Тематические векторные представления текста и разведочный информационный поиск

Воронцов Константин Вячеславович

(Лаборатория машинного интеллекта МФТИ)

Математический кружок ФПМИ МФТИ

МФТИ • 20 сентября 2019

Содержание

Что такое «тема» в коллекции текстовых документов?

Выделение тем — первый шаг к пониманию смысла текста

- *тема* — специальная терминология предметной области
- *тема* — кластер семантически близких фраз

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

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

Пример. Мультиязычная тематическая модель Википедии

216 175 русско-английских пар статей.
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. Мультиязычная тематическая модель Википедии

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

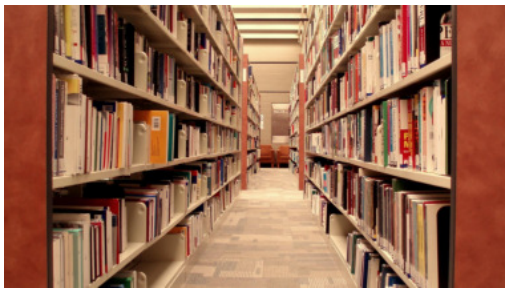
Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	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.

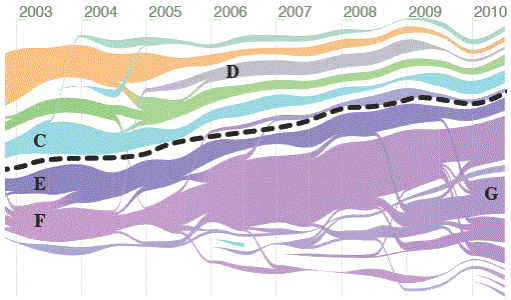
Приложения тематического моделирования

разведочный поиск в электронных библиотеках

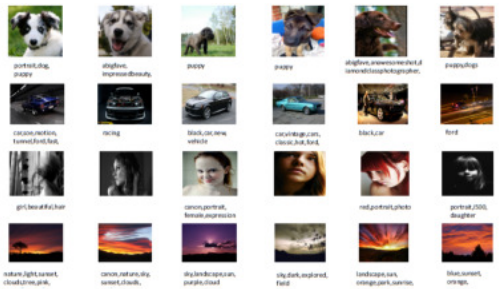


поиск тематического контента в соцсетях

детектирование и трекинг новостных сюжетов

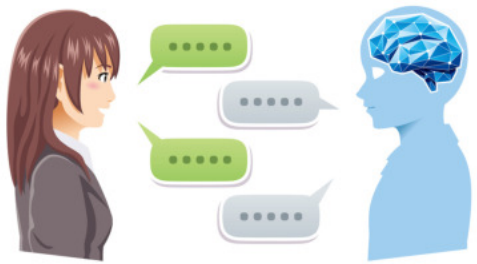


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

управлением диалогом в разговорном интеллекте



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: $p(w|d,t) = p(w|t)$

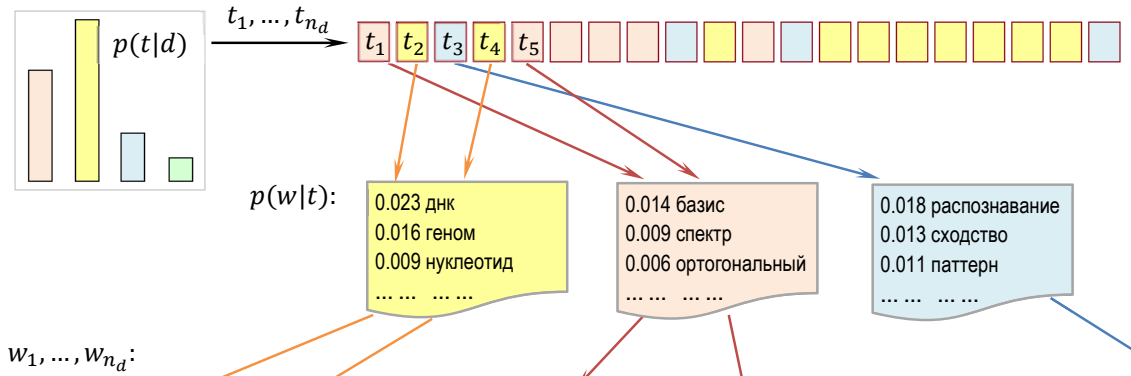
Тематическая модель, по формуле полной вероятности:

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

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:

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



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

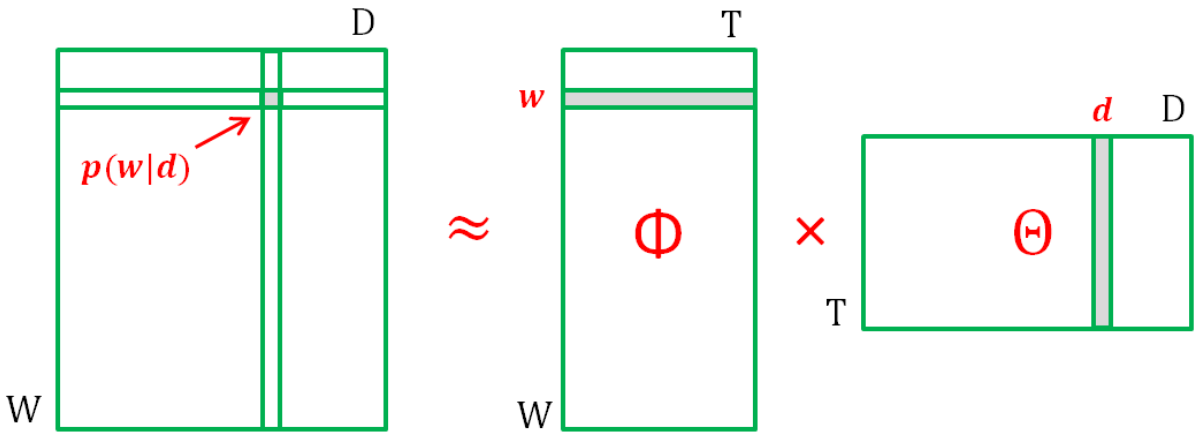
Дано: коллекция текстовых документов

- 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

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

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

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

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

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

Условия вырожденности модели для тем и документов

Решение может быть вырожденным для некоторых тем (столбцов матриц Φ) и документов (столбцов матрицы Θ).

Тема t вырождена, если для всех терминов $w \in W$

$$n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \leqslant 0.$$

Если тема t вырождена, то $p(w|t) = \phi_{wt} \equiv 0$; это означает, что тема исключается из модели (происходит отбор тем).

Документ d вырожден, если для всех тем $t \in T$

$$n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \leqslant 0.$$

Если документ d вырожден, то $p(t|d) = \theta_{td} \equiv 0$; это означает, что модель не в состоянии описать данный документ.

Напоминание. Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_x; \\ g_i(x) \leqslant 0, \quad i = 1, \ldots, m; \\ h_j(x) = 0, \quad j = 1, \ldots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители μ_i, $i = 1, \ldots, m$, λ_j, $j = 1, \ldots, k$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0; \ h_j(x) = 0; \ \text{(исходные ограничения)} \\ \mu_i \geqslant 0; \ \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; \ \text{(условие дополняющей нежёсткости)} \end{cases}$$

Вывод системы уравнений из условий Каруша–Куна–Таккера

1. Условия ККТ для ϕ_{wt} (для θ_{td} всё аналогично):

$$\sum_d n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \mu_{wt}; \quad \mu_{wt} \geqslant 0; \quad \mu_{wt}\phi_{wt} = 0.$$

2. Умножим обе части равенства на ϕ_{wt} и выделим p_{tdw}:

$$\phi_{wt}\lambda_t = \sum_d n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}.$$

3. Если $\lambda_t \leqslant 0$, то тема t вырождена, $\phi_{wt} \equiv 0$ для всех w.
4. Если $\lambda_t > 0$, то либо $\phi_{wt} = 0$, либо $n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} > 0$:

$$\phi_{wt}\lambda_t = \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

5. Суммируем обе части равенства по $w \in W$:

$$\lambda_t = \sum_{w \in W} \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

6. Подставим λ_t из (5) в (4), получим требуемое. ■

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:

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

Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

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

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

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

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

Байесовское обучение — доминирующий подход в TM

Основа подхода — байесовский вывод:

$$\text{Posterior}(\Phi, \Theta|\text{data}) \propto \text{Prior}(\Phi, \Theta)\, \text{P}(\text{data}|\Phi, \Theta)$$

В модели LDA Prior и Posterior — распределения Дирихле.

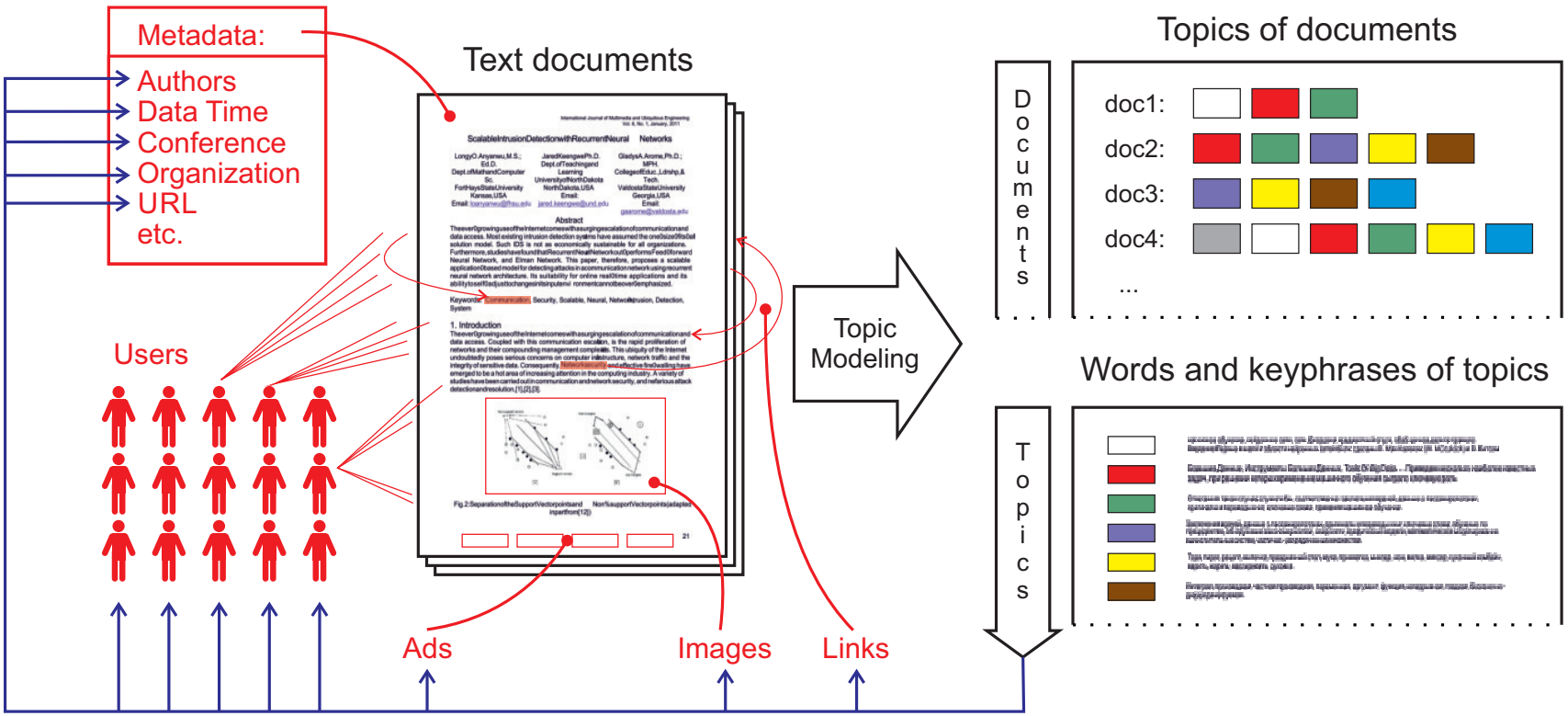
Проблемы:

- Нам нужны лишь значения Φ, Θ, а не их распределения
- Prior Дирихле имеет слабые лингвистические обоснования
- Задача сильно усложняется для несопряжённых Prior
- Байесовский вывод уникален для каждой модели
- Невозможно модульное комбинирование моделей

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

Задачи мультимодального тематического моделирования

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



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

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

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

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

Модальность биграмм улучшает интерпретируемость тем

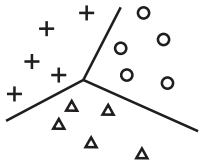
Коллекция 1000 статей конференций ММРО, ИОИ на русском

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.

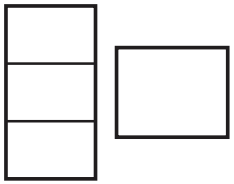
Специальные случаи мультимодальных тематических моделей

supervised



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

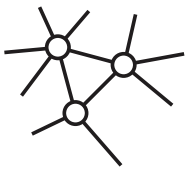
multilanguage



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 $S_{gg'}$:

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

Напоминание. Дивергенция Кульбака–Лейблера

Функция расстояния между распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

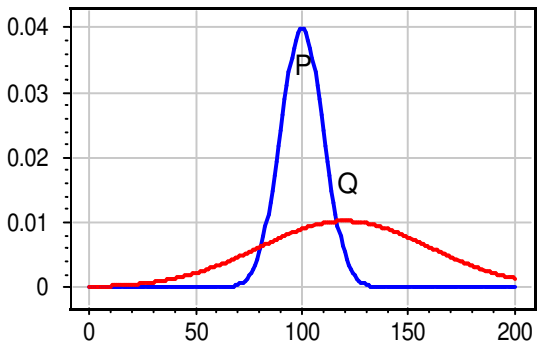
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

1. $\mathrm{KL}(P\|Q) \geqslant 0$; $\quad \mathrm{KL}(P\|Q) = 0 \;\Leftrightarrow\; P = Q$;

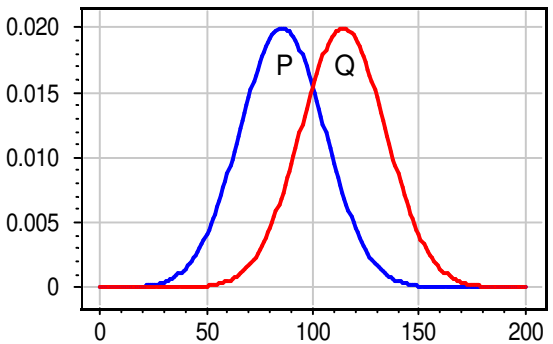
2. Минимизация KL эквивалентна максимизации правдоподобия:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_\alpha \quad \Longleftrightarrow \quad \sum_{i=1}^n p_i \ln q_i(\alpha) \to \max_\alpha.$$

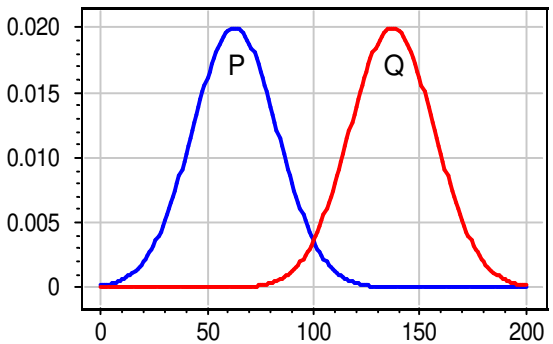
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



$\mathrm{KL}(P\|Q) = 0.442$
$\mathrm{KL}(Q\|P) = 2.966$

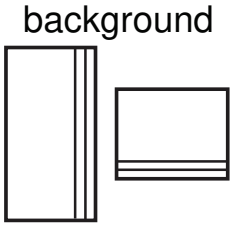


$\mathrm{KL}(P\|Q) = 0.444$
$\mathrm{KL}(Q\|P) = 0.444$



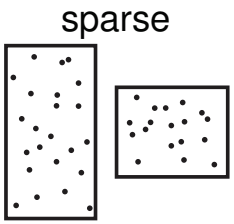
$\mathrm{KL}(P\|Q) = 2.969$
$\mathrm{KL}(Q\|P) = 2.969$

Регуляризаторы для улучшения интерпретируемости тем



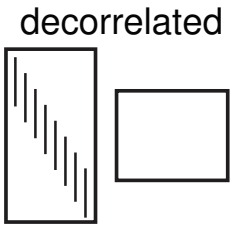
Сглаживание фоновых тем $B \subset T$:

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



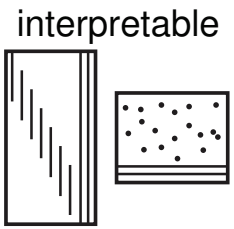
Разреживание предметных тем $S = T \setminus B$:

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



Декоррелирование для повышения различности тем:

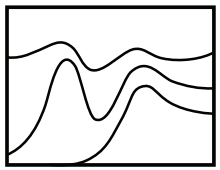
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем

Темпоральные, регрессионные, иерархические модели

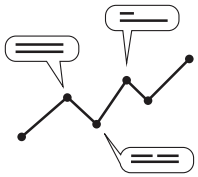
temporal



Темпоральные модели с модальностью времени i:

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

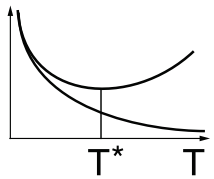
regression



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

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

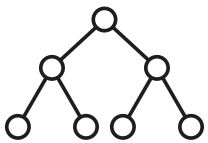
n of topics



Разреживание $p(t)$ для отбора тем:

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

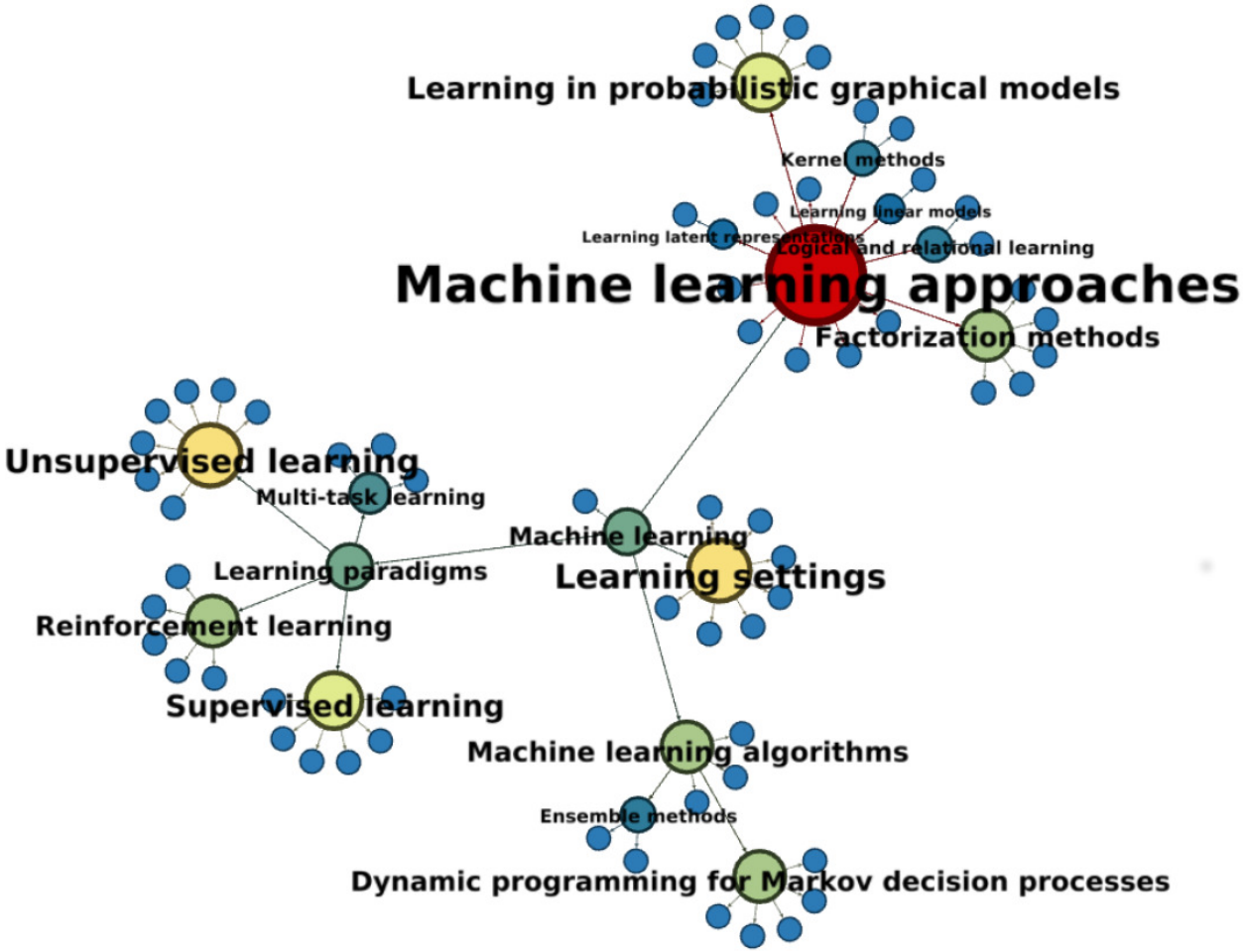
hierarchy



Связь родительских тем t с дочерними подтемами s:

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

Пример тематической иерархии



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

Послойное построение уровней тематической иерархии

Шаг 1. Строим модель с небольшим числом тем.

Шаг *k*. Пусть модель с множеством тем T уже построена. Строим множество дочерних тем S (subtopics), $|S| > |T|$.

Родительские темы приближаются смесями дочерних тем:

$$\sum_{t\in T} n_t \,\mathrm{KL}_w\Big(p(w|t) \,\Big\|\, \sum_{s\in S} p(w|s)p(s|t)\Big) \;\to\; \min_{\Phi,\Psi},$$

где $p(s|t) = \psi_{st}$, $\Psi = (\psi_{st})_{S\times T}$ — *матрица связей*.

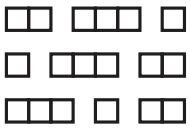
Родительская $\Phi^p \approx \Phi\Psi$, отсюда регуляризатор матрицы Φ:

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

Родительские темы t — *псевдо-документы* с частотами слов n_{wt}.

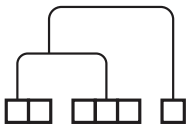
Тематические модели связного текста (beyond bag-of-words)

n-gram



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

syntax

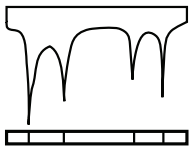


Модели, учитывающие результаты автоматического синтаксического разбора (UDPipe или SyntaxNet)

sentence

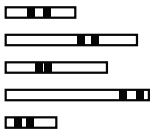
Тематические модели, учитывающие границы предложений, абзацев и секций документов

segmentation



Тематические модели сегментации с автоматическим определением границ сегментов

coherence



Модели дистрибутивной семантики на основе частоты совместной встречаемости слов

Дистрибутивная гипотеза и виды семантической близости слов

- Words that occur in the same contexts tend to have similar meanings [Harris, 1954].
- You shall know a word by the company it keeps [Firth, 1957].

Синтагматическая близость слов:
со-встречаемость слов в одном контексте.

здание–строитель, кран–вода, функция–точка

Парадигматическая близость слов:
взаимозаменяемость слов в одном контексте.

здание–дом, кран–смеситель, функция–отображение

Z.Harris. Distributional structure. 1954.
J.R.Firth. A synopsis of linguistic theory 1930-1955. Oxford, 1957.
P.D.Turney, P.Pantel. From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research (JAIR), 2010.

Модели векторных представлений для текстов и графов

word2vec: векторные представления (эмбединги) слов
T.Mikolov et al. Efficient estimation of word representations in vector space. 2013.

paragraph2vec: эмбединги фрагментов или документов
Q.Le, T.Mikolov. Distributed representations of sentences and documents. 2014.

sent2vec: эмбединги предложений
M.Pagliardini et al. Unsupervised learning of sentence embeddings using compositional n-gram features. 2017.

FastText: эмбединги символьных n-грамм
`https://github.com/facebookresearch/fastText`

node2vec: эмбединги вершин графа
A.Grover, J.Leskovec. Node2vec: scalable feature learning for networks. 2016.

graph2vec: более общие эмбединги на графах
A.Narayanan et al. Graph2vec: learning distributed representations of graphs. 2017.

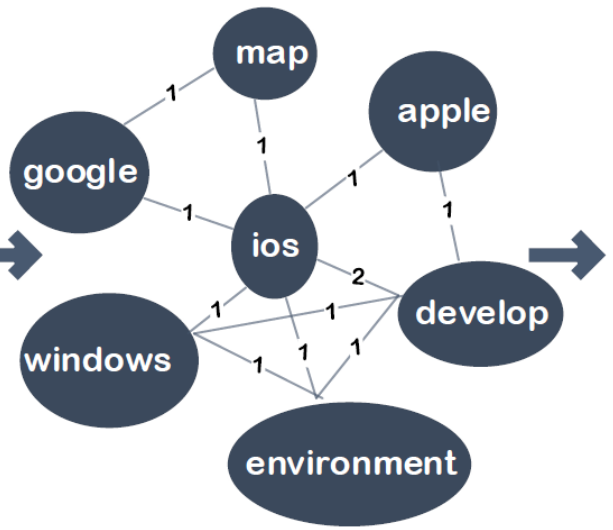
StarSpace: эмбединги чего угодно от Facebook AI Research
L.Wu, A.Fisch, S.Chopra, K.Adams, A.B.J.Weston. StarSpace: embed all the things! 2018.

Недостаток: координаты векторов не интерпретируемы

Модель сети слов — Word Network Topic Model

Идея: моделировать не документы, а связи между словами.
d_u — псевдо-документ, объединение всех контекстов слова u.
n_{uw} — число вхождений слова w в псевдо-документ d_u.
Контекст — короткое сообщение / предложение / окно $\pm h$ слов.

doc1	Google Map for iOS
doc2	iOS is developed by Apple
doc3	iOS develop environment for Windows



word	pseudo-document
map	google ios
google	map ios
apple	develop ios
ios	google map apple windows environment develop develop
windows	environment ios develop
develop	apple ios ios windows environment
environment	ios windows develop

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Модели WNTM и WTM (Word Topic Model)

Тематическая модель контекстов, разложение $W \times W$-матрицы:

$$p(w|d_u) = \sum_{t \in T} p(w|t)p(t|d_u) = \sum_{t \in T} \phi_{wt}\theta_{tu},$$

где d_u — псевдо-документ слова u.

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{uw} \log \sum_{t \in T} \phi_{wt}\theta_{tu} \to \max_{\Phi,\Theta},$$

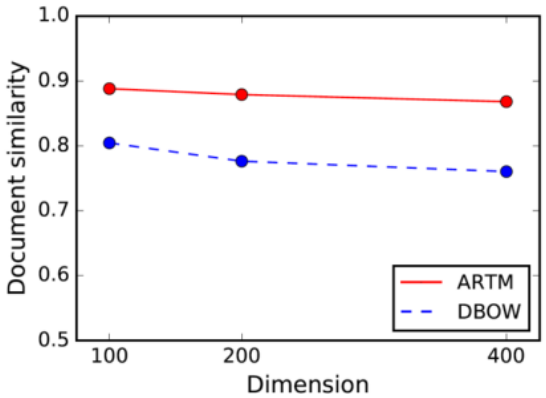
где n_{uw} — совстречаемость слов u, w (кстати, $n_{uw} = n_{wu}$).

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Berlin Chen. Word Topic Models for spoken document retrieval and transcription. ACM Trans., 2009.

word2vec и ARTM в задаче семантической близости документов

ArXiv triplets dataset: 20K троек статей:
⟨статья A, схожая статья B, непохожая статья C⟩



- обучение по 1M текстов статей ArXiv
- тестирование на триплетах ArXiv
- Конкурент DBOW: paragraph2vec [Dai et. al, 2015]

ARTM превосходит модель DBOW (distributed bag-of-words).

Andrew Dai, Cristopher Olah, Quoc Le. Document Embedding with Paragraph Vectors, CoRR, 2015

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Две коллекции новостей про технологии

Habrahabr.ru

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

TechCrunch.com

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

Предобработка текстов

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

Методика оценивания качества разведочного поиска

Поисковый запрос

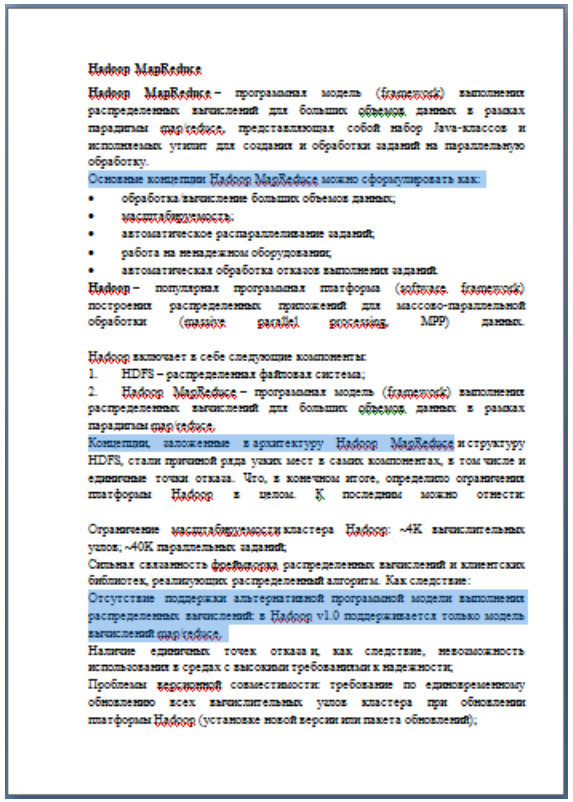
набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

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

Два задания асессорам

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



Hadoop MapReduce

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

Основные концепции Hadoop MapReduce можно сформулировать как:

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

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

Hadoop включает в себе следующие компоненты:

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

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

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

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

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

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

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

Пример запроса для разведочного поиска

Пример: фрагмент запроса «Система IBM Watson»

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

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

... ...

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

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

Тематика запросов разведочного поиска

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

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

Поиск документов, тематически близких к запросу

$\theta_{tq} = p(t|q)$ — тематический вектор запроса q
$\theta_{td} = p(t|d)$ — тематические векторы 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

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

Оценивание качества поиска

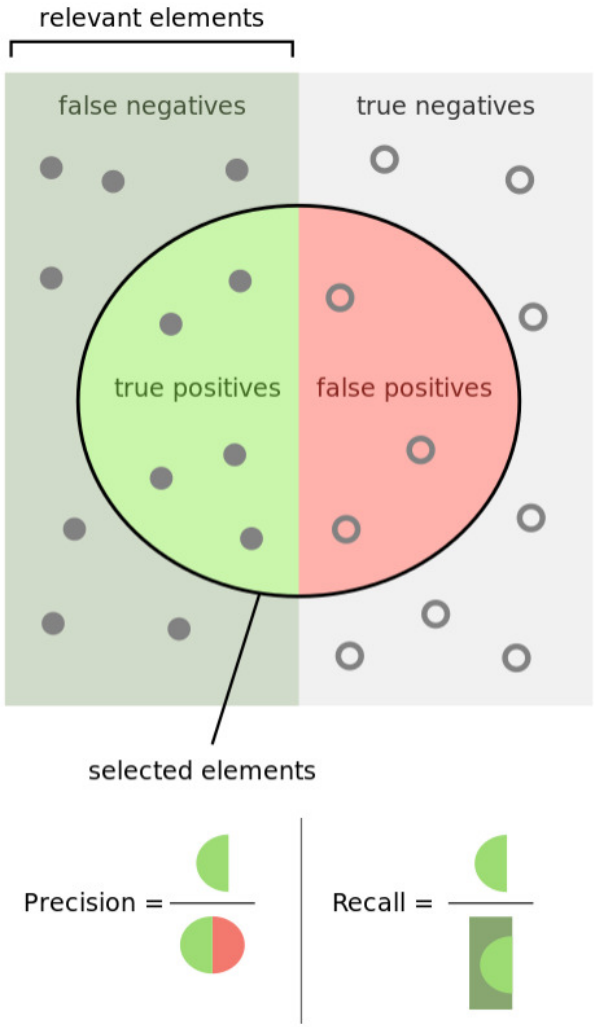
Precision — доля релевантных среди найденных
Recall — доля найденных среди релевантных

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

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

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

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



Какие модели поиска сравнивались

- **assessors:** результаты поиска, выполненного асессорами
- **TF-IDF, BM25:** сравнение документов по частотам слов
- **word2vec:** нетематические векторные представления слов
- **PLSA:** Probabilistic Latent Semantic Analysis (1999)
- **LDA:** Latent Dirichlet Allocation (2003)
- **ARTM:** тематическая модель с тремя регуляризаторами
- **hARTM:** двухуровневая иерархическая модель ARTM

Задачи регуляризаторов в ARTM и hARTM:

- сделать темы как можно более различными
- сделать векторы $p(t|d)$ как можно более разреженными
- не допустить вырожденности распределений $p(w|t)$

Стратегия регуляризации

Последовательное применение трёх регуляризаторов

1. декоррелирование тем:

$$R(\Phi) = -\tau \sum_{s,t \in T} \sum_{w \in W} \phi_{wt} \phi_{ws}$$

2. разреживание распределений $p(t|d)$:

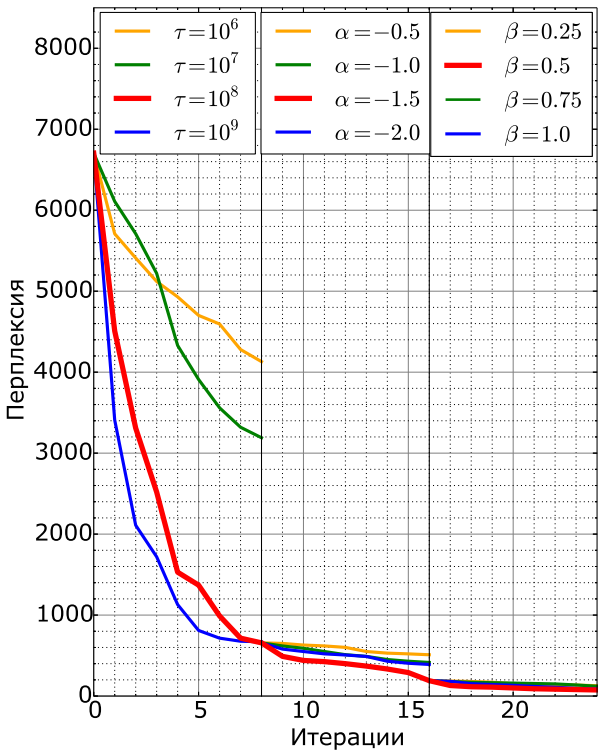
$$R(\Theta) = -\alpha \sum_{d,t} \ln \theta_{td}$$

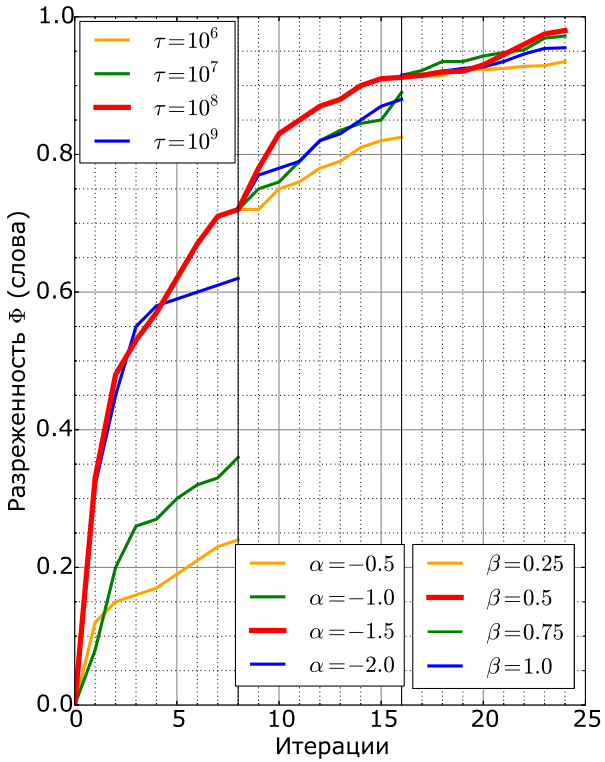
3. сглаживание распределений $p(w|t)$:

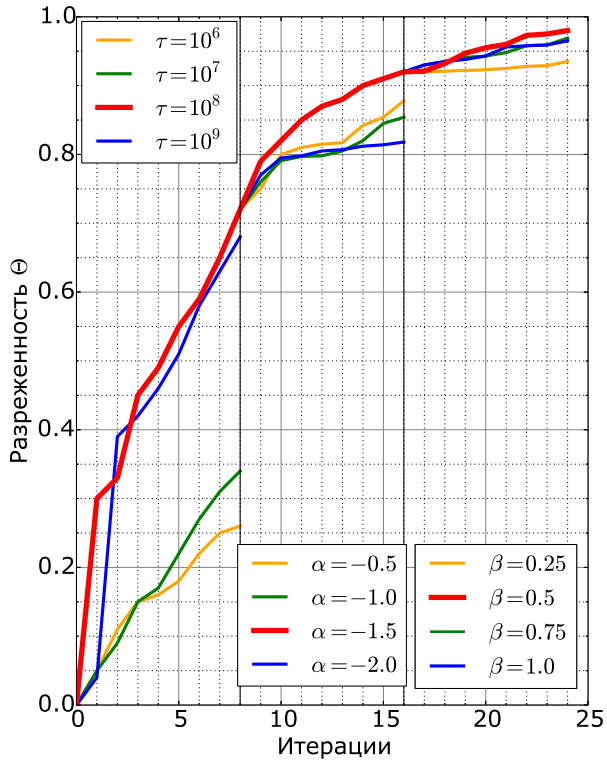
$$R(\Phi) = \beta \sum_{t,w} \ln \phi_{wt}$$

Последовательный подбор коэффициентов регуляризации

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



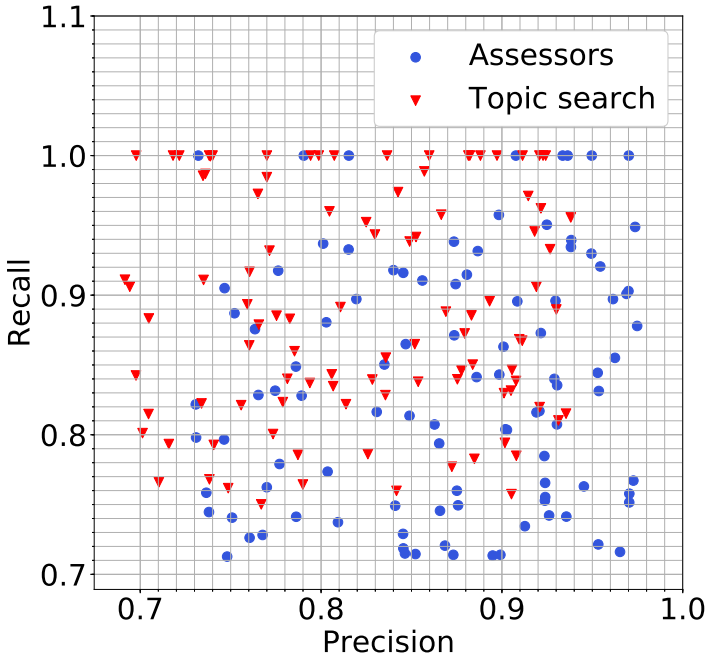




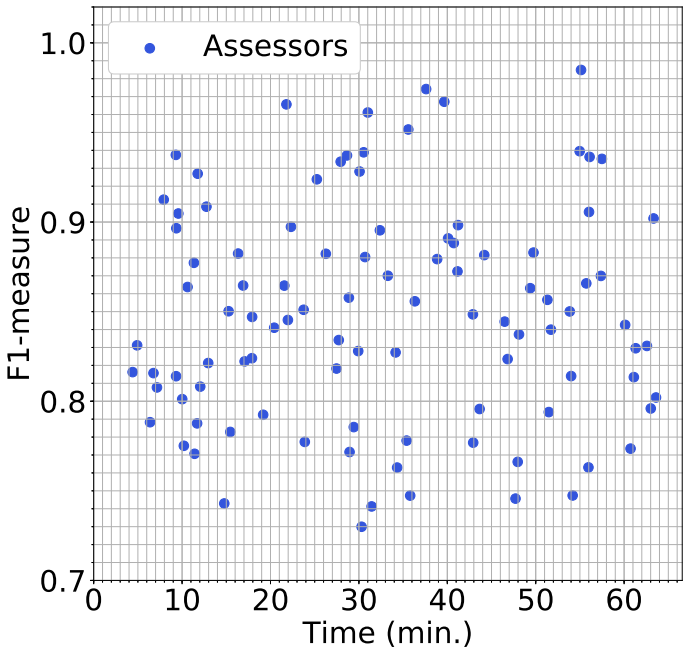
Результаты измерения точности и полноты по запросам

100 запросов, 3 асессора на запрос

точность и полнота поиска



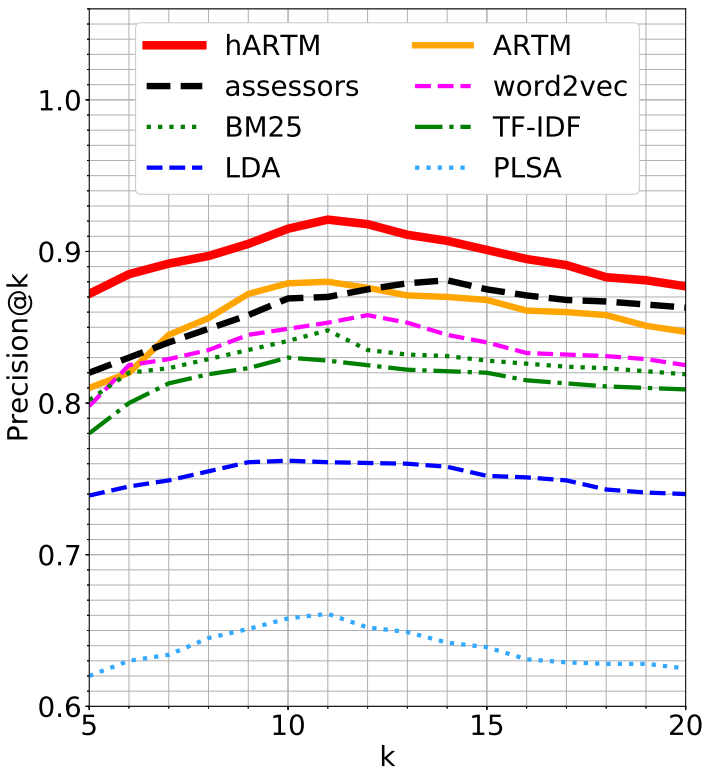
время и F_1-мера (асессоры)

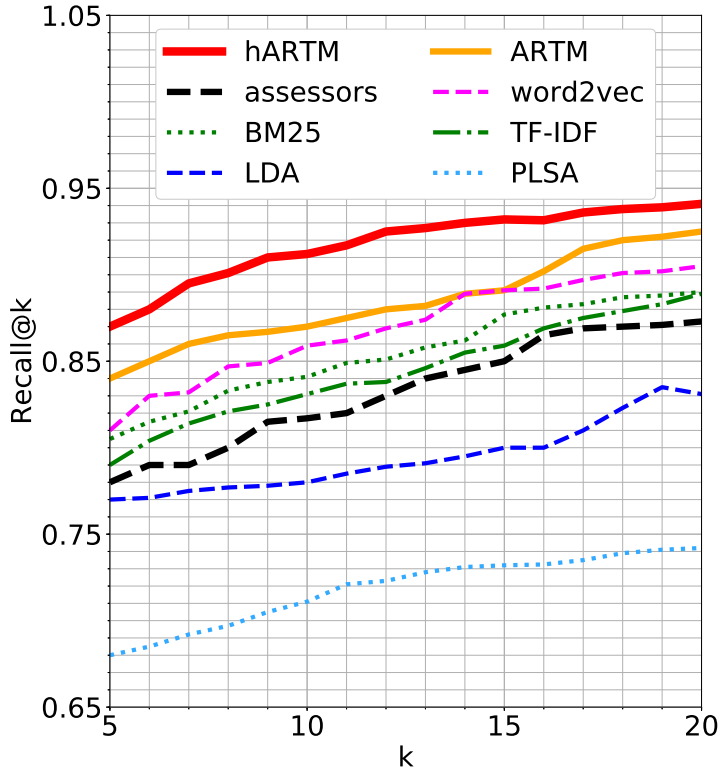


- среднее время обработки запроса асессором — 30 минут
- точность выше у асессоров, полнота — у поисковика

Сравнение с асессорами по качеству поиска

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

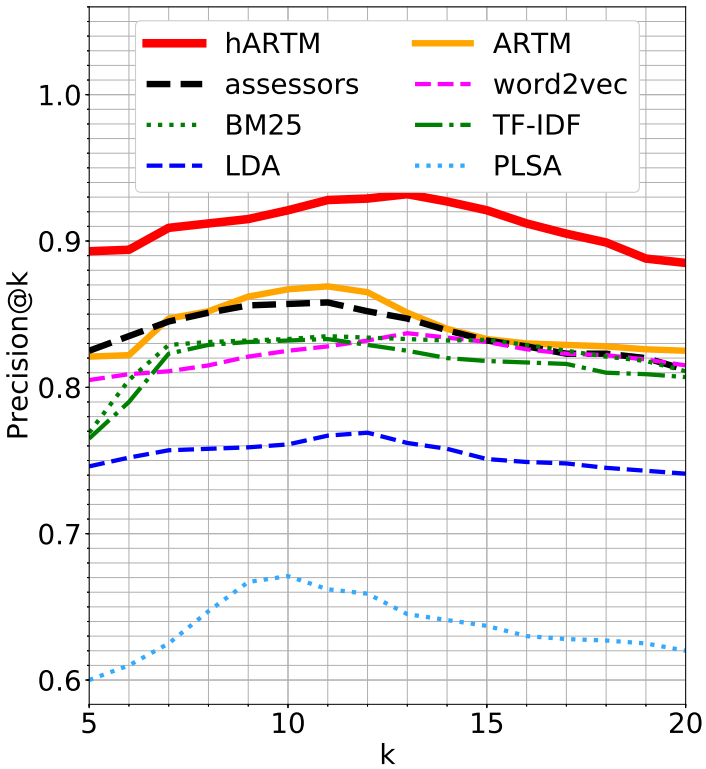


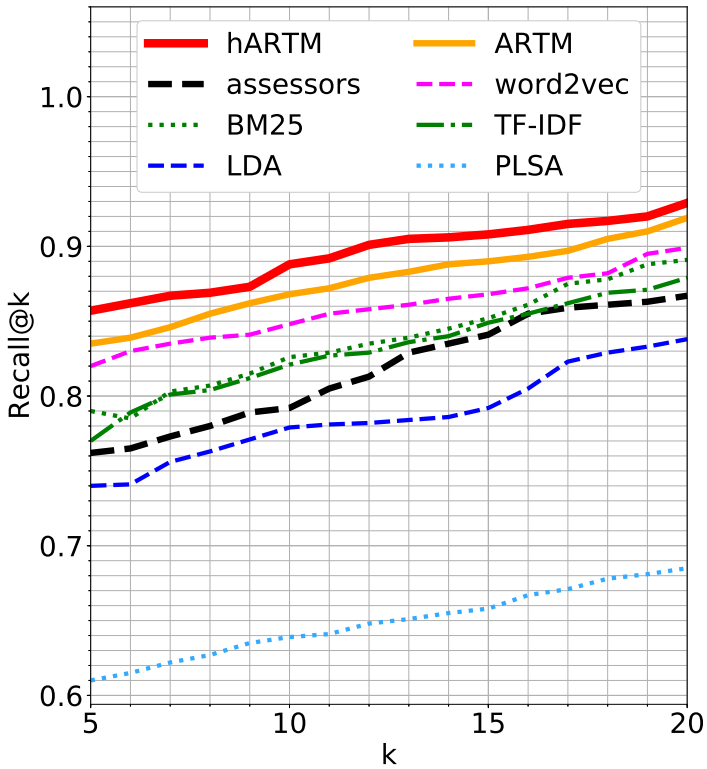


A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

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 числа тем на качество поиска

Все регуляризаторы и модальности, плоская модель

	Habrahabr						TechCrunch					
	acecc	100	150	**200**	250	400	acecc	350	400	450	**475**	500
Pr@5	0.821	0.662	0.721	**0.810**	0.761	0.693	0.822	0.653	0.725	0.752	**0.819**	0.777
Pr@10	0.869	0.761	0.812	**0.879**	0.825	0.673	0.851	0.663	0.732	0.762	**0.867**	0.811
Pr@15	0.875	0.733	0.795	**0.868**	0.791	0.651	0.835	0.682	0.743	0.787	**0.833**	0.793
Pr@20	0.863	0.724	0.795	**0.847**	0.792	0.642	0.813	0.650	0.743	0.773	**0.825**	0.793
R@5	0.780	0.732	0.807	**0.840**	0.821	0.721	0.762	0.731	0.762	0.793	**0.835**	0.817
R@10	0.817	0.771	0.843	**0.870**	0.851	0.751	0.792	0.763	0.793	0.812	**0.868**	0.855
R@15	0.850	0.824	**0.895**	0.891	0.871	0.773	0.835	0.782	0.807	0.855	**0.890**	0.882
R@20	0.873	0.857	0.905	**0.925**	0.892	0.771	0.867	0.792	0.823	0.862	**0.919**	0.903

- существует оптимальное число тем
- чем больше коллекция, тем больше оптимум числа тем

Влияние числа тем на качество поиска

Habrahabr. Все регуляризаторы и модальности, три уровня

$\|T_1\|$	**20**		**25**							**30**	
$\|T_2\|$	150	200	**250**		**275**			**300**		400	450
$\|T_3\|$	750	800	1200	1300	1300	**1400**	1500	1500	1600	3000	3500
Pr@5	0.625	0.743	0.840	0.852	0.869	**0.872**	0.870	0.805	0.771	0.705	0.672
Pr@10	0.648	0.754	0.851	0.867	0.882	**0.915**	0.901	0.811	0.799	0.722	0.694
Pr@15	0.632	0.752	0.850	0.872	0.878	**0.895**	0.889	0.809	0.785	0.729	0.703
Pr@20	0.629	0.745	0.845	0.861	0.871	0.877	**0.882**	0.803	0.778	0.710	0.681
R@5	0.632	0.780	0.845	0.869	0.883	**0.889**	0.872	0.851	0.841	0.721	0.695
R@10	0.654	0.792	0.859	0.873	0.905	**0.922**	0.881	0.873	0.850	0.749	0.703
R@15	0.675	0.805	0.874	0.892	0.932	**0.942**	0.905	0.889	0.863	0.787	0.725
R@20	0.684	0.824	0.889	0.901	0.958	**0.961**	0.912	0.904	0.878	0.805	0.734

- существует оптимальное число тем на каждом уровне
- 3 уровня сильно лучше, чем 1, и немного лучше, чем 2
- оптимальное число тем увеличивается

Влияние числа тем на качество поиска

TechCrunch. Все регуляризаторы и модальности, три уровня

$\|T_1\|$	**80**		**100**							**120**	
$\|T_2\|$	300	350	**500**		**550**			**600**		700	750
$\|T_3\|$	1500	1700	2500	2600	2600	**2800**	3000	3000	3200	4500	4700
Pr@5	0.655	0.707	0.751	0.792	0.887	**0.893**	0.890	0.789	0.722	0.703	0.678
Pr@10	0.678	0.712	0.773	0.823	0.895	**0.922**	0.905	0.805	0.741	0.722	0.692
Pr@15	0.692	0.715	0.775	0.831	0.902	**0.921**	0.907	0.821	0.743	0.725	0.703
Pr@20	0.687	0.709	0.761	0.819	0.889	0.885	**0.898**	0.809	0.736	0.719	0.683
R@5	0.751	0.795	0.802	0.856	0.871	**0.877**	0.863	0.852	0.831	0.738	0.705
R@10	0.767	0.812	0.825	0.875	0.892	**0.908**	0.879	0.871	0.842	0.751	0.711
R@15	0.772	0.824	0.841	0.887	0.912	**0.927**	0.901	0.893	0.854	0.772	0.721
R@20	0.783	0.830	0.854	0.892	0.931	**0.949**	0.935	0.905	0.871	0.790	0.732

- существует оптимальное число тем на каждом уровне
- 3 уровня сильно лучше, чем 1, и немного лучше, чем 2
- оптимальное число тем увеличивается

Влияние модальностей на качество поиска

Все регуляризаторы и модальности, 3 уровня, оптимальное $|T|$

Модальности: Words, Bigrams, Authors, Comments, Tags, Hubs, Categories

	Habrahabr						TechCrunch					
	acecc	W	Com	WB	WBTH	**All**	acecc	W	C	WB	WBC	**All**
Pr@5	0.821	0.621	0.558	0.673	0.871	**0.872**	0.822	0.718	0.569	0.795	0.891	**0.893**
Pr@10	0.869	0.645	0.567	0.712	0.911	**0.915**	0.851	0.729	0.592	0.807	0.919	**0.922**
Pr@15	0.875	0.631	0.532	0.693	0.894	**0.895**	0.835	0.737	0.603	0.803	0.920	**0.921**
Pr@20	0.863	0.628	0.531	0.688	0.877	**0.877**	0.813	0.729	0.594	0.792	0.883	**0.885**
R@5	0.780	0.725	0.645	0.797	0.888	**0.889**	0.762	0.754	0.659	0.775	0.874	**0.877**
R@10	0.817	0.748	0.652	0.812	0.921	**0.922**	0.792	0.778	0.671	0.808	0.908	**0.908**
R@15	0.850	0.782	0.679	0.842	0.941	**0.942**	0.835	0.783	0.679	0.825	0.927	**0.927**
R@20	0.873	0.789	0.672	0.852	0.960	**0.961**	0.867	0.785	0.711	0.837	0.949	**0.949**

- лучше использовать все модальности
- биграммы и категории выигрывают у асессоров
- авторы и комментаторы наименее важны

Влияние регуляризаторов на качество поиска

Все регуляризаторы и модальности, 3 уровня, оптимальное $|T|$

Регуляризаторы: Decorrelation, Θ-sparsing, Φ-smoothing, Hierarchy

	Habrahabr					**TechCrunch**				
	нет	D	DΘ	DΘΦ	DΘΦH	нет	D	DΘ	DΘΦ	DΘΦH
Pr@5	0.628	0.772	0.771	0.865	**0.872**	0.652	0.777	0.779	0.879	**0.893**
Pr@10	0.653	0.781	0.812	0.883	**0.915**	0.679	0.788	0.819	0.895	**0.922**
Pr@15	0.642	0.785	0.792	0.891	**0.895**	0.669	0.791	0.798	0.901	**0.921**
Pr@20	0.643	0.771	0.783	0.875	**0.877**	0.673	0.775	0.792	**0.892**	0.885
R@5	0.692	0.820	0.805	0.875	**0.889**	0.673	0.825	0.812	0.869	**0.877**
R@10	0.714	0.831	0.834	0.905	**0.922**	0.685	0.856	0.845	0.881	**0.908**
R@15	0.725	0.847	0.867	0.921	**0.942**	0.712	0.877	0.869	0.912	**0.927**
R@20	0.735	0.873	0.891	0.943	**0.961**	0.723	0.892	0.895	0.934	**0.949**

- Лучше использовать все регуляризаторы
- В моделях PLSA, LDA регуляризация слишком слабая

Выводы по результатам экспериментов

- Регуляризаторы, улучшающие интерпретируемость модели, повышают также и качество поиска
- Иерархия улучшает качество поиска (в основном точность) благодаря постепенному сужению области поиска
- Подбор траектории регуляризации и оптимизация коэффициентов регуляризации влияет на качество поиска
- При тщательной оптимизации тематический поиск превосходит как асессоров, так и конкурирующие модели
- Асессорские данные относятся не к темам, а к коллекции; поэтому с их помощью можно оценивать новые модели
- Небольших асессорских данных хватает для оценивания тематических моделей, которые обучаются *без учителя*

Гипотеза условной независимости

Основное вероятностное допущение тематической модели:

$$\text{гипотеза } H_0: \quad p(w|d,t) = p(w|t), \quad d \in D, t \in T,$$

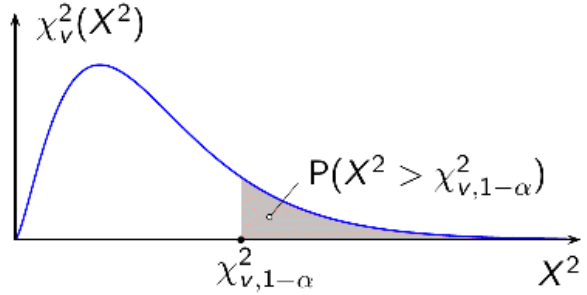
$p(w|t) = \phi_{wt} = \frac{n_{wt}}{n_t}$ — вероятностная порождающая модель,
$p(w|d,t) = \frac{n_{dwt}}{n_{td}}$ — эмпирическое распределение по n_{td} словам.

Критерий согласия хи-квадрат для проверки гипотезы H_0:

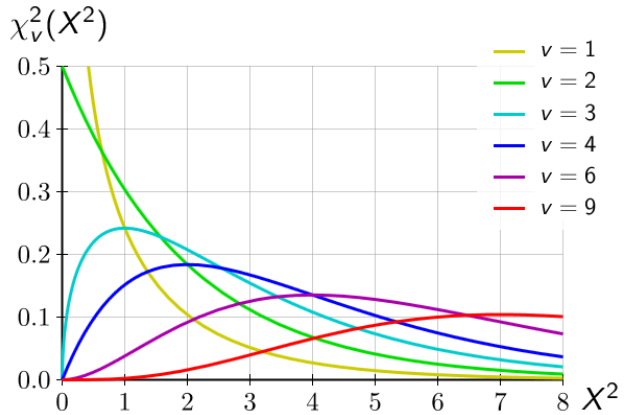
1. *Статистика хи-квадрат*:
$$X^2(d,t) = n_{dt} \sum_{w:\ n_{wt}>0} \frac{\big(p(w|d,t) - p(w|t)\big)^2}{p(w|t)}.$$
2. $X^2(d,t) \sim \chi^2$-распределение с $v = |W| - 1$ степенями свободы.
3. Если $X^2(d,t) > \chi^2_{v,1-\alpha}$, то гипотеза H_0 отвергается.

Стандартная схема проверки статистической гипотезы

Если $X^2 > \chi^2_{v,1-\alpha}$,
то гипотеза H_0 отвергается
при уровне значимости α.



Статистика X^2 распределена
как χ^2_v лишь асимптотически
и лишь при условиях
1) $n_{dt} \geqslant 50$,
2) $n_{dt} p(w|t) \geqslant 5, \ \forall w \in W$.



Проблема 1: если распределения $p(w|t)$ сильно разрежены, то асимптотика χ^2_v неприменима к статистике X^2.

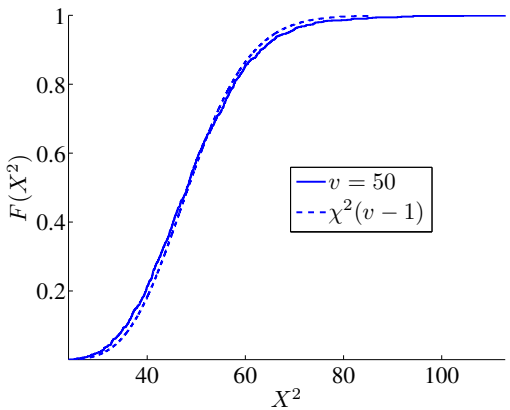
Проблема 2: распределение статистики может зависеть от n_{dt}.

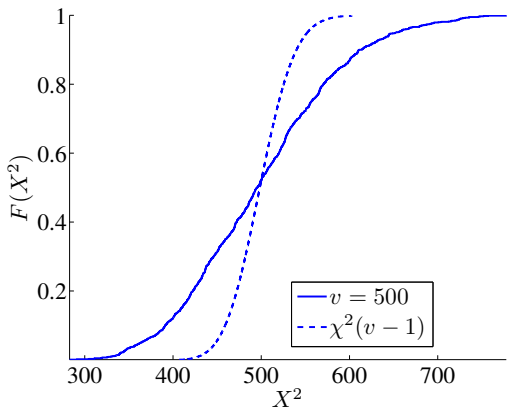
Проблема разреженности распределения

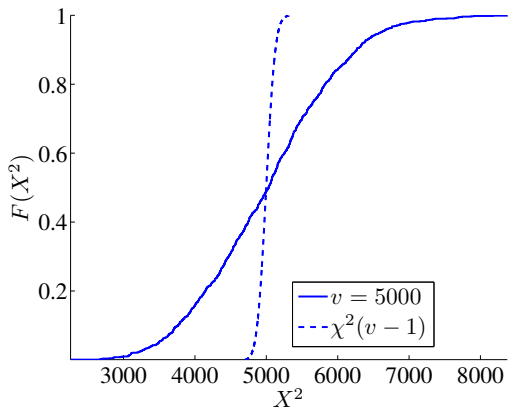
Закон Ципфа $p(i) = Ai^{-s}$ — распределение частот слов в языке по их номерам i в порядке убывания частот,
A — нормировочный множитель, s — параметр, обычно $s \approx 1$.

the	of	to	a	and	in	said	for	that	was	on	he
6.49	2.80	2.60	2.39	2.32	2.27	1.35	0.97	0.93	0.79	0.78	0.67

Эмпирические функции распределения статистики X^2,
$K = 1000$ выборок, $n = 100$, $s = 1$, $v = |W| \in \{50, 500, 5000\}$:







Чем выше разреженность, тем больше расхождение с χ^2.

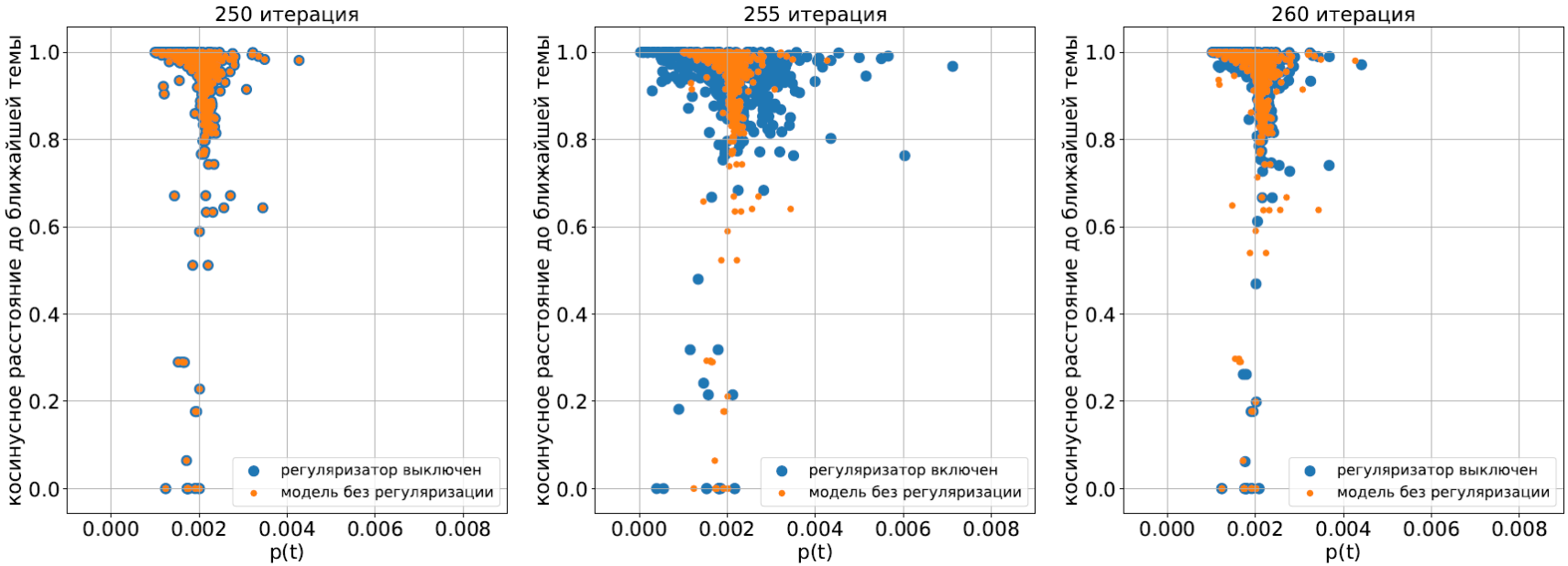
Открытые проблемы

- Можно ли эффективно вычислять эмпирическое распределение статистики хи-квадрат в EM-алгоритме?
- Можно ли придумать другую статистику, распределение которой выводится аналитически при сколь угодно разреженных распределениях?

Проблема малых тем и тем-дубликатов

Эксперимент на коллекции postnauka.ru

- Самой модели не выгодно производить малые темы!
- Регуляризатор отбора тем плохо устраняет дубликаты!

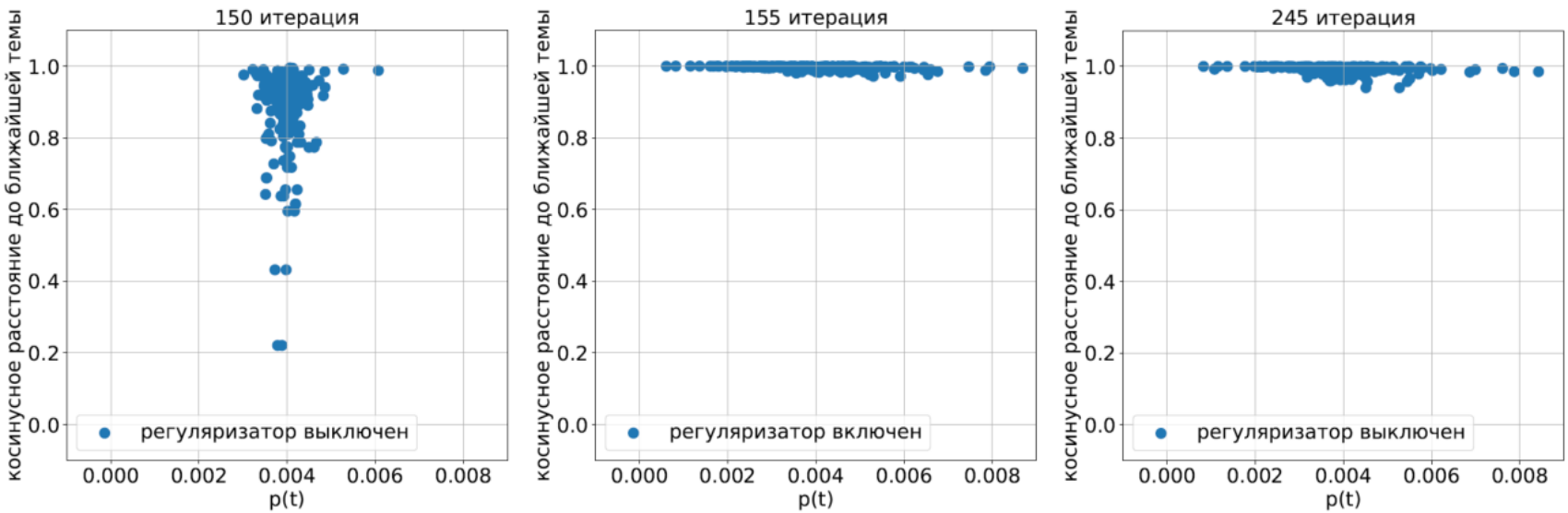


Г.Фоминская. Выявление тем-дубликатов в тематических моделях. Курсовая работа, ВМК МГУ, 2018.

Проблема малых тем и тем-дубликатов

Эксперимент на коллекции postnauka.ru

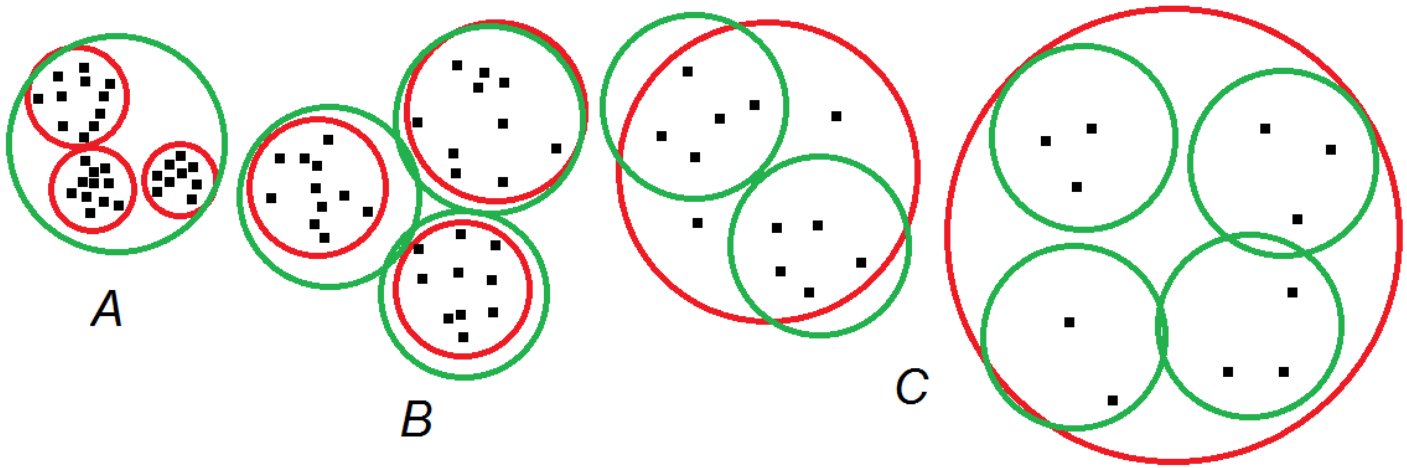
- Регуляризатор декоррелирования удаляет дубликаты лучше!
- Заодно он усиливает разброс тем по их мощности $p(t)$



Г.Фоминская. Выявление тем-дубликатов в тематических моделях. Курсовая работа, ВМК МГУ, 2018.

Проблема расщепления и слияния тем

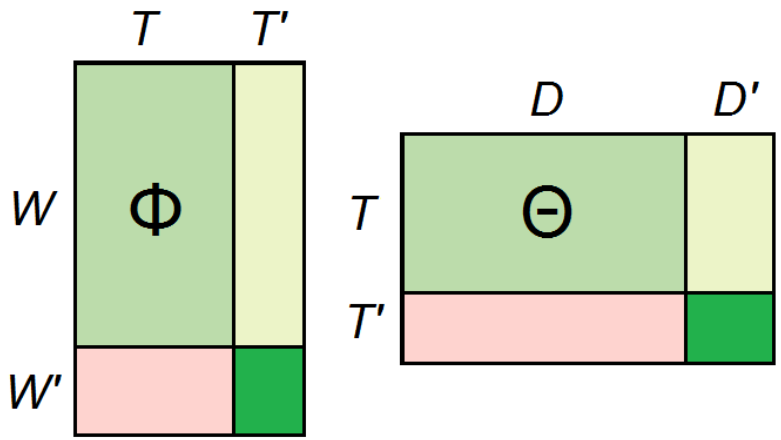
- Тематические модели стремятся выравнивать темы по их мощности (красные кластеры).
- Это приводит к появлению тем-дубликатов (A) и семантически разнородных тем (C).



- Как выровнять темы по *радиусу семантической однородности* — среднему расстоянию от $p(w|t)$ до $p(w|d,t)$?

Открытые проблемы при создании новых тем

- добавление пакета документов D' к коллекции D
- в словарь W добавляются новые слова W'
- множество тем T наращивается новыми темами T'



- сколько новых тем $|T'|$ создать?
- как инициализировать новые блоки в матрицах?
- какие документы $d \in D'$ содержат новые темы?

Вместо заключения. Алгоритм НИР

Итеративно повторять, в произвольном порядке:

- погружение в современную научную литературу
- поиск противоречий и их аккуратная формализация
- поиск лаконичных обозначений и простых доказательств
- проверка предположений в экспериментах
- анализ простых частных или крайних случаев
- изменение постановки задачи на близкие
- аккуратное письменное изложение всего
- семинары, обсуждения, диспуты, брейн-штормы

http://bigartm.org
voron@forecsys.ru

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.
- *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST 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.
- *V.Alekseev, V.Bulatov, K.Vorontsov.* Intra-Text Coherence as a Measure of Topic Models Interpretability. Dialogue, 2018.
- *A.Belyy, M.Seleznova, A.Sholokhov, K.Vorontsov.* Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue, 2018.
- *N.Skachkov, K.Vorontsov.* Improving topic models with segmental structure of texts. Dialogue, 2018.