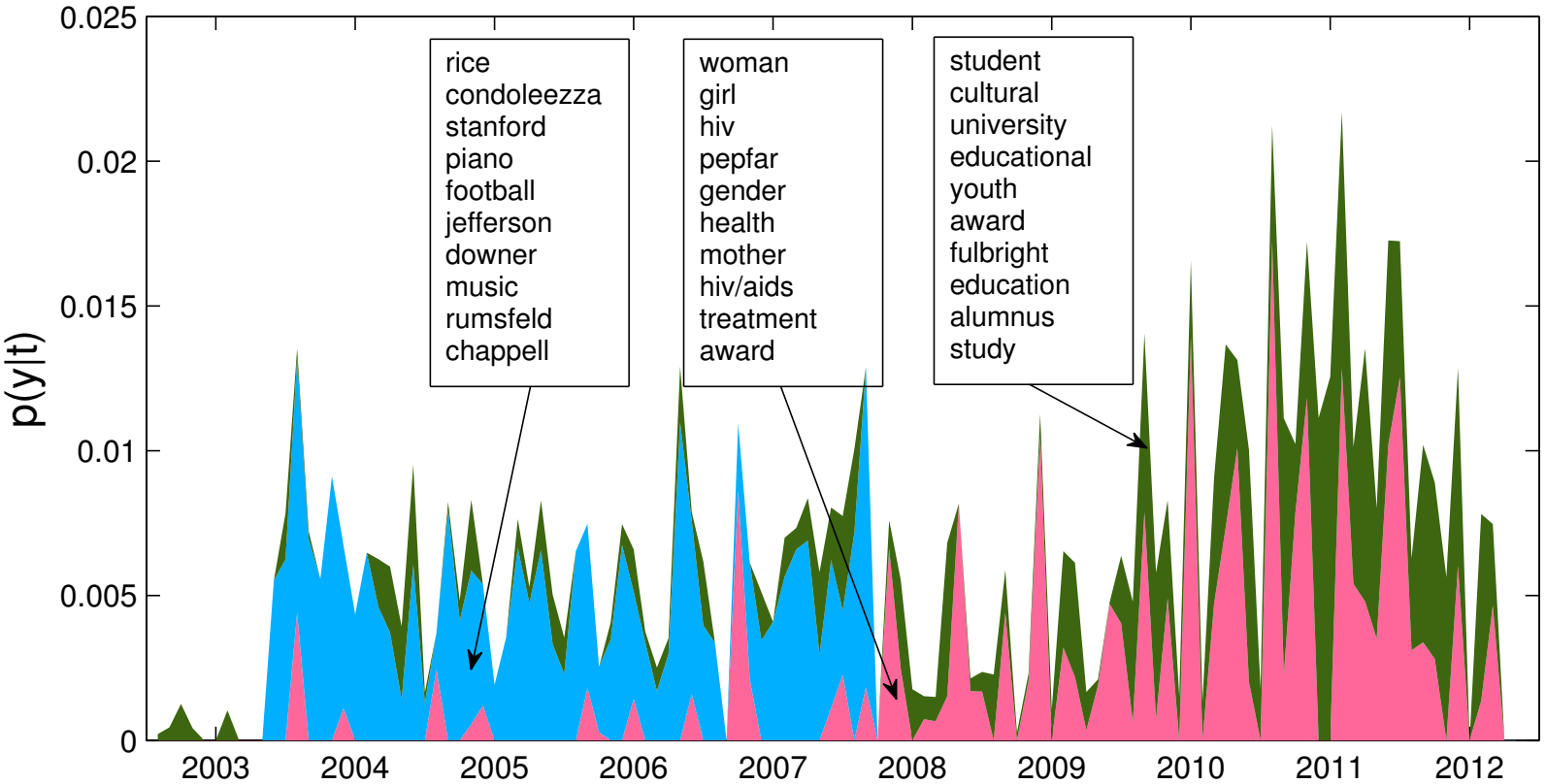
Пример 3. Темпоральная модель коллекции пресс-релизов

Примеры событийных тем и момента их совместного всплеска



Пример 3. Темпоральная модель коллекции пресс-релизов

Примеры перманентных тем



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
- Оценивание враждебности, конфликности
- Поддержка социологических исследований

Вспомогательные задачи:

- Фильтрация (обогащение) потока данных
- Обеспечение полноты поиска этничных тем
- Выявление тематических сообществ
- Выделение событийных и региональных тем
- Решение проблемы коротких сообщений

Пример 4. Поиск этно-релевантных тем в социальных сетях

Примеры этнонимов, используемых для формирования тем

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

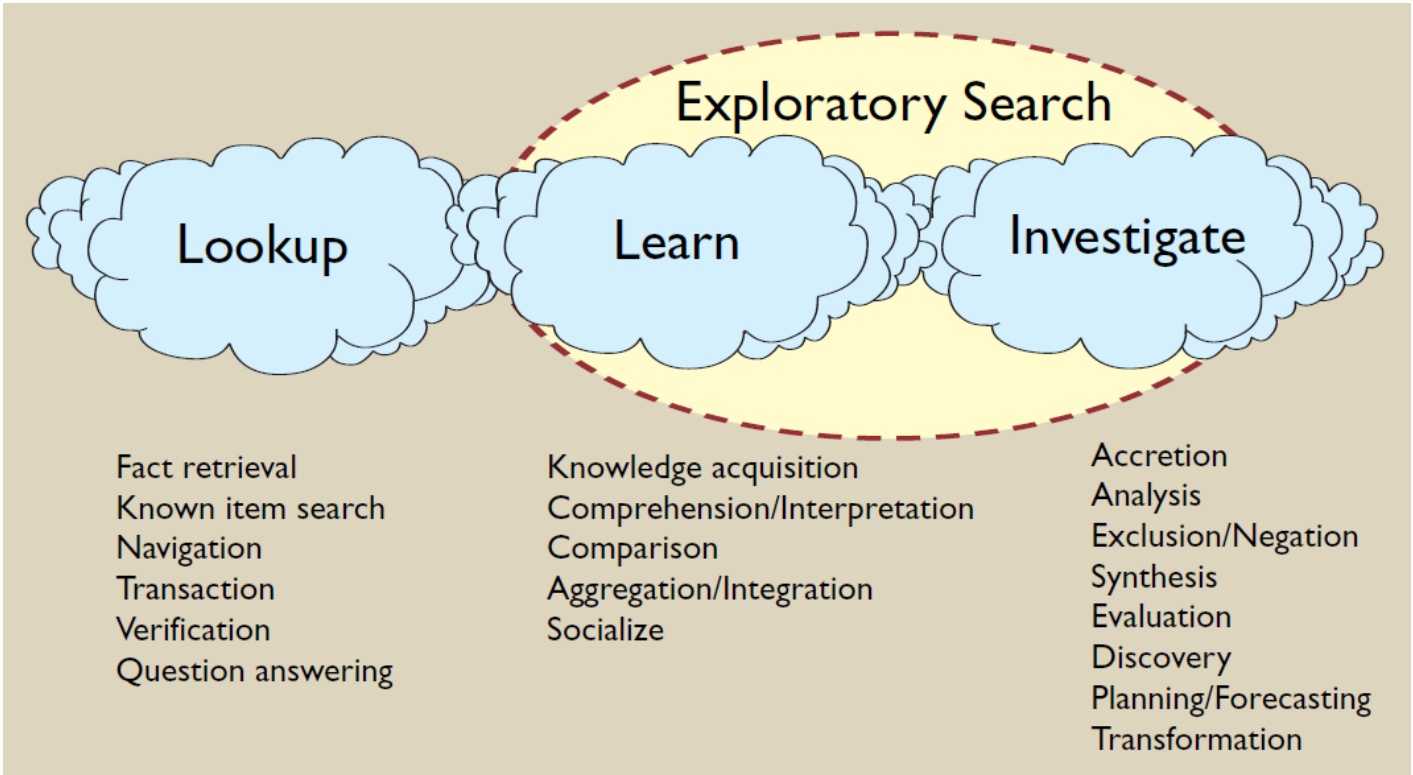
(евреи): израиль, израильский, страна, война, нетаньяху, тель-авив, время, сша, сирия, египет, случай, самолет, еврейский, военный, ближний,
(американцы): американский, американка, война, россия, военный, страна, вашингтон, америка, армия, конгресс, сирия, союзный, российский, обама, войска, русский, 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,

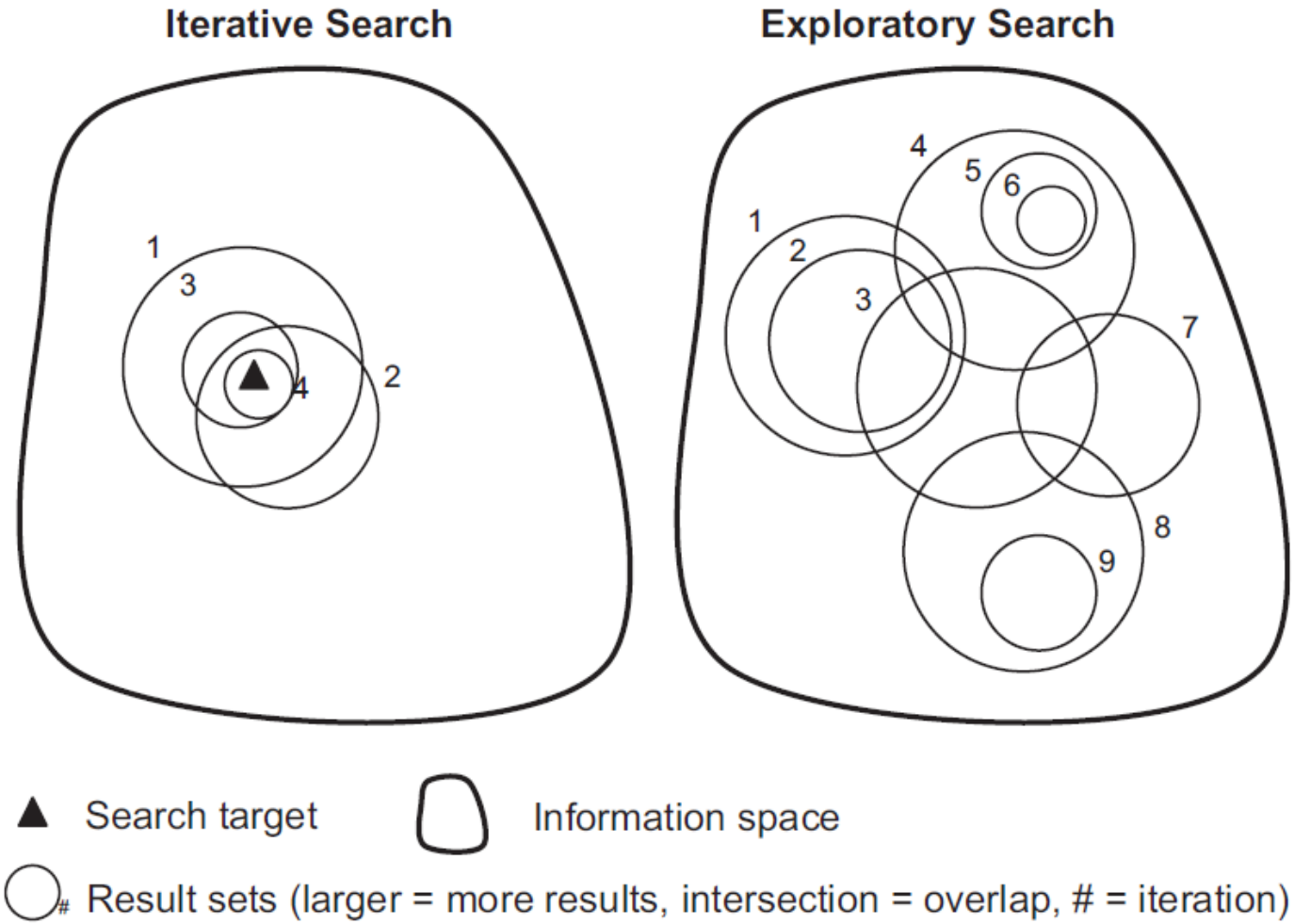
Разведочный поиск — знания «на кончиках пальцев»

- пользователь может не знать ключевых терминов
- пользователя может интересовать множество ответов



Gary Marchionini. Exploratory Search: from finding to understanding. Communications of the ACM. 2006, 49(4), p. 41–46.

От поиска «query-browse-refine» к разведочному поиску



R.W.White, R.A.Roth. Exploratory Search: beyond the Query-Response paradigm. San Rafael, CA: Morgan and Claypool, 2009.

Возможный сценарий разведочного поиска

Поисковый запрос:

- документ любой длины или даже коллекция документов

Цели поиска:

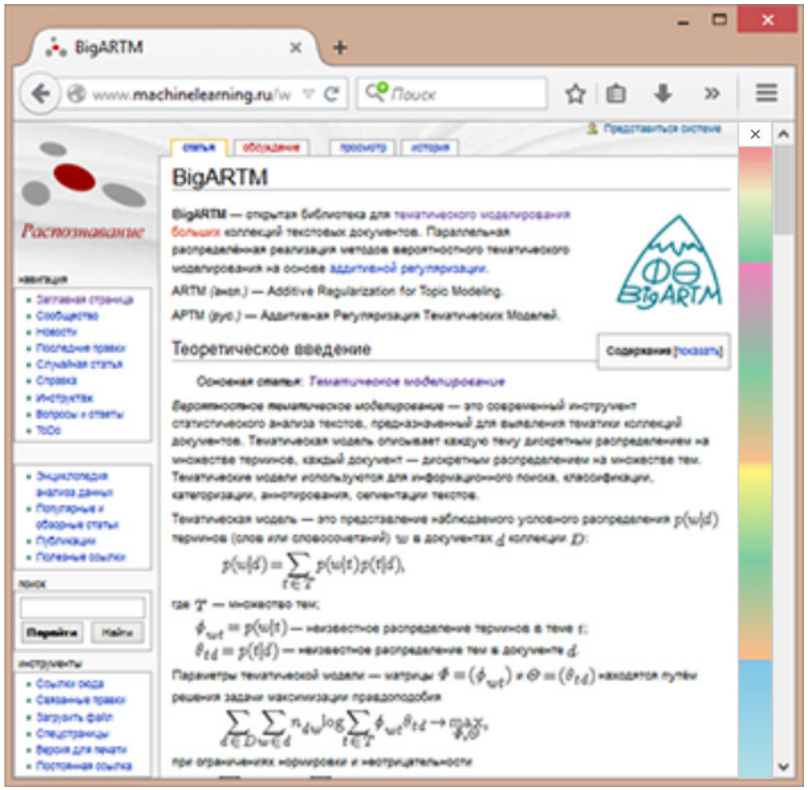
- к каким темам относится мой запрос?
- что ещё известно по этим темам?
- какова тематическая структура этой предметной области?
- что ещё есть понятного, обзорного, важного, свежего?

Сценарий поиска:

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

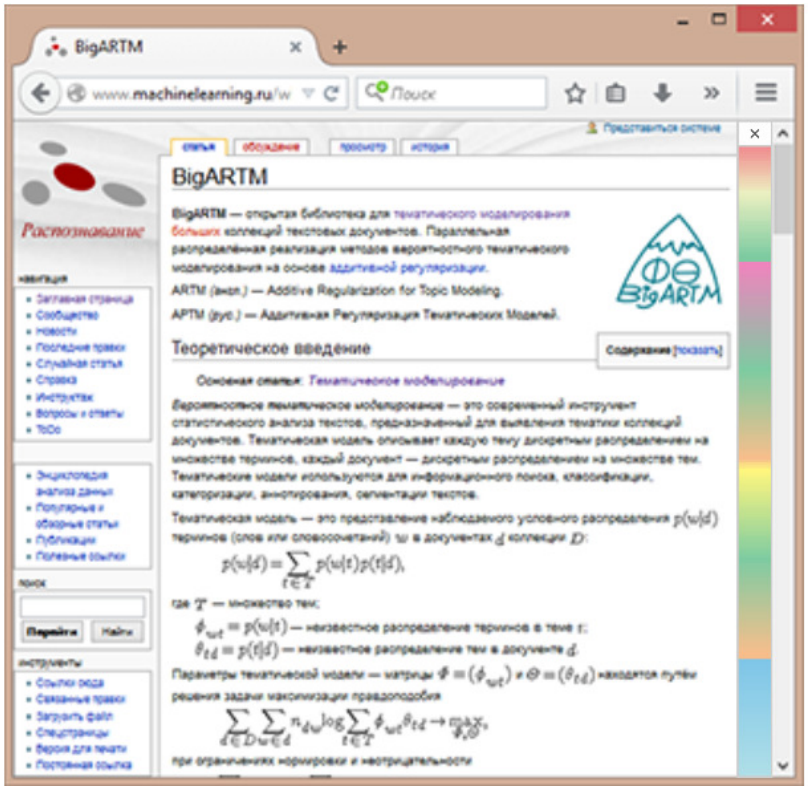
Разведочный поиск: прототип интерфейса

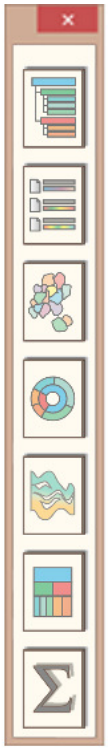
Радужная полоса напоминает, что знания всегда под рукой



Разведочный поиск: прототип интерфейса

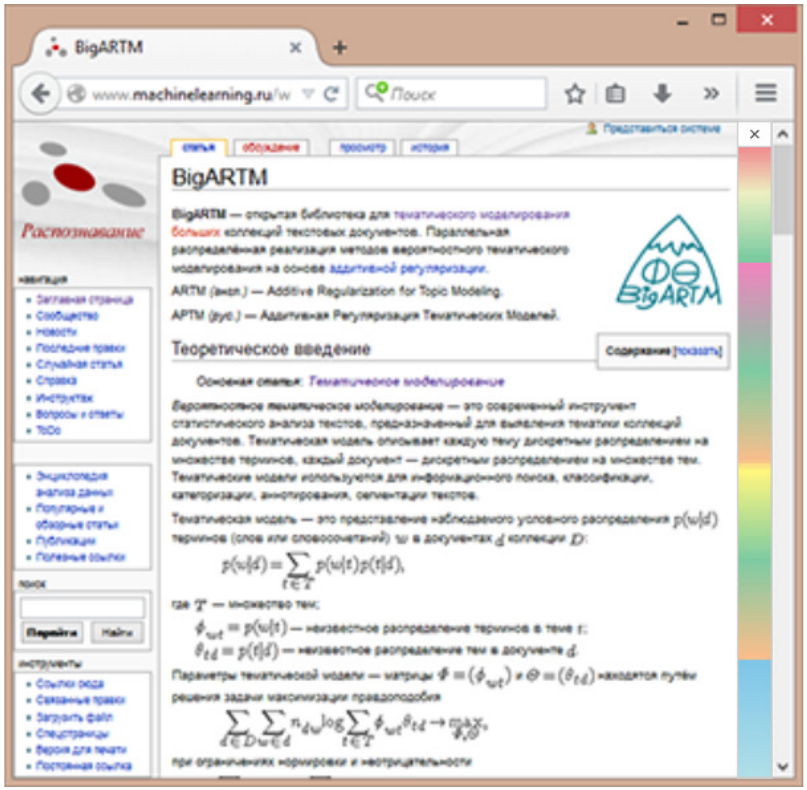
Клик по радужной полосе — тематический поисковый запрос

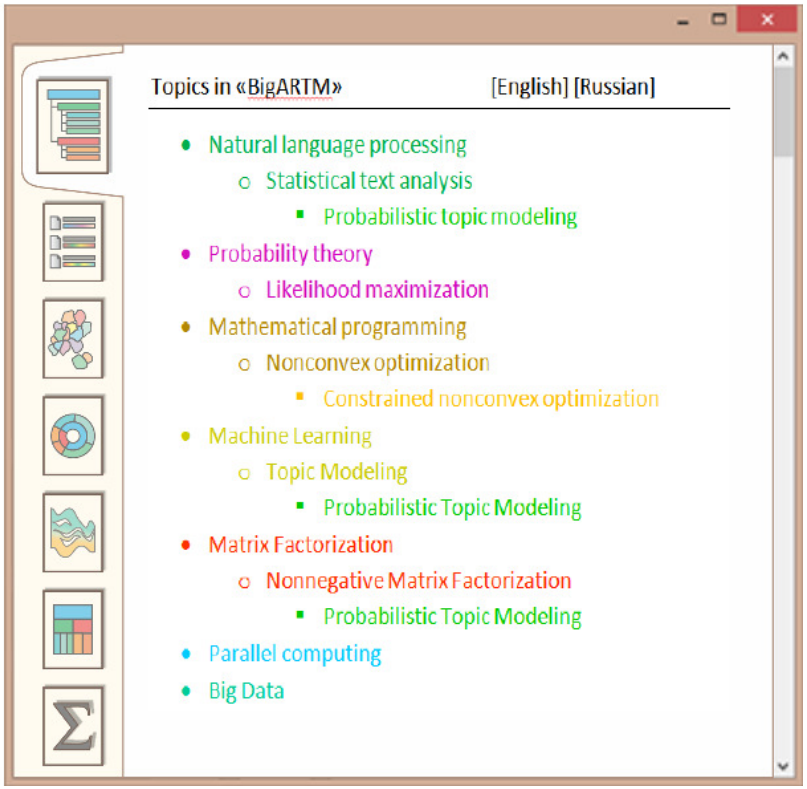




Разведочный поиск: прототип интерфейса

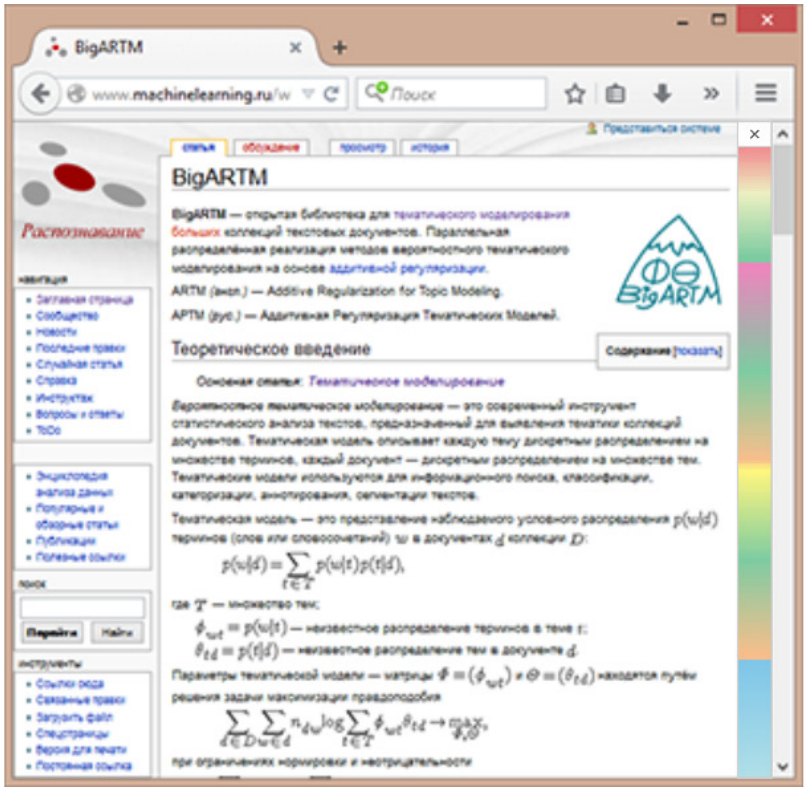
Темы-подтемы выбранного фрагмента текста

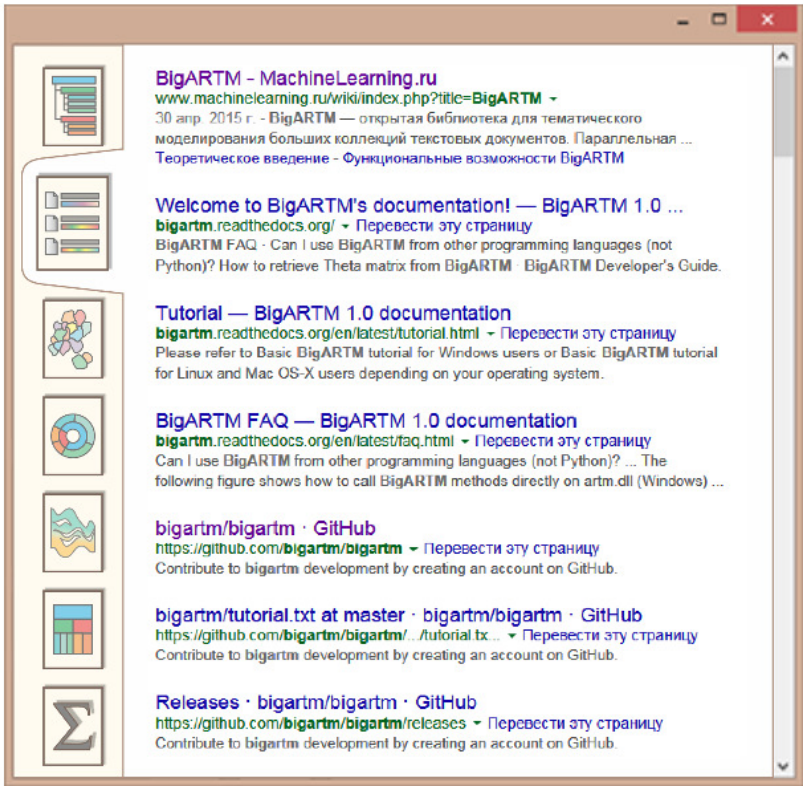




Разведочный поиск: прототип интерфейса

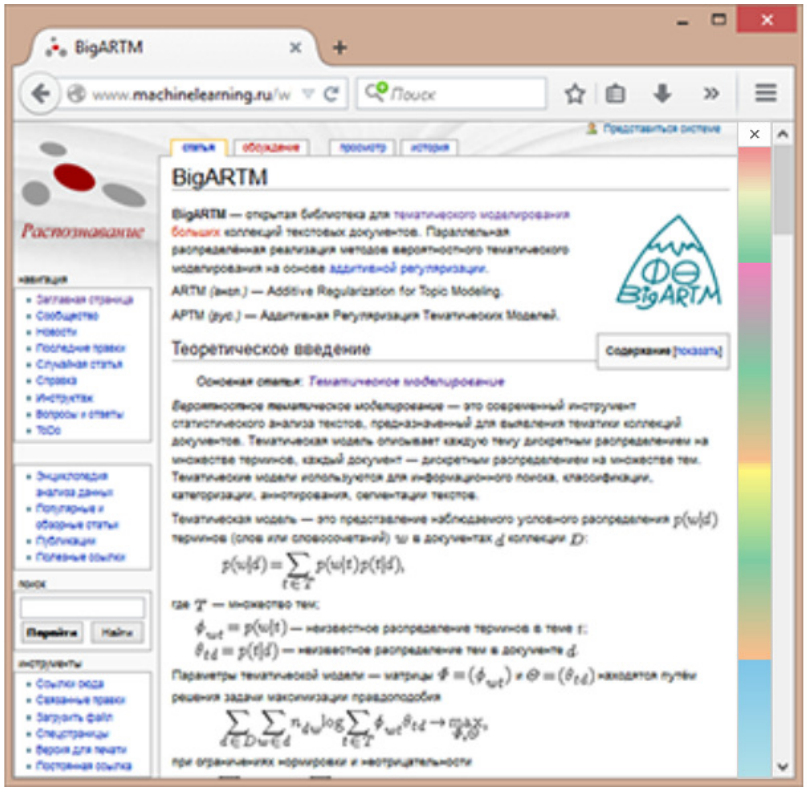
Документы и иные объекты, ранжированные по релевантности

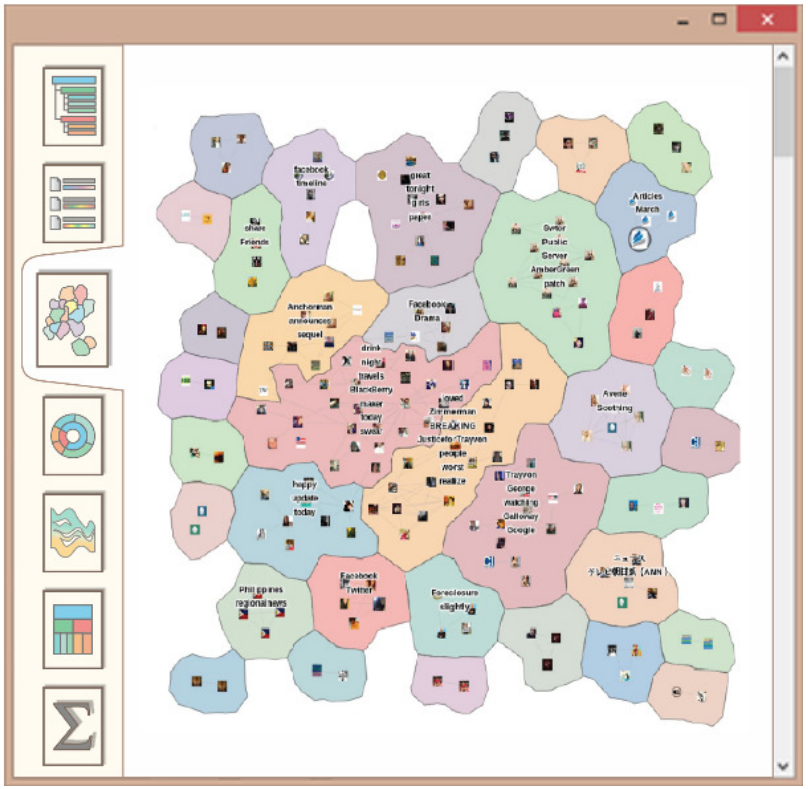




Разведочный поиск: прототип интерфейса

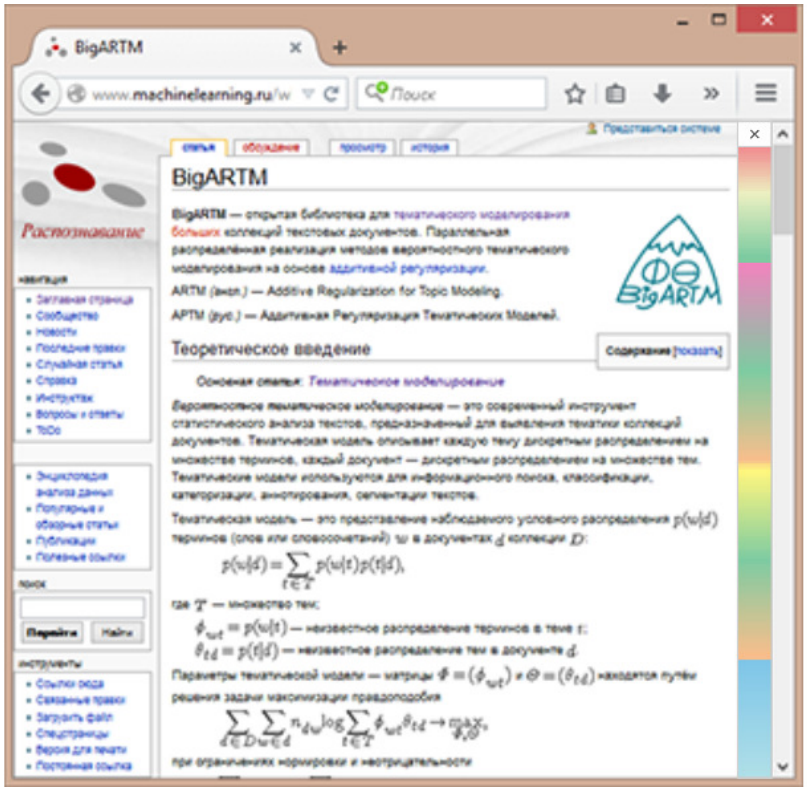
Дорожная карта: кластеризация релевантных документов

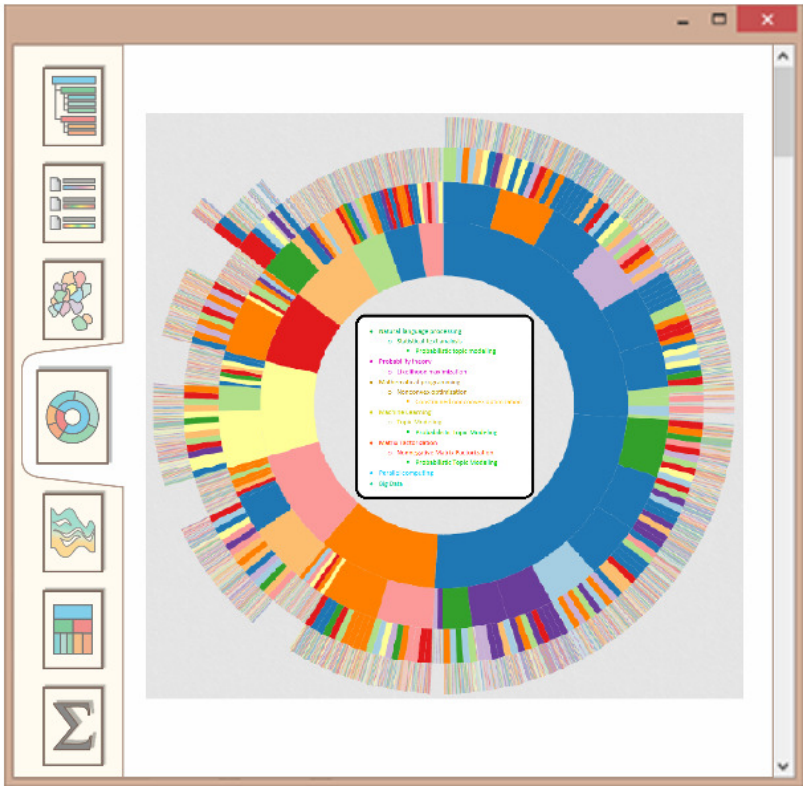




Разведочный поиск: прототип интерфейса

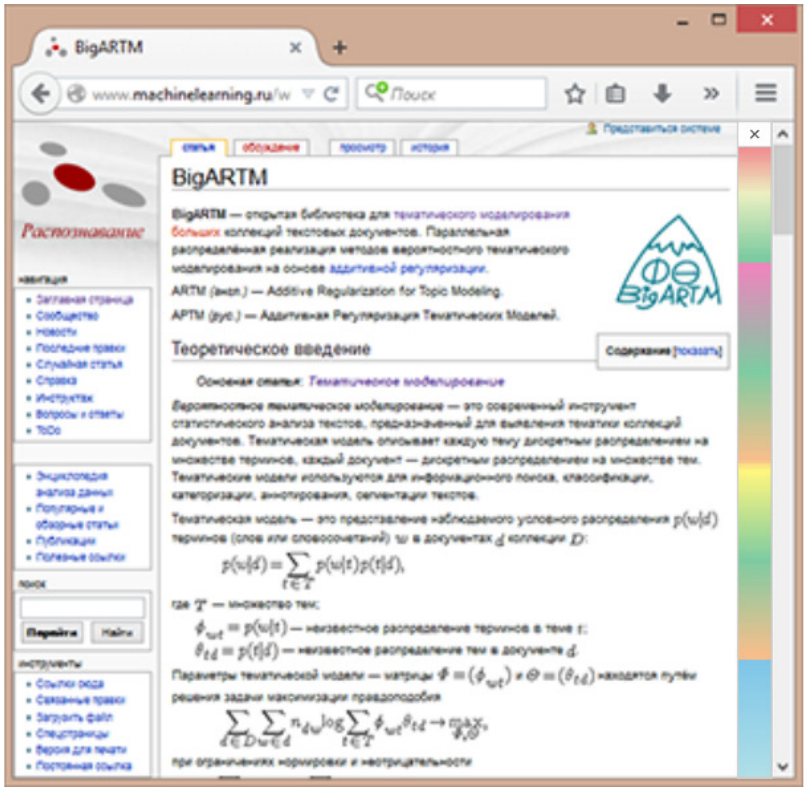
Тематическая иерархия: структура предметной области

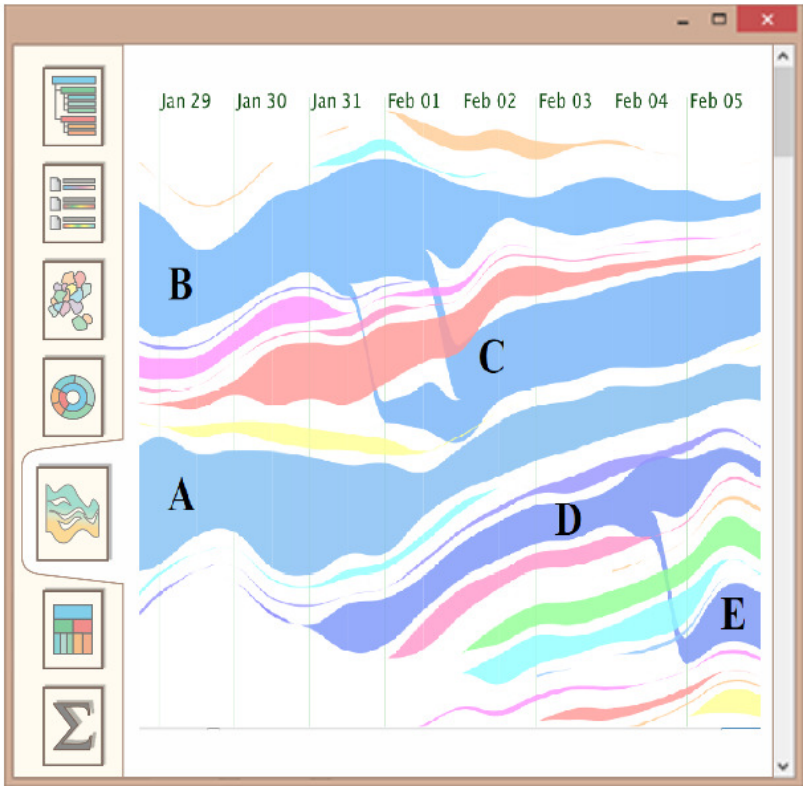




Разведочный поиск: прототип интерфейса

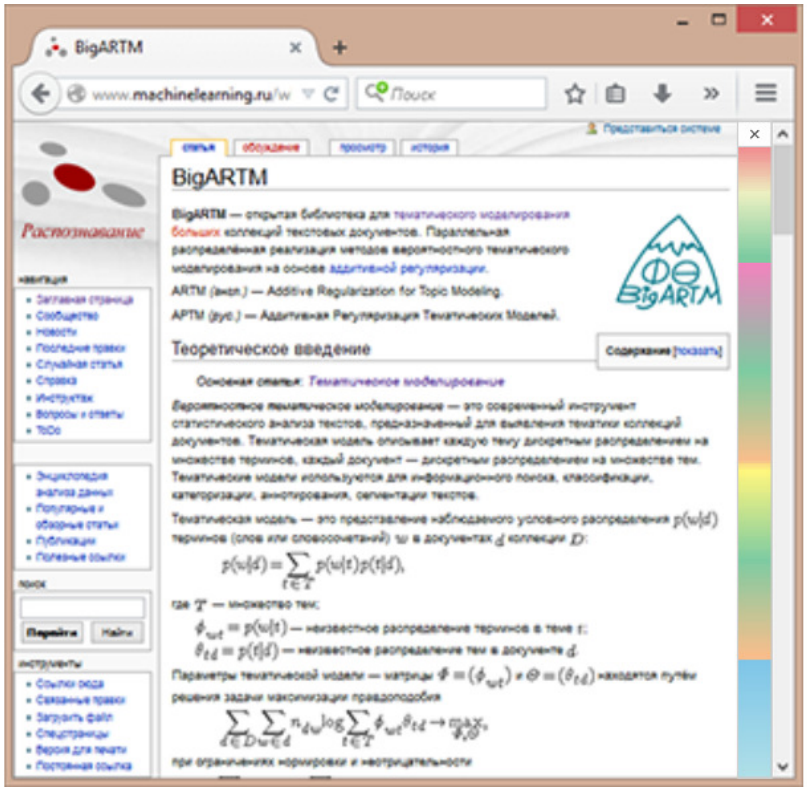
Динамика тем: эволюция предметной области

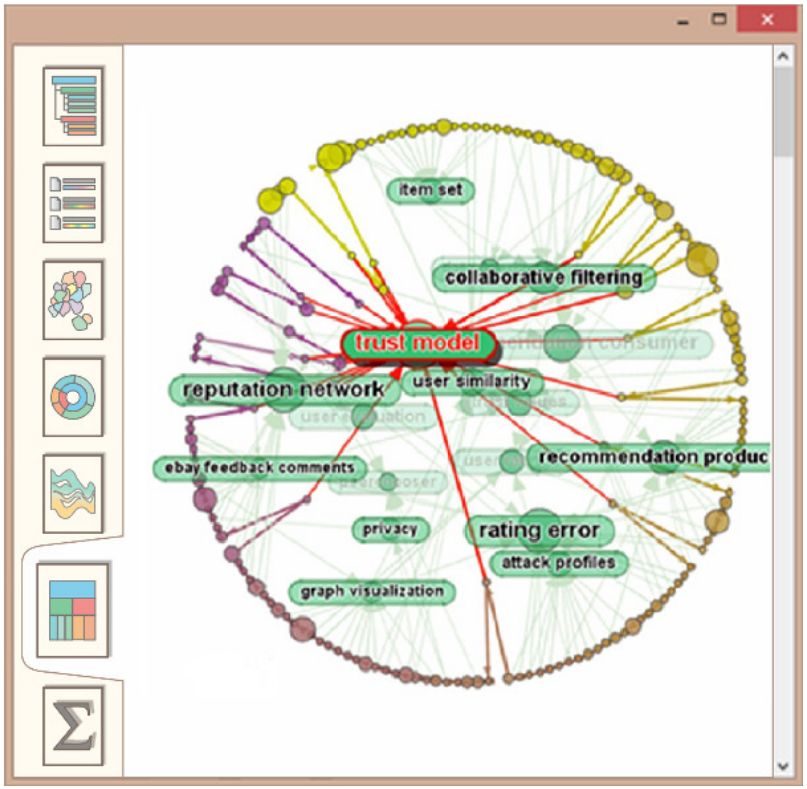




Разведочный поиск: прототип интерфейса

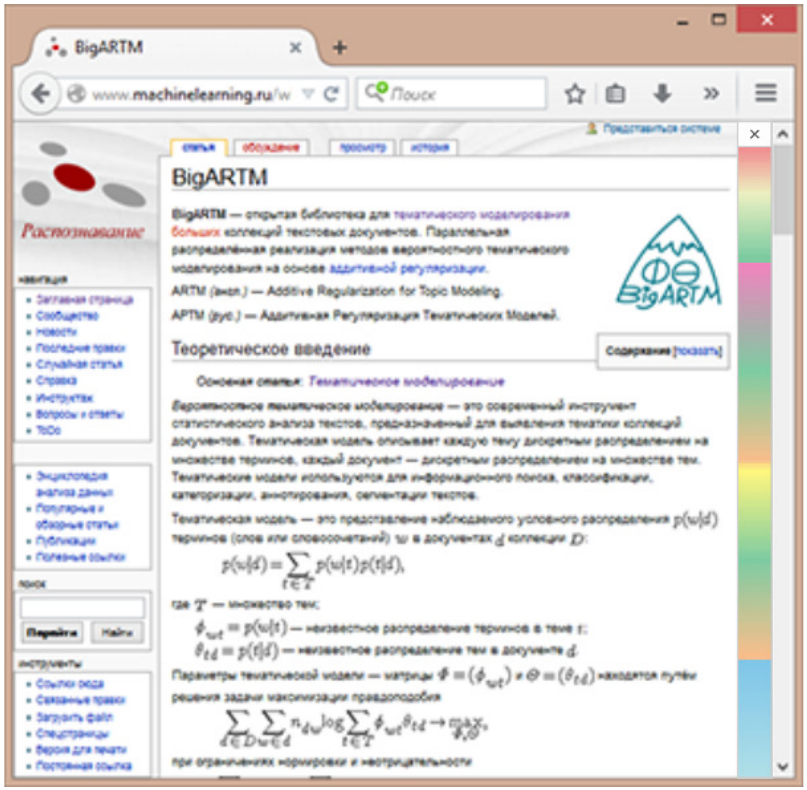
Тематическая сегментация документа запроса

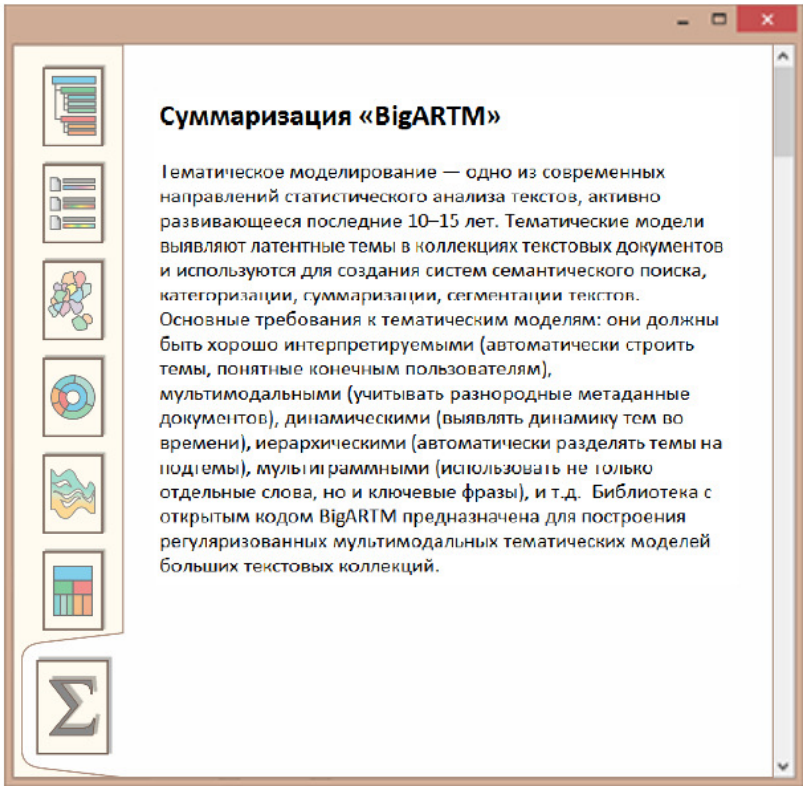




Разведочный поиск: прототип интерфейса

Суммаризация документа запроса





http://textvis.lnu.se

Интерактивный обзор 272 средств визуализации текстов

Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // Машинное обучение и анализ данных (http://jmlda.org). 2015. Т. 1, № 11. С. 1584-1618.

Технологические элементы разведочного поиска

1. Интернет-краулинг ..... имеются готовые решения
2. Фильтрация контента ..... имеются готовые решения
3. Тематическое моделирование ..... **математика здесь**
4. Инвертированный индекс ..... имеются готовые решения
5. Ранжирование ..... имеются готовые решения
6. Визуализация ..... имеются готовые решения

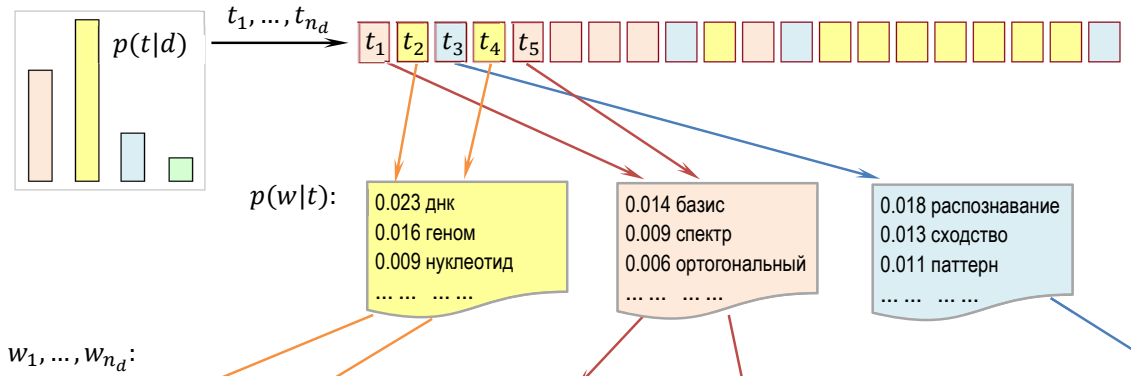
Тематическая модель для разведочного поиска должна быть...

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

Прямая задача — порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D описывает появление терминов w в документах d темами t:

$$p(w|d) = \sum_t p(w|t)p(t|d), \quad d \in D$$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

Дано: W — словарь терминов (слов или словосочетаний),
D — коллекция текстовых документов $d \subset W$,
n_{dw} — сколько раз термин w встретился в документе d.

Найти: модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$ с параметрами $\underset{W \times T}{\Phi}$ и $\underset{T \times D}{\Theta}$:
$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t,
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d.

Критерий максимума логарифма правдоподобия:

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

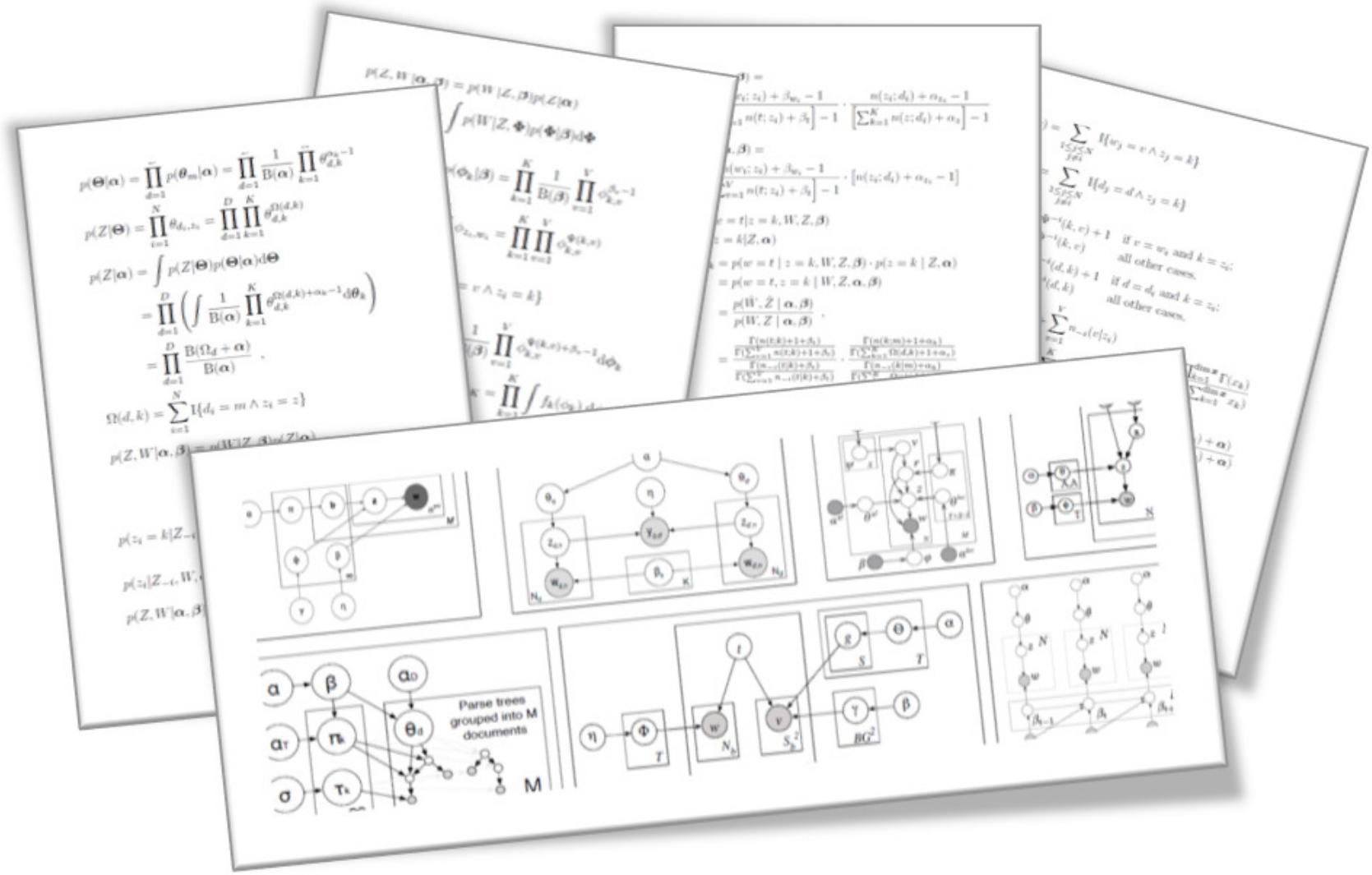
$$\phi_{wt} \geqslant 0; \quad \textstyle\sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \textstyle\sum_t \theta_{td} = 1.$$

Проблема: задача стохастического матричного разложения *некорректно поставлена*: $\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$.

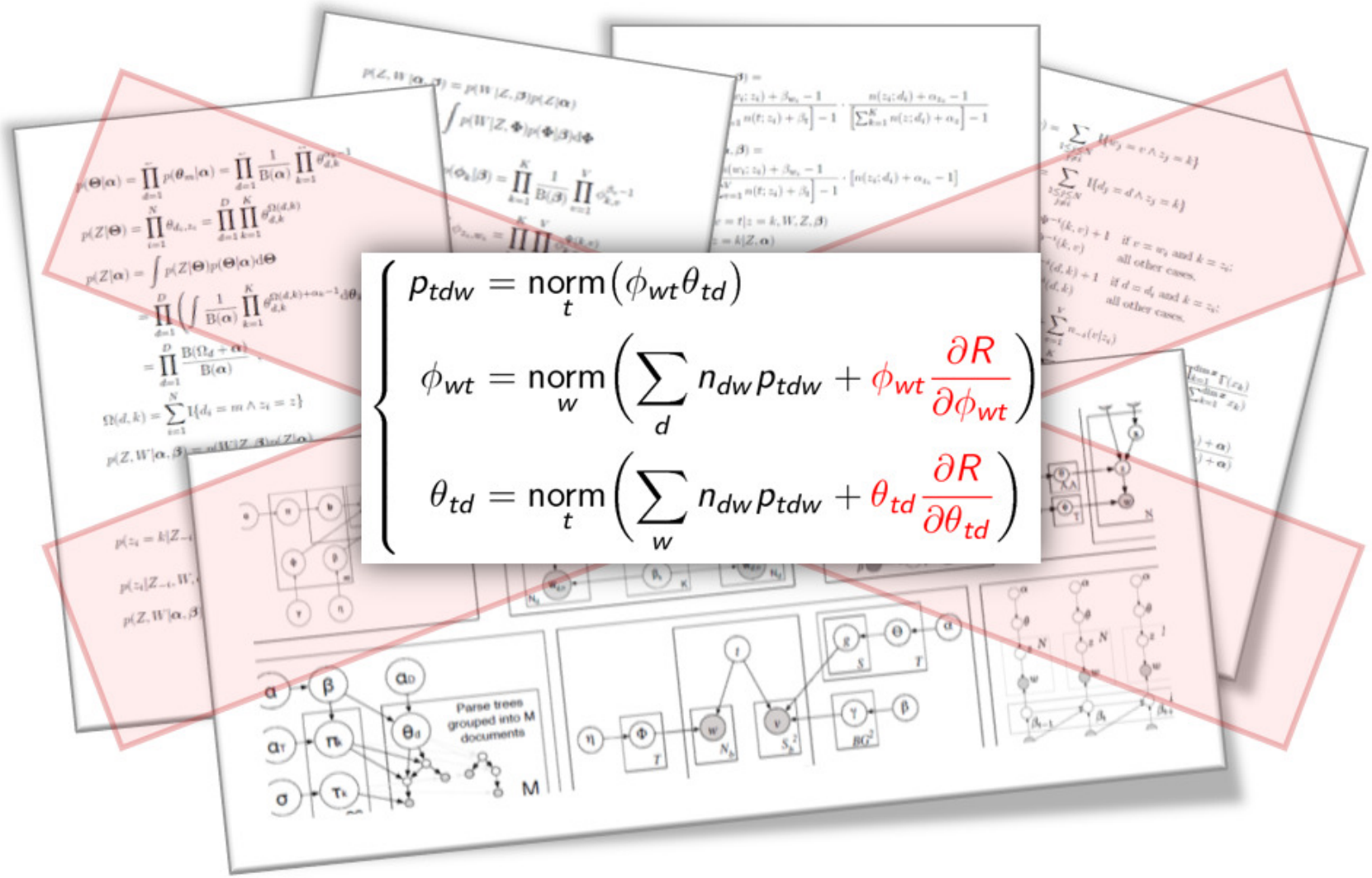
Кризис байесовского обучения в тематическом моделировании

- сотни тематических моделей, начиная с LDA (Blei, 2003),
- создаются скорее ради теории, а не ради приложений,
- часто не имеют достаточных лингвистических обоснований,
- опираются на избыточные вероятностные допущения и
- уникальные математические выкладки для каждой модели,
- слишком сложны для понимания, вывода, сравнения,
- имеют проблемы неединственности и неустойчивости,
- не комбинируются и взаимно не заменяются,
- не имеют полнофункциональных библиотек в открытом коде,
- что создаёт барьеры вхождения для прикладников,
- которые предпочитают устаревшие но понятные PLSA и LDA

Байесовское обучение — доминирующий подход в BTM



ARTM — альтернатива байесовскому обучению



ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация ln правдоподобия с регуляризатором R:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

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

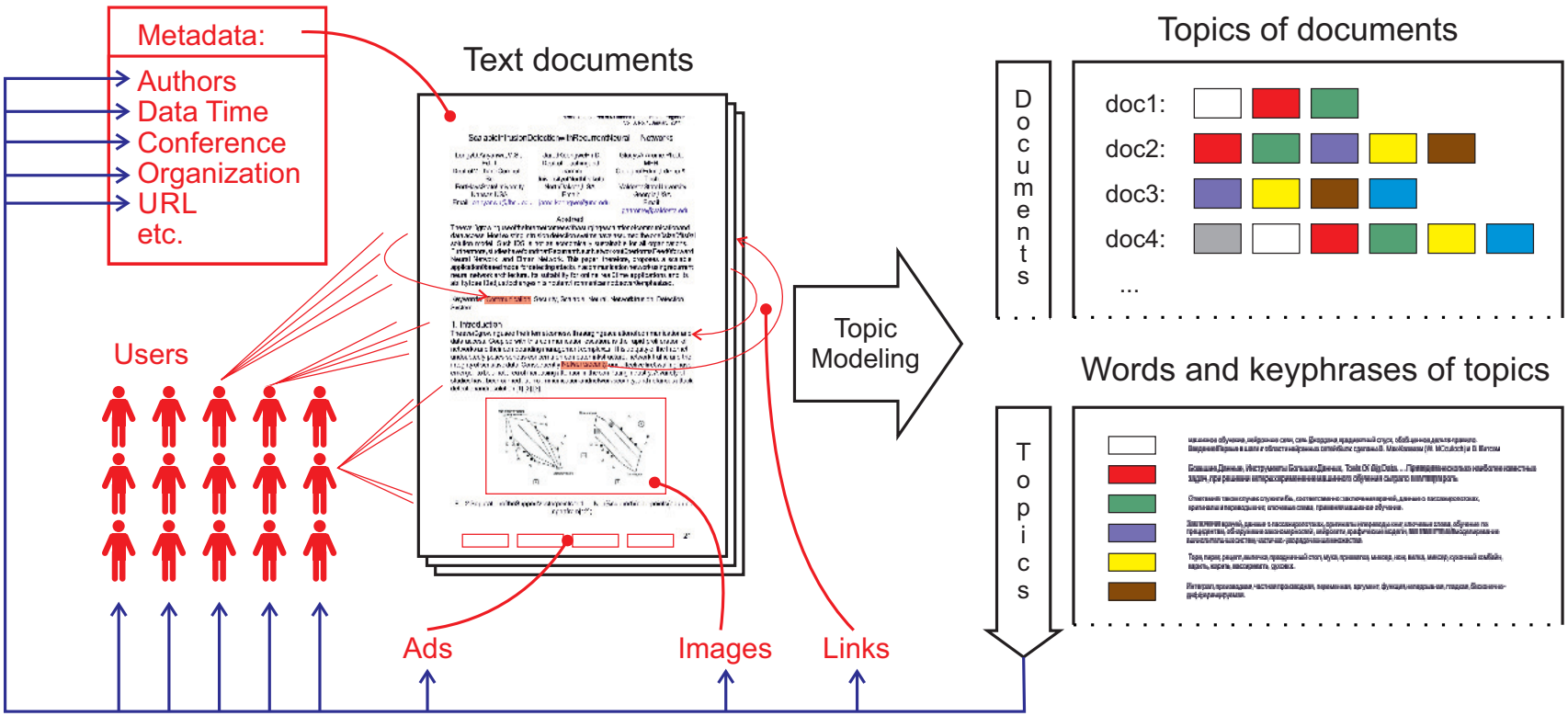
$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), & n_{wt} = \sum_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right),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

Модель PLSA: $R(\Phi,\Theta) = 0$

Модель LDA: $R(\Phi,\Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}$

ARTM легко обобщается на мультимодальные задачи

Выявление тематики документов $p(t|d)$, терминов $p(t|w)$,
и различных *модальностей*: $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{ссылка})$,
$p(t|\text{баннер})$, $p(t|\text{элемент изображения})$, $p(t|\text{пользователь})$,...



Преимущества ARTM

- Отказ от избыточных вероятностных допущений
- Общие формулы для любых регуляризаторов и модальностей
- Простота формализации, вывода, понимания моделей
- Комбинирование моделей путём сложения регуляризаторов
- Быстрый однопроходный онлайновый EM-алгоритм
- Библиотека BigARTM с открытым кодом

Воронцов К.В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014. Т. 455., № 3. 268–271.

Vorontsov K.V., Potapenko A.A. Additive Regularization of Topic Models // Machine Learning. Springer, 2015. Volume 101, Issue 1-3. Pp. 303–323.

Vorontsov K.V., Frei O.I., Apishev M.A., Romov P.A., Suvorova M.A., Yanina A.O. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections // Workshop on Topic Models, 2015, Melbourne, Australia. Pp. 29–37.

Примеры регуляризаторов (сглаживание и разреживание)

1. разреживание предметных тем $S \subset T$:

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

2. сглаживание фоновых тем $B \subset T$:

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

3. частичное обучение по подмножествам $W_t \subset W$, $D_t \subset D$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi_{wt} + \alpha_0 \sum_{t \in T_0} \sum_{d \in D_t} \theta_{td} \to \max$$

Примеры регуляризаторов (корреляции и декорреляции)

4. декоррелирование тем как столбцов Φ:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in S} \sum_{s \in S \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max$$

5. максимизация когерентности тем:

$$R(\Phi) = \tau \sum_{t \in T} \sum_{u,w \in W} C_{uw} n_{ut} \ln \phi_{wt} \to \max$$

6. учёт связей между документами $n_{dd'}$:

$$R(\Theta) = \tau \sum_{d,d'} n_{dd'} \sum_{t \in T} \theta_{td} \theta_{td'} \to \max$$

7. учёт корреляций между темами как строками Θ:

$$R(\Theta) = -\frac{\tau}{2} \sum_{d \in D} (\ln \theta_d - \mu)^{\mathrm{T}} \Sigma^{-1} (\ln \theta_d - \mu) \to \max$$

Примеры регуляризаторов (определение числа тем)

8. удаление неинформативных тем:

$$R(\Theta) = -\tau \sum_{t \in S} \ln p(t) \to \max, \qquad p(t) = \sum_{d \in D} \theta_{td} p(d)$$

9. разреживание тем во времени:

$$R(\Theta) = -\tau \sum_{y \in Y} \sum_{t \in T} \ln p(t|y) \to \max, \quad p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$$

10. сглаживание тем во времени:

$$R(\Theta) = -\tau \sum_{y \in Y} \sum_{t \in T} \left| p(y|t) - p(y-1|t) \right| \to \max$$

BigARTM: библиотека тематического моделирования

Ключевые возможности:

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

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: унификация разработки тематических моделей

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*

Разработка тематических моделей в среде IPython Notebook

http://nbviewer.ipython.org/github/bigartm/bigartm-book/tree/master/

Коллекция:

Воспользуемся небольшой коллекцией 'kos', доступной в репозитории UCI https://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/. Параметры коллекции следующие:

- 3430 документов;
- 6906 слов в словаре;
- 467714 слов в коллекции.

Для начала подключим все необходимые модули (убедитесь, что путь к Python API BigARTM находится в вашей переменной PATH):

In [1]:
```
%matplotlib inline
import glob
import matplotlib.pyplot as plt

import artm
```

Прежде всего необходимо подготовить входные данные. BigARTM имеет собственный формат документов для обработки, называемый батчами. В библиотеки присутствуют средства по созданию батчей из файлов Bag-Of-Words в форматах UCI и Vowpal Wabbit (подробности можно найти в http://docs.bigartm.org/en/latest/formats.html).

В Python API, по аналогии с алгоритмами из scikit-learn, входные данные представлены одним классом BatchVectorizer. Объект этого класса принимает на вход батчи или файлы с Bag-Of-Words и подаётся на вход всем методам. В случае, если входные данные не являются батчами, он создаёт их и сохраняет на диск для последующего быстрого использования.

Итак, создадим объект BatchVectorizer:

In [2]:
```
batch_vectorizer = None
if len(glob.glob('kos' + "/*.batch")) < 1:
    batch_vectorizer = artm.BatchVectorizer(data_path='', data_format='bow
_uci', collection_name='kos', target_folder='kos')
else:
    batch_vectorizer = artm.BatchVectorizer(data_path='kos', data_format='
batches')
```

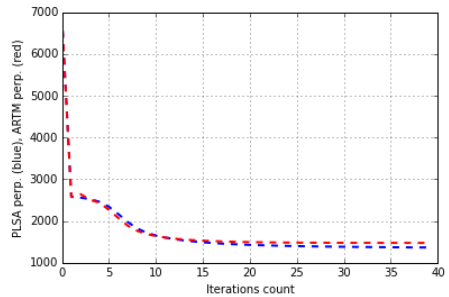
ARTM --- это класс, представляющий собой Python API BigARTM, и позволяющий использовать практически все возможности библиотеки в стиле scikit-learn. Создадим две тематические модели для нашего эксперимента. Наиболее важным параметром модели является число тем. Опционально можно указать списки регуляризаторов и функционалов качества, которые следует использовать для данной модели. Если этого не сделать, то регуляризаторы и функционалы всегда можно добавить позднее. Обратите внимание, что каждая модель задаёт

Продолжим обучение моделей, инициировав 25 проходов по коллекции, после чего снова посмотрим на значения функционалов качества:

In [11]:
```
model_plsa.fit_offline(batch_vectorizer=batch_vectorizer, num_collection_p
asses=25, num_document_passes=1)
model_artm.fit_offline(batch_vectorizer=batch_vectorizer, num_collection_p
asses=25, num_document_passes=1)
```

In [12]:
```
print_measures(model_plsa, model_artm)
```

```
Sparsity Phi: 0.332 (PLSA) vs. 0.740 (ARTM)
Sparsity Theta: 0.082 (PLSA) vs. 0.602 (ARTM)
Kernel contrast: 0.530 (PLSA) vs. 0.568 (ARTM)
Kernel purity: 0.396 (PLSA) vs. 0.531 (ARTM)
Perplexity: 1365.804 (PLSA) vs. 1475.455 (ARTM)
```



Кроме того, для наглядности построим графики изменения разреженностей матриц по итерациям:

In [13]:
```
plt.plot(xrange(model_plsa.num_phi_updates), model_plsa.score_tracker['Spa
rsityPhiScore'].value, 'b--',
                 xrange(model_artm.num_phi_updates), model_artm.score_trac
ker['SparsityPhiScore'].value, 'r--', linewidth=2)
plt.xlabel('Iterations count')
plt.ylabel('PLSA Phi sp. (blue), ARTM Phi sp. (red)')
plt.grid(True)
plt.show()

plt.plot(xrange(model_plsa.num_phi_updates), model_plsa.score_tracker['Spa
rsityThetaScore'].value, 'b--',
                 xrange(model_artm.num_phi_updates), model_artm.score_trac
ker['SparsityThetaScore'].value, 'r--', linewidth=2)
```

Обгоняем конкурентов по скорости

- 3.7M статей английской Вики, 100K уникальных слов

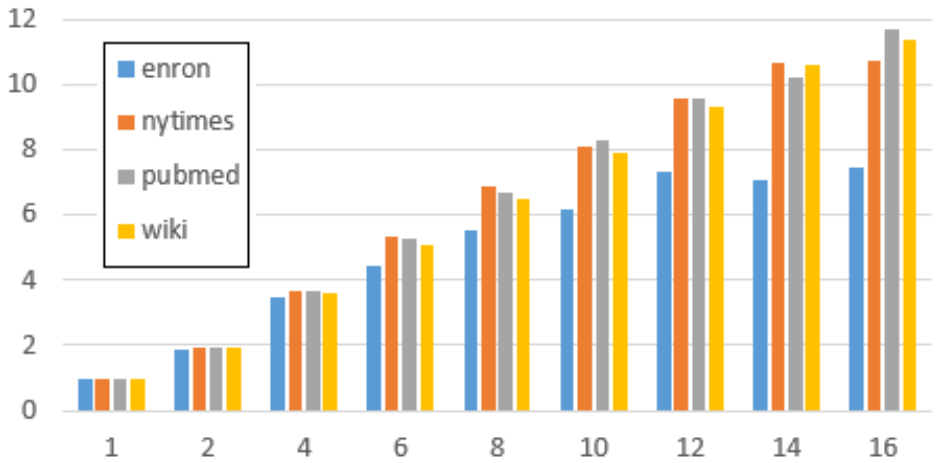
	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

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

Масштабируемость по числу потоков

коллекция	$\|W\|$, 10^3	$\|D\|$, 10^6	n, 10^6	размер, Гб
enron	28	0.04	6.4	0.07
nytimes	103	0.3	100	0.13
pubmed	141	8.2	738	1.0
wiki	100	3.7	1009	1.2

ускорение



процессоров

Amazon EC2 cc2.8xlarge instance:
16 cores + hyperthreading, Intel® Xeon® CPU E5-2670 2.6GHz.

Приложения

- Разведочный поиск на habrahabr.ru и его оценивание
- Кросс-язычный разведочный поиск arXiv.org+Википедия
- Тематизация коллекции научных статей ММРО/ИОИ
- Тематизация текстов и изображений из соцсетей
- Тематизация картин британского музея и их описаний
- Классификация авторефератов по областям знаний
- Конкурс Kaggle «The Allen AI Science Challenge»
- Информационный анализ электрокардиосигналов
- Поиск мотивов в задачах биоинформатики

Направления текущих исследований

- обеспечивать полноту, устойчивость, интерпретируемость
- объединить ARTM с моделями дистрибутивной семантики
- разделять текст на тематически однородные сегменты
- формировать суммаризацию темы
- автоматически создавать и именовать темы
- строить мелкозернистую тематическую иерархию
- строить 50 тысяч хорошо интерпретируемых тем
- применять гиперграфовые модели к данным соцсетей
- адаптивно управлять траекторией регуляризации
- разработать визуальные средства систематизации знаний
- создать систему тематического разведочного поиска

http://bigartm.org