Математические методы и прикладные задачи анализа текстов

Воронцов Константин Вячеславович

• лаборатория Машинного интеллекта МФТИ •
• кафедра Интеллектуальных систем ФУПМ МФТИ •

Долгопрудный, МФТИ • 29 января 2018

Содержание

Пирамида NLP (Natural Language Processing)

Задачи анализа текстов

- Морфологический анализ и лемматизация (lemmatization)
- Синтаксический анализ (syntax analysis)

...

Задачи анализа текстов

- Автоматическое выделение терминов (automatic term extraction)
- Распознавание именованных сущностей (named entity recognition)
- Семантические сети и онтологии (ontology learning)
- Семантические векторные представления слов (word embedding)
- Тематическое моделирование (topic modeling)
- Кластеризация текстов (text clustering)
- Классификация текстов (text classification)
- Сегментация текста (text segmentation)
- Аннотирование и суммаризация (text summarization)
- Информационный поиск (information retrieval)
- Обучаемое ранжирование (learning to rank)
- Ответы на вопросы, машинный перевод, чат-боты (sequence-to-sequence)

Прикладные задачи анализа текстов

- Новостной мониторинг: выявление и классификация новостей, относящихся к данной компании
- Анализ новостей или твитов для предсказания скачков цены на биржевых торгах
- Анализ текстов отзывов о компании и её продуктах для выявления и категоризации проблем
- Анализ текстов комментариев в опросах сотрудников крупных компаний для выявления и категоризации проблем
- Анализ записей голосовых обращений в контакт-центр для классификации целей клиента (автосалоны, недвижимость, запись к врачу, заказ такси)
- Сегментация записей разговоров контакт-центра для мониторинга работы операторов и оптимизации скриптов
- Анализ данных о посещениях сайтов для выявления интересов пользователей и таргетирования рекламы

Другие прикладные исследования

- Анализ банковских транзакционных данных физических лиц для выявления типов потребления и прогнозирования расходов
- Анализ банковских транзакционных данных юридических лиц для выявления видов экономической деятельности по отраслям
- Распознавание движений животного или человека по данным электрокортикограммы
- Распознавание действий рабочих по данным с носимых мобильных устройств

---

www.MachineLearning.ru: Численные методы обучения по прецедентам (практика, В.В. Стрижов)

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

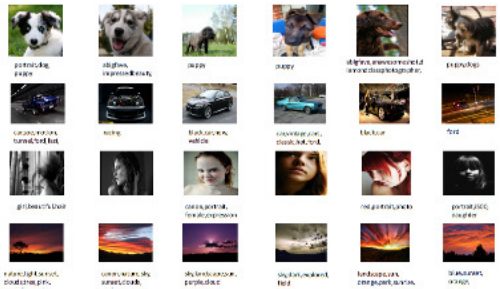
Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели // Магистерская диссертация, МФТИ, 2015.

Приложения тематического моделирования

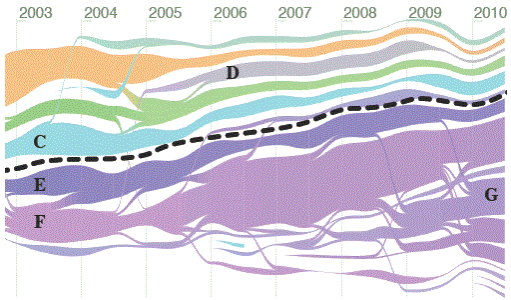
разведочный поиск в электронных библиотеках

персонализированный поиск в соцсетях

мультимодальный поиск текстов и изображений

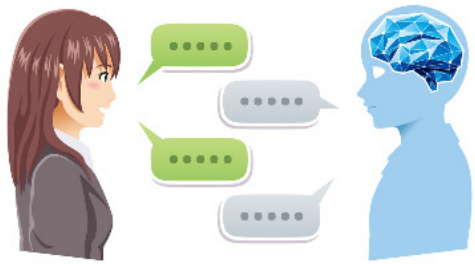


детектирование и трекинг новостных сюжетов



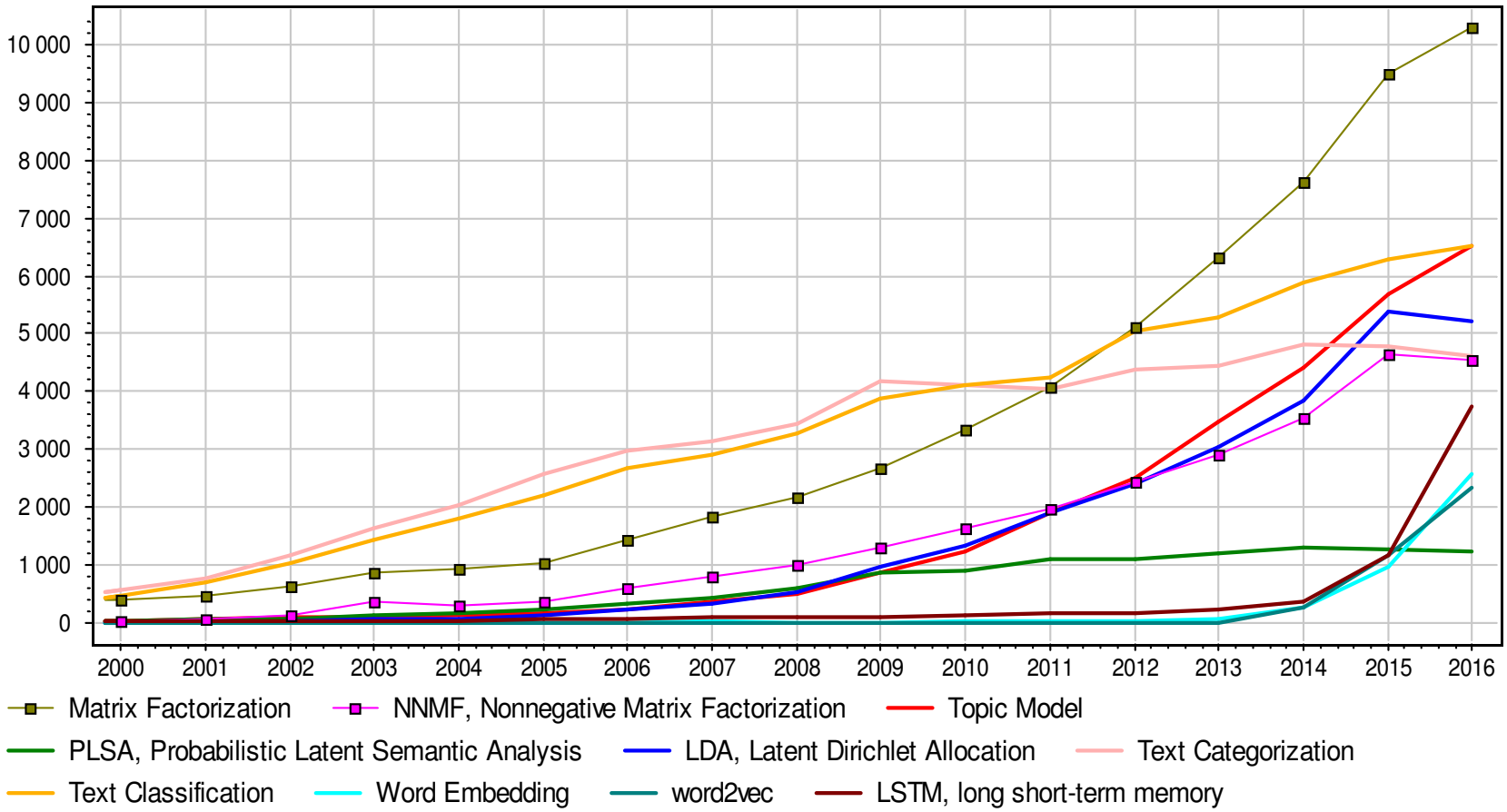
навигация по большим текстовым коллекциям

управлением диалогом в разговорном интеллекте



Тематическое моделирование и смежные области исследований

Динамика цитирования, по данным Google Scholar:



Пусть

- W — конечное множество слов (терминов, токенов)
- D — конечное множество текстовых документов
- T — конечное множество тем
- каждое слово w в документе d связано с некоторой темой t
- $D \times W \times T$ — дискретное вероятностное пространство
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- коллекция — это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d,t) = p(w|t)$

Тематическая модель, по формуле полной вероятности:

$$p(w|d) = \sum_{t \in T} p(w \mid \cancel{d}, t)\, p(t|d)$$

Постановка задачи тематического моделирования

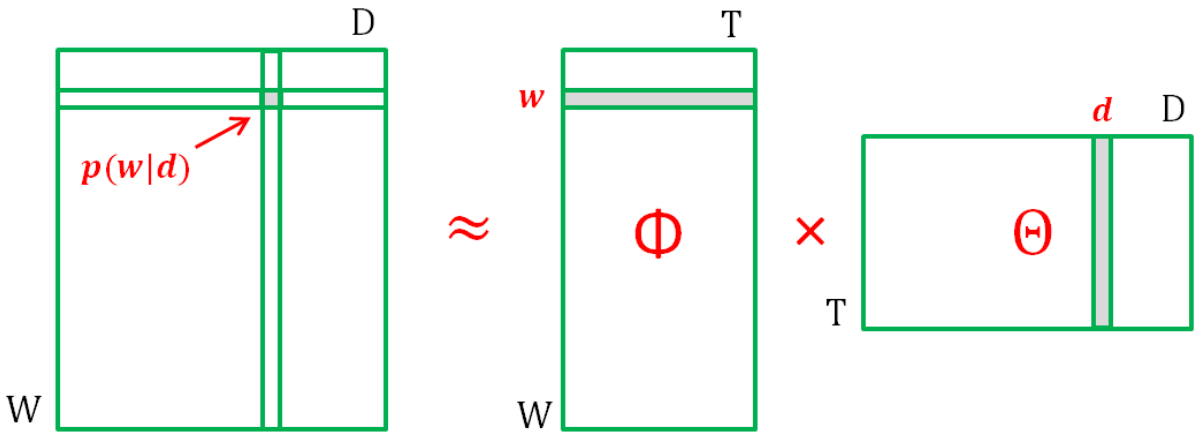
Дано: коллекция текстовых документов

- n_{dw} — частоты терминов в документах, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

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

- $\phi_{wt} = p(w|t)$ — вероятности 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

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

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)

Наша задача матричного разложения *некорректно поставлена*: если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta), \quad \operatorname{rank} S = |T|$
- $\mathscr{L}(\Phi', \Theta') = \mathscr{L}(\Phi, \Theta)$
- $\mathscr{L}(\Phi', \Theta') \leqslant \mathscr{L}(\Phi, \Theta) + \varepsilon$ — приближённые решения

Регуляризация — стандартный приём доопределения решения с помощью дополнительных критериев.

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

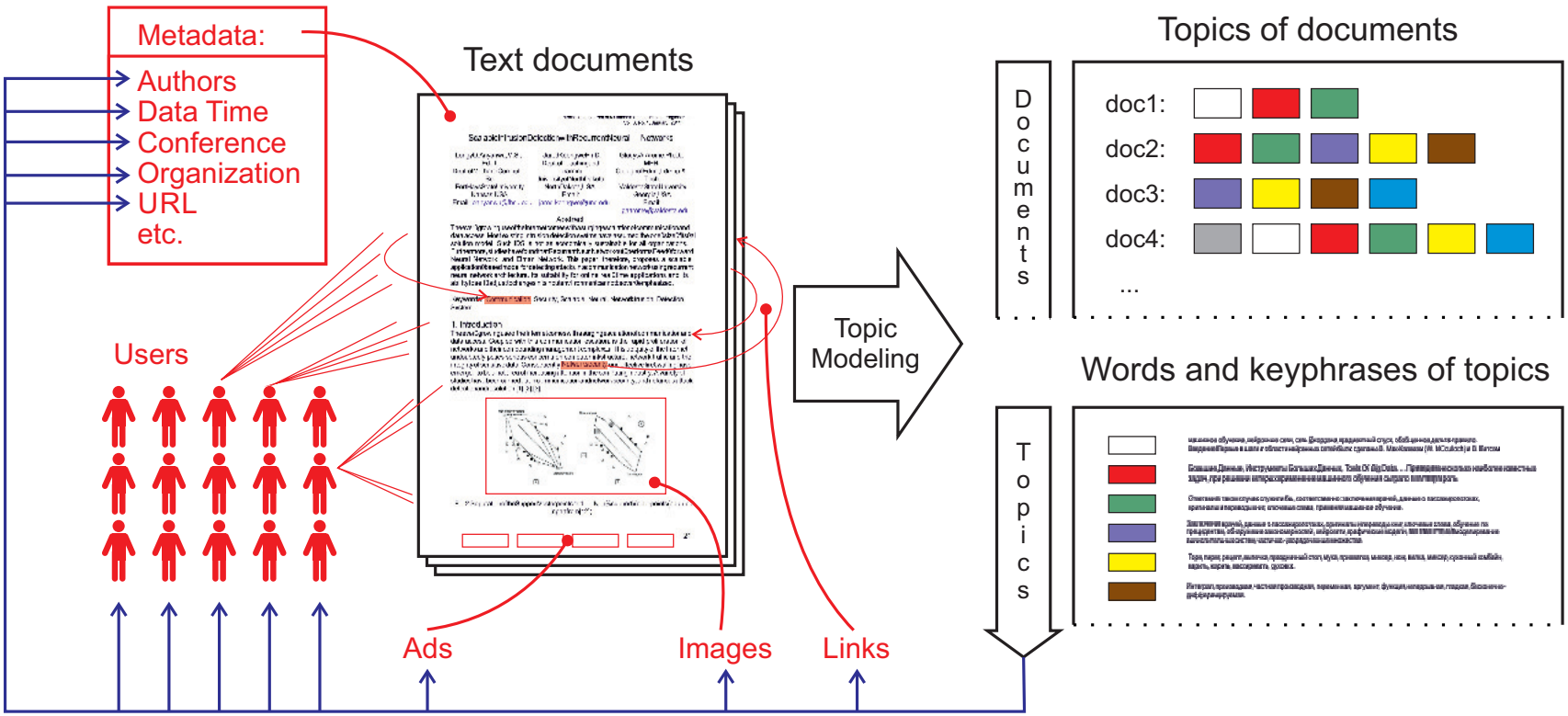
$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W}{\text{norm}}\left(\sum_{d \in D} n_{dw} p(t|d,w) + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} n_{dw} p(t|d,w) + \theta_{td} \frac{\partial R}{\partial \theta_{td}}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

где $\underset{t \in T}{\text{norm}}(x_t) = \frac{\max\{x_t,0\}}{\sum_{s \in T} \max\{x_s,0\}}$ — операция нормировки вектора.

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

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

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

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

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

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- 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

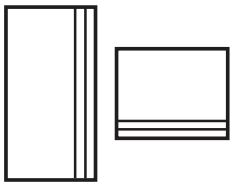
Для построения сложных моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».

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

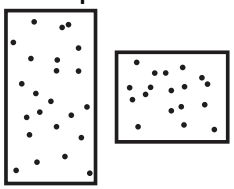
-- стандартизуемые этапы

Регуляризаторы для улучшения интерпретируемости тем



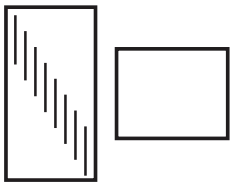
Сглаживание фоновых тем $B \subset T$:

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



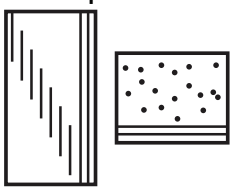
Разреживание предметных тем $S = T \setminus B$:

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



Декоррелирование для повышения различности тем:

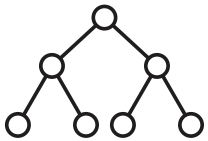
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем

Иерархические, темпоральные, регрессионные модели

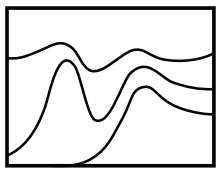
hierarchy



Связь родительских тем t с дочерними подтемами s:

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

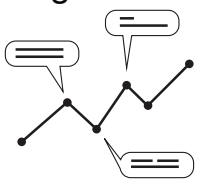
temporal



Темпоральные модели с модальностью времени i:

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

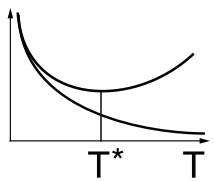
regression



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

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

n of topics

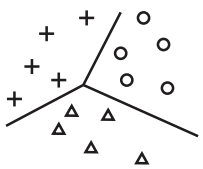


Разреживание $p(t)$ для отбора тем:

$$R(\Theta) = -\tau \sum_{t\in T} \frac{1}{|T|} \ln p(t), \quad p(t) = \sum_{d\in D} p(d)\theta_{td}.$$

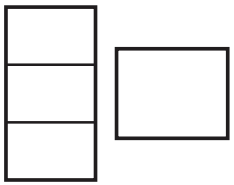
Специальные случаи мультимодальных тематических моделей

supervised



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

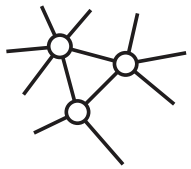
multilanguage



Модальность языков и регуляризация со словарём $\pi_{uwt} = p(u|w,t)$ переводов с языка k на ℓ:

$$R(\Phi,\Pi) = \tau \sum_{u\in W^k} \sum_{t\in T} n_{ut} \ln \sum_{w\in W^\ell} \pi_{uwt}\phi_{wt}$$

graph



Модальность вершин графа v, содержащих D_v:

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v)\in E} S_{uv} \sum_{t\in T} n_t^2 \left(\frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|}\right)^2.$$

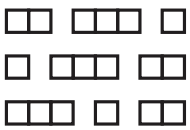
geospatial

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

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

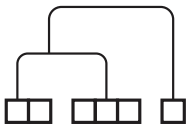
Тематические модели связного текста (beyond bag-of-words)

n-gram



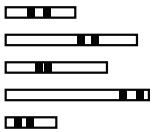
Модели с модальностями *n*-грамм, коллокаций, именованных сущностей

syntax



Модели, учитывающие результаты автоматического синтаксического разбора (SyntaxNet)

coherence

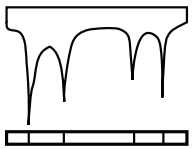


Модели дистрибутивной семантики на основе совстречаемости слов (битермы, когерентность)

sentence

Тематические модели, учитывающие границы предложений, абзацев и секций документов

segmentation



Тематические модели сегментации с автоматическим определением границ сегментов

Поиск и классификация этничного дискурса в соцсетях

Задача: найти все этно-релевантные темы для мониторинга межнациональных отношений.

Используем словарь из 300 этнонимов для обучения тем.

Мешок регуляризаторов:

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{temporal}\right) + R\left(\text{geospatial}\right) + R\left(\text{sentiment}\right) \to \max$$

Результаты: число релевантных тем выросло с 45 для LDA до 83 для 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.

Разведочный поиск в коллективных блогах

Задача: поиск документов по длинному запросу.

Мешок регуляризаторов:

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

Результаты:

- Точность и полнота увеличились
  с $(65\%, 73\%)$ для LDA до $(85\%, 92\%)$ для ARTM
  на данных Habrahabr.ru и TechCrunch.com.
- Точность и полнота сравнимы с результатами асессоров.
- Тематический поиск даёт результат мгновенно,
  асессоры тратят на эту же работу в среднем 30 минут.

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

Иерархическая темпоральная модель новостного потока

Задачи:

- наращивать 3х-уровневую иерархию динамически
- обеспечить интерпретируемость и именование всех тем
- управлять медиакомпаниями и творческими заданиями

Мешок регуляризаторов:

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

Результат: ... (исследование продолжается)

Сценарный анализ записей разговоров контакт-центра

Задачи:

- выделить сценарии диалогов оператор–клиент
- автоматизировать оценивание качества работы операторов
- выработать онлайновые подсказки для оператора

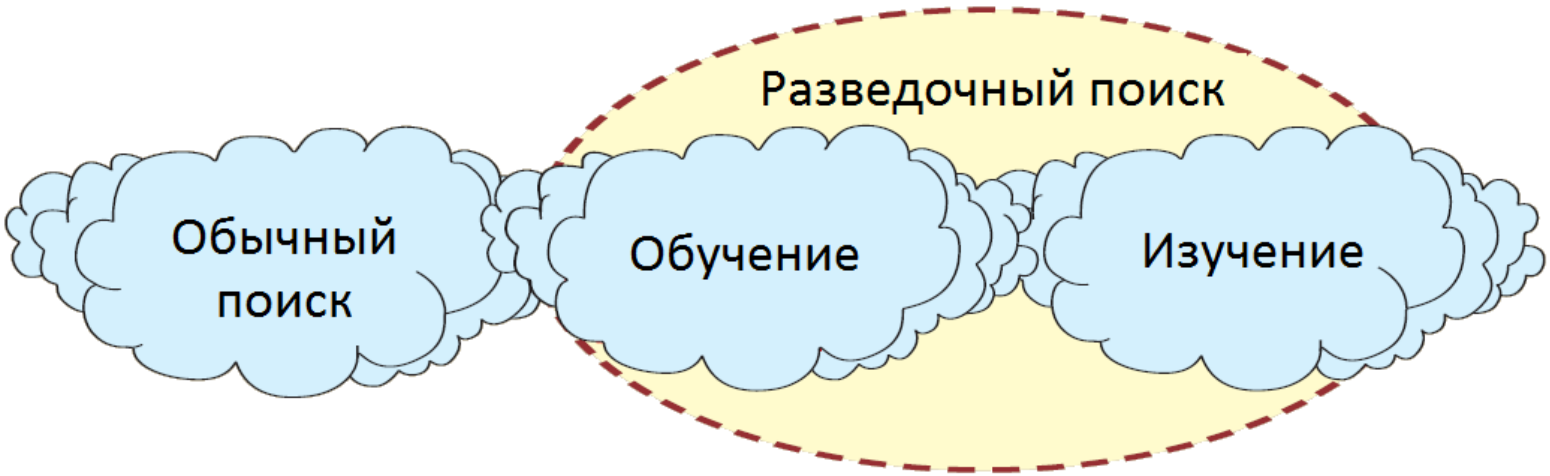
Мешок регуляризаторов:

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

Результат: качество сегментации выросло
с 40% у базового решения до 75% у ARTM

Концепция разведочного поиска (Exploratory Search)

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



Обычный поиск	Обучение	Изучение
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.

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

TechCrunch

Методика оценивания качества разведочного поиска

Поисковый запрос

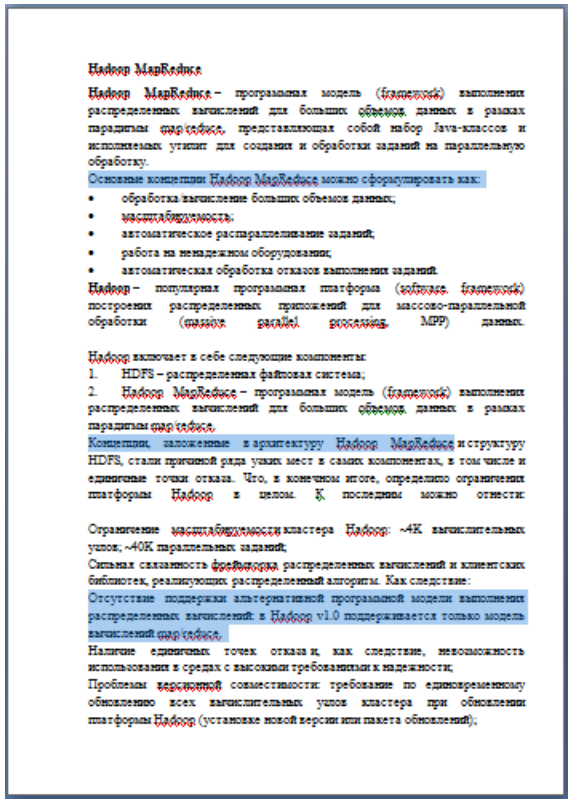
набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

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

Два задания асессорам

- найти как можно больше статей, пользуясь любыми средствами поиска (и засечь время)
- 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 себя 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

Разведочный тематический поиск

$q = (w_1, \ldots, w_{n_q})$ — текст запроса произвольной длины n_q

$\theta_{tq} = p(t|q)$ — тематический профиль запроса q
$\theta_{td} = p(t|d)$ — тематические профили документов $d \in D$

Косинусная мера близости документа d и запроса q:

$$\text{sim}(q,d) = \frac{\sum_t \theta_{tq}\theta_{td}}{\left(\sum_t \theta_{tq}^2\right)^{1/2}\left(\sum_t \theta_{td}^2\right)^{1/2}}.$$

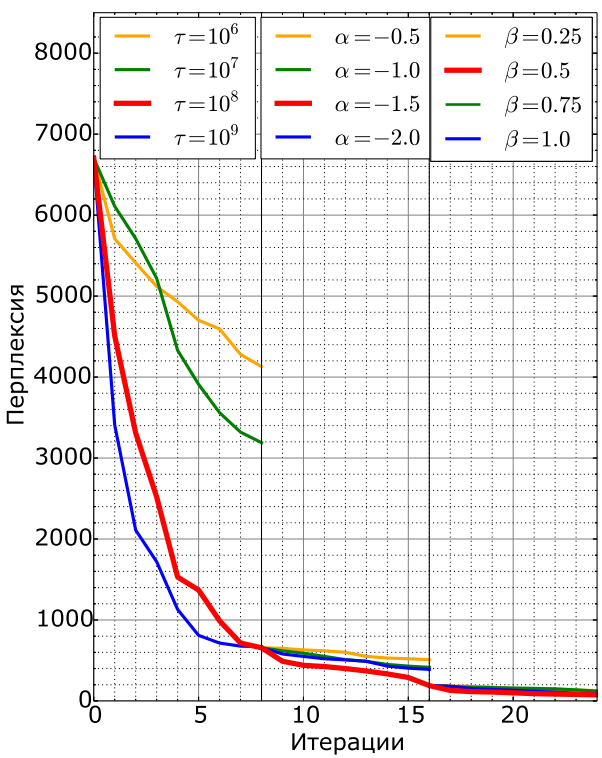
Ранжируем документы коллекции $d \in D$ по убыванию $\text{sim}(q,d)$

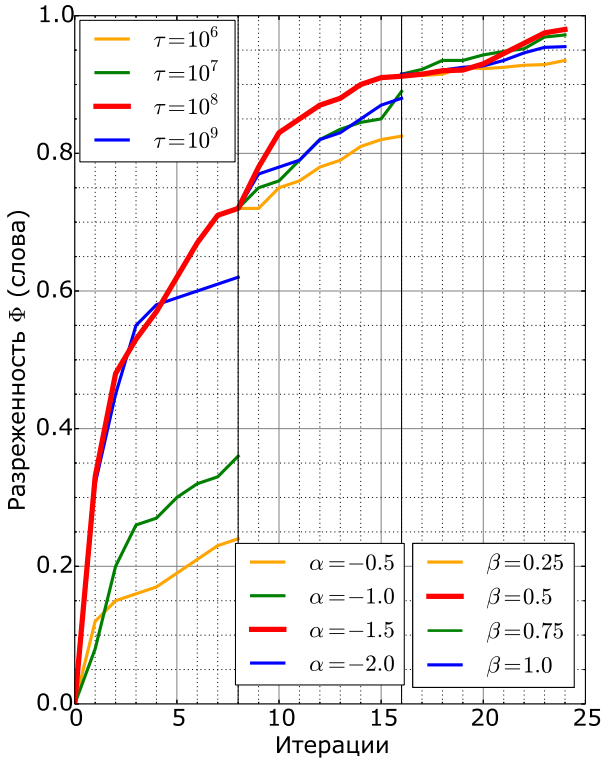
Выдача тематического поиска — k первых документов.

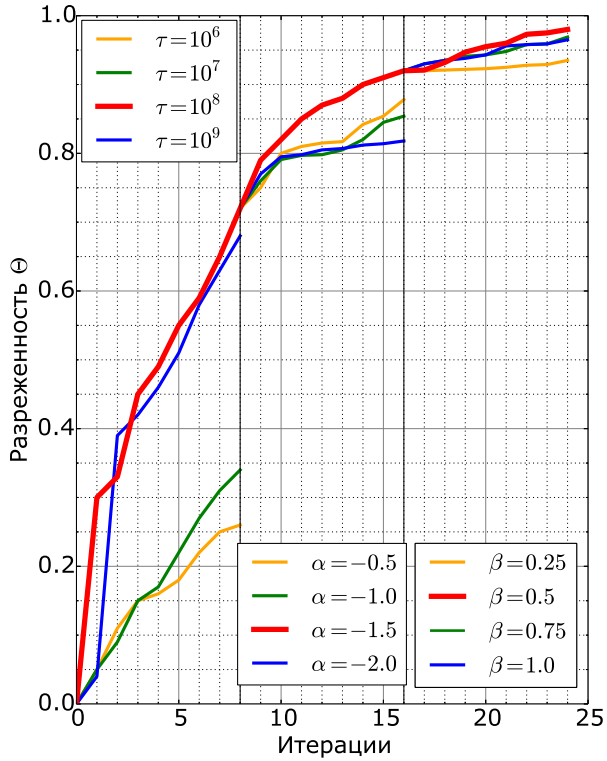
Реализация: *инвертированный индекс* для быстрого поиска документов d по каждой из тем t запроса

Последовательный подбор коэффициентов регуляризации

- декоррелирование распределений терминов в темах (τ),
- разреживание распределений тем в документах (α),
- сглаживание распределений терминов в темах (β).







Оценки качества поиска

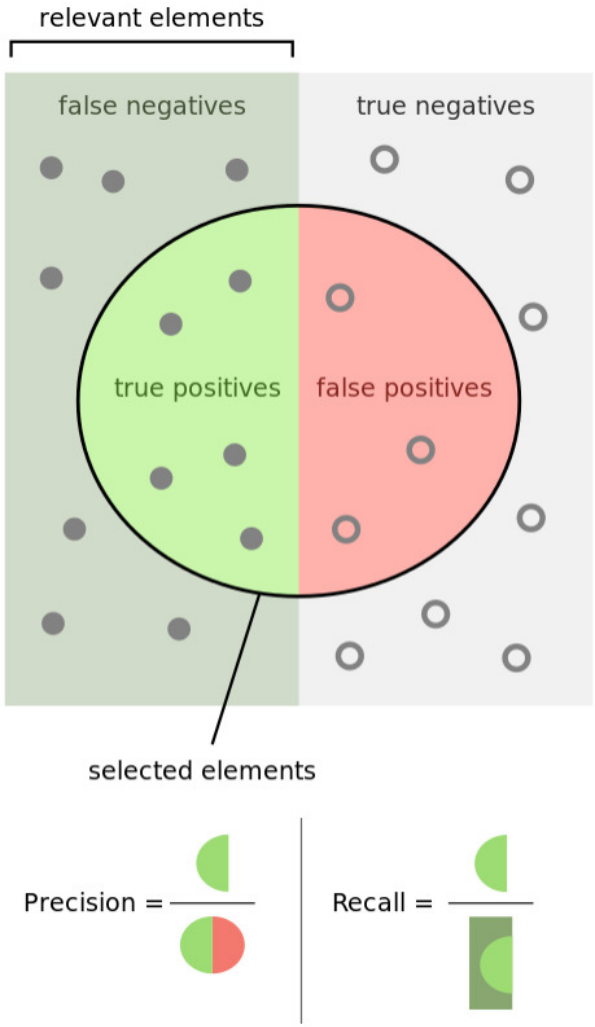
Precision — доля релевантных среди найденных
Recall — доля найденных среди релевантных

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \text{ — точность (precision)}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ — полнота, (recall)}$$

$$F_1 = \frac{P + R}{2PR} \text{ — F1-мера}$$

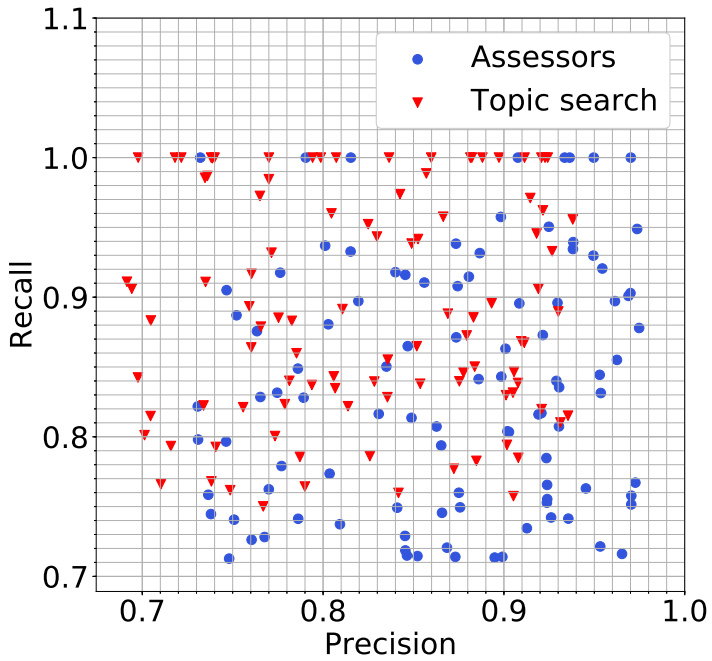
TP (true positive) — найденные релевантные
FP (false positive) — найденные нерелевантные
FN (false negative) — ненайденные релевантные



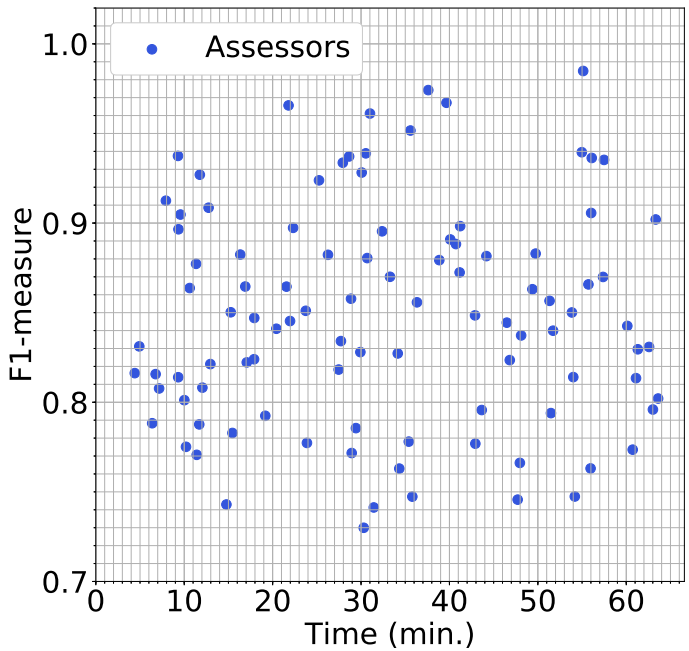
Результаты измерения точности и полноты по запросам

100 запросов, 3 асессора на запрос

точность и полнота поиска



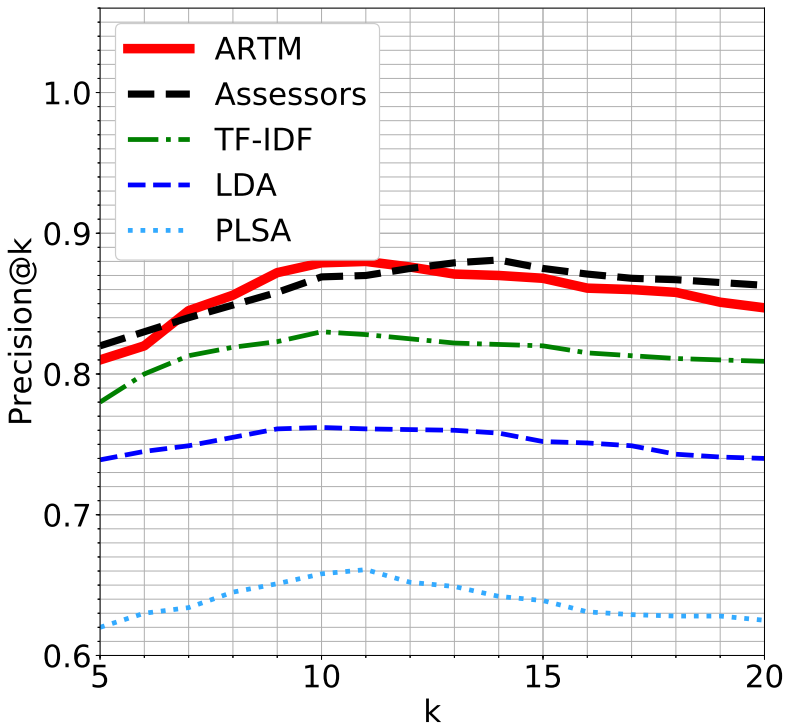
время и F_1-мера (асессоры)

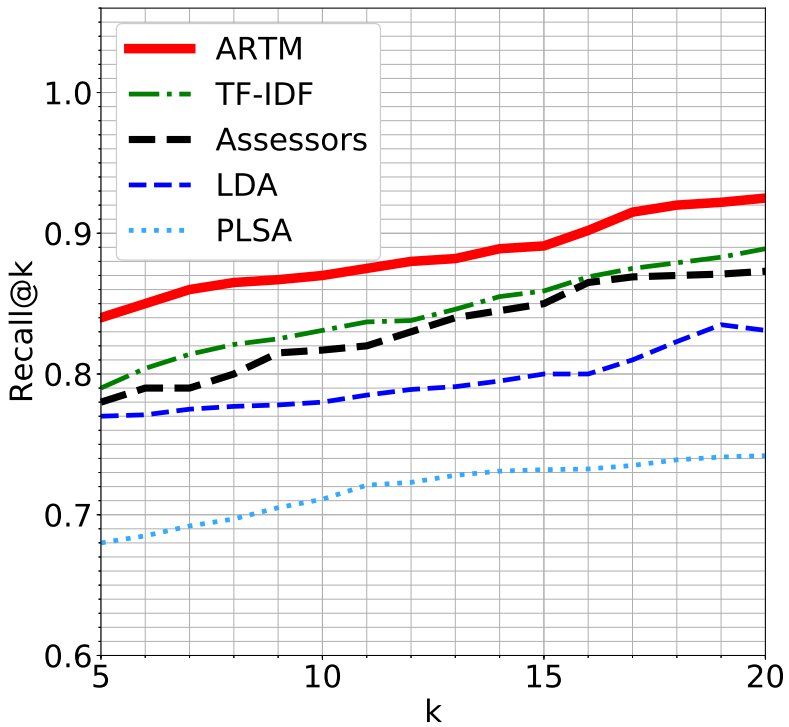


- среднее время обработки запроса асессором — 30 минут
- полнота чуть лучше, точность чуть хуже, чем у асессоров

Сравнение с асессорами по качеству поиска

Точность и полнота по первым k позициям поисковой выдачи (коллекция Habrahabr.ru)

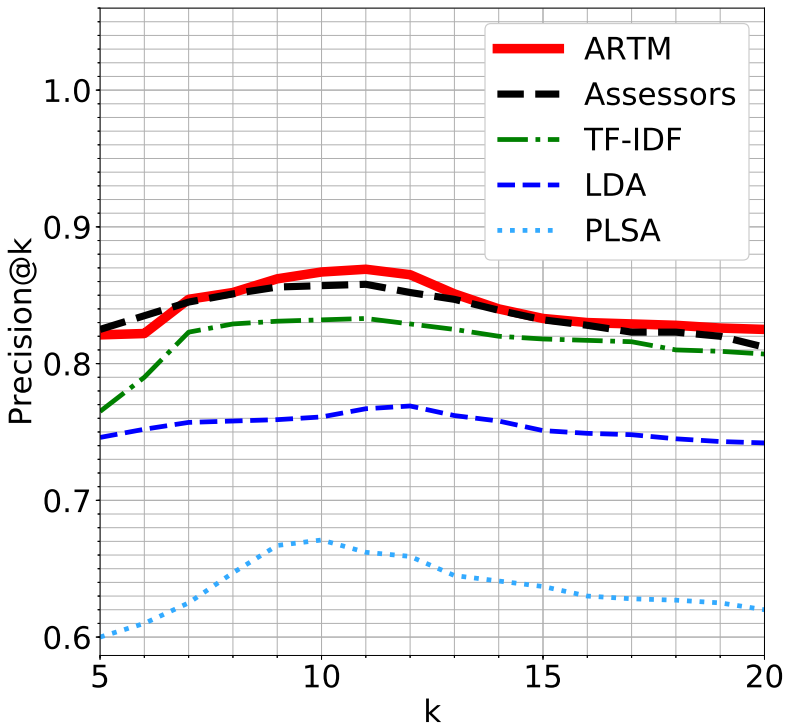


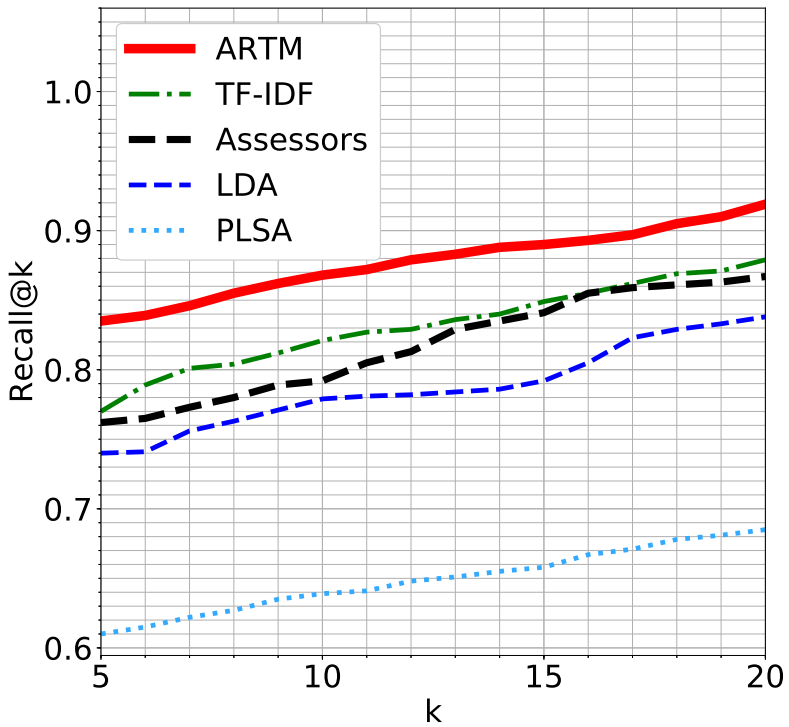


A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Сравнение с асессорами по качеству поиска

Точность и полнота по первым k позициям поисковой выдачи (коллекция TechCrunch.com)





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

Влияние комбинаций регуляризаторов на качество поиска

Декоррелирование, Θ-разреживание, Φ-сглаживание

	Habrahabr				TechCrunch			
	$R = 0$	Д	ДΘ	ДΘΦ	$R = 0$	Д	ДΘ	ДΘΦ
Prec@5	0.628	0.748	0.771	**0.810**	0.652	0.775	0.779	**0.819**
Prec@10	0.653	0.776	0.812	**0.879**	0.679	0.787	0.819	**0.867**
Prec@15	0.642	0.765	0.792	**0.868**	0.669	0.773	0.798	**0.833**
Prec@20	0.643	0.759	0.783	**0.847**	0.673	0.777	0.792	**0.825**
Recall@5	0.692	0.784	0.805	**0.840**	0.673	0.812	0.812	**0.835**
Recall@10	0.714	0.814	0.834	**0.870**	0.685	0.821	0.845	**0.868**
Recall@15	0.725	0.835	0.867	**0.891**	0.712	0.859	0.869	**0.890**
Recall@20	0.735	0.862	0.891	**0.925**	0.723	0.882	0.895	**0.919**

- комбинирование регуляризаторов улучшает качество поиска
- хотя исходно все регуляризаторы нацелены на улучшение интерпретируемости тем и не оптимизируют поиск явно

Влияние сочетания модальностей на качество поиска

Коллекция Habrahabr.ru. Число тем $|T| = 200$. Модальности: Слова, Биграммы, Теги, Хабы, Комментаторы, Авторы.

	асессоры	С	К	СБ	СБТХ	все
Prec@5	0.821	0.612	0.549	0.654	0.737	**0.810**
Prec@10	0.869	0.635	0.568	0.701	0.752	**0.879**
Prec@15	0.875	0.625	0.532	0.685	0.682	**0.868**
Prec@20	0.863	0.616	0.533	0.682	0.687	**0.847**
Recall@5	0.780	0.722	0.636	0.797	0.827	**0.840**
Recall@10	0.817	0.744	0.648	0.812	0.875	**0.870**
Recall@15	0.850	0.778	0.677	0.842	0.893	**0.891**
Recall@20	0.873	0.803	0.685	0.852	0.898	**0.925**

- Наилучшее качество поиска — по всем модальностям
- Наиболее полезные модальности — слова и теги

Влияние числа тем на качество поиска

Коллекция Habrahabr.ru

Используем все 5 модальностей, меняем $|T|$

	асессоры	100	150	**200**	250	400
Prec@5	0.821	0.662	0.721	**0.810**	0.761	0.693
Prec@10	0.869	0.761	0.812	**0.879**	0.825	0.673
Prec@15	0.875	0.733	0.795	**0.868**	0.791	0.651
Prec@20	0.863	0.724	0.795	**0.847**	0.792	0.642
Recall@5	0.780	0.732	0.807	**0.840**	0.821	0.721
Recall@10	0.817	0.771	0.843	**0.870**	0.851	0.751
Recall@15	0.850	0.824	**0.895**	0.891	0.871	0.773
Recall@20	0.873	0.857	0.905	**0.925**	0.892	0.771

- Наилучшее качество поиска — при 200 темах
- Тематический поиск превосходит асессоров по полноте

Резюме

- Тематическое моделирование — это восстановление латентных тем по коллекции текстовых документов
- Задача сводится к стохастическому матричному разложению
- Задача является некорректно поставленной, так как множество её решений в общем случае бесконечно
- Аддитивная регуляризация (ARTM) доопределяет задачу и позволяет строить модели с заданными свойствами
- Онлайновый EM-алгоритм хорошо распараллеливается и тематизирует большие коллекции за один проход
- Разведочный тематический поиск против асессоров: точность та же, полнота на 5% выше, 1 сек. вместо 30 мин.

Литература

- *К.В.Воронцов.* Обзор вероятностных тематических моделей. 2017. – NEW!
  http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf
- *К.В.Воронцов.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 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.Chirkova, K.Vorontsov.* Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.
- *A.Ianina, K.Vorontsov.* Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.
- *A.Potapenko, A.Popov, K.Vorontsov.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.