Тематические векторные представления текста: от «мешка слов» к моделям связного текста

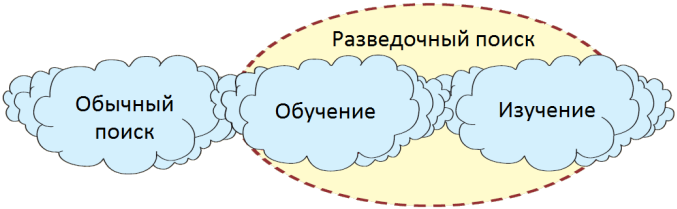
Воронцов Константин Вячеславович
(МФТИ • ФИЦ ИУ РАН • AITHEA)

Сколково.Роботикс • 16 апреля 2019

Содержание

Концепция разведочного информационного поиска (exploratory search)

- пользователь может не знать ключевых терминов предметной области
- запросом может быть текст произвольной длины или даже подборка текстов
- информационная потребность пользователя — систематизация знаний



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

Что такое «тема» в коллекции текстовых документов?

Выделение тем — первый шаг к пониманию смысла текста

- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся слов и словосочетаний

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.

О какой теме t думал автор, когда писал термин w в документе d?

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

Пример. Мультиязычная модель Википедии. Интерпретируемость тем.

216K русско-английских пар статей. Первые 10 слов и их вероятности в теме, %:

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.

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

Пример. Мультиязычная модель Википедии. Интерпретируемость тем.

216K русско-английских пар статей. Первые 10 слов и их вероятности в теме, %:

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.

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

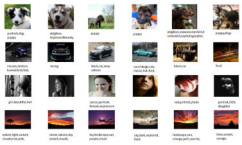
Приложения тематического моделирования

Тематическое моделирование — «мягкая кластеризация» коллекции текстов

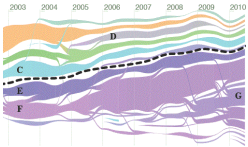
разведочный поиск в электронных библиотеках

персонализированный поиск в соцсетях

мультимодальный поиск текстов и изображений



детектирование и трекинг новостных сюжетов

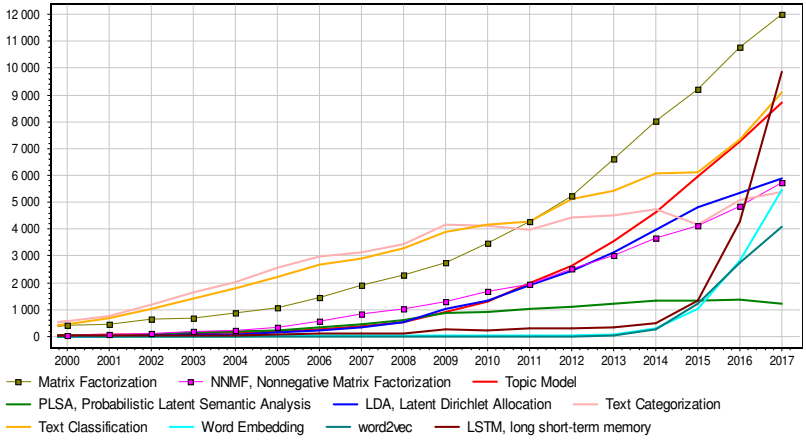


навигация по большим текстовым коллекциям

управлением диалогом в разговорном интеллекте

Тематическое моделирование и смежные области исследований

Динамика цитирования в академических публикациях, по данным Google Scholar:



Поиск тематически близких документов

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

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.

Две коллекции новостей про технологии

Habrahabr.ru

175 143 статей на русском языке

Шесть модальностей в текстах:

- 10 552 слов (униграмм)
- 742 000 биграмм
- 524 авторов статей
- 10 000 авторов комментариев
- 2546 тегов
- 123 хаба

TechCrunch.com

759 324 статей на английском языке

Четыре модальности в текстах:

- 11 523 слов (униграмм)
- 1.2 млн. биграмм
- 605 авторов
- 184 категорий

Методика оценивания качества разведочного поиска

Поисковый запрос

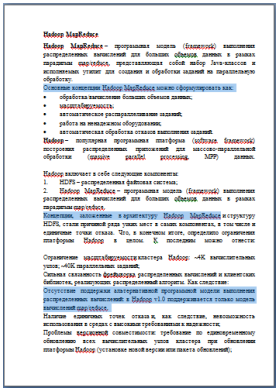
ключевые слова или фрагменты текста, одна страница A4

Поисковая выдача

документы, тематически близкие к документу-запросу

Два задания асессорам

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



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 А4):

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

Оценивание качества поиска

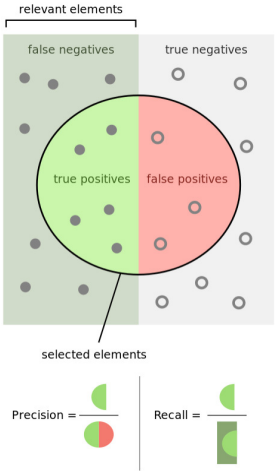
Precision — доля релевантных среди найденных
Recall — доля найденных среди релевантных

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

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

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

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



Какие модели поиска сравнивались

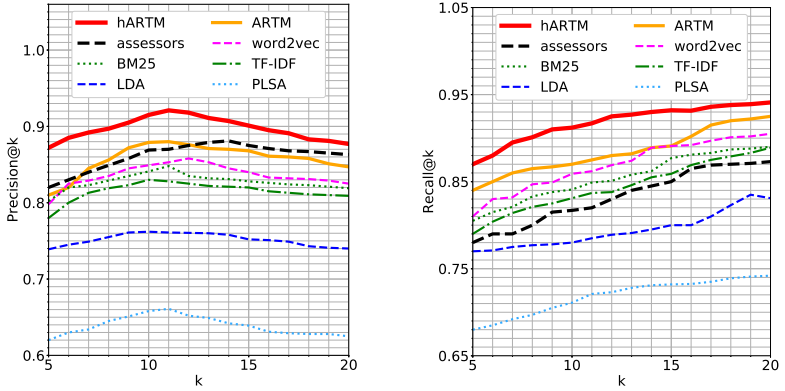
- **assessors:** результаты поиска, выполненного людьми (асессорами)
- **TF-IDF, BM25:** сравнение документов по векторам частот слов
- **word2vec:** нетематические векторные представления слов
- **PLSA:** Probabilistic Latent Semantic Analysis [T.Hofmann, 1999]
- **LDA:** Latent Dirichlet Allocation [D.Blei, A.Ng, M.Jordan, 2003]
- **ARTM:** тематическая модель с тремя регуляризаторами
- **hARTM:** двухуровневая иерархическая тематическая модель

Дополнительные критерии (регуляризаторы) в ARTM и hARTM:

- сделать темы как можно более различными
- сделать профили $p(t|d)$ как можно более разреженными
- сужать область поиска с помощью иерархических профилей $p(t|d)$

Сравнение качества поиска с асессорами и простыми моделями

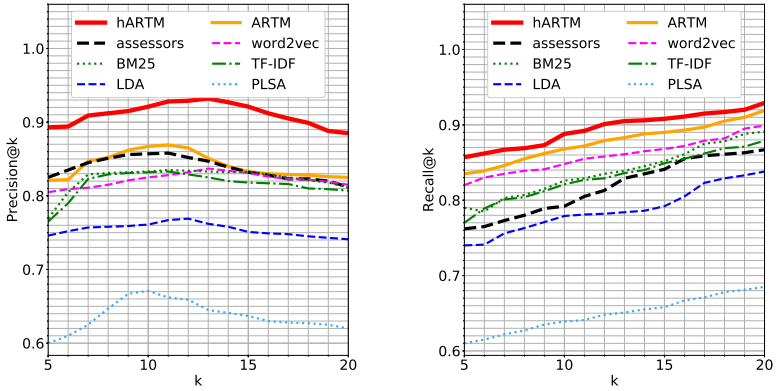
Точность и полнота по первым k позициям поисковой выдачи (Habrahabr.ru)



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

Сравнение качества поиска с асессорами и простыми моделями

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

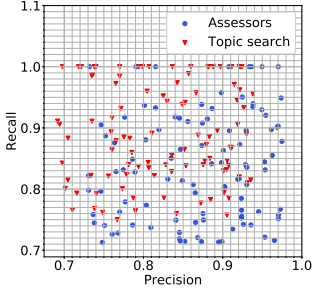


A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

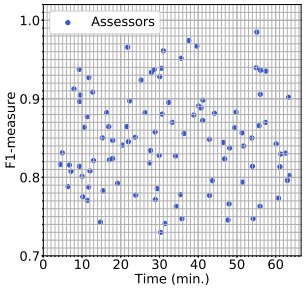
Результаты измерения точности и полноты по запросам

Точность, полнота и время поиска (100 запросов, 3 асессора на запрос, Habrahabr.ru)

точность и полнота поиска



время и F_1-мера (асессоры)



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

Влияние числа тем на качество поиска

Коллекция Habrahabr.ru

Используем 3 регуляризатора, 5 модальностей, меняем число тем

	acессоры	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

- Существует оптимальное по критерию качества поиска число тем

Влияние числа тем на качество поиска

Коллекция TechCrunch.com

Используем 3 регуляризатора, 4 модальности, меняем число тем

	acессоры	350	400	450	**475**	500
Prec@5	0.822	0.653	0.725	0.752	**0.819**	0.777
Prec@10	0.851	0.663	0.732	0.762	**0.867**	0.811
Prec@15	0.835	0.682	0.743	0.787	**0.833**	0.793
Prec@20	0.813	0.650	0.743	0.773	**0.825**	0.793
Recall@5	0.762	0.731	0.762	0.793	**0.835**	0.817
Recall@10	0.792	0.763	0.793	0.812	**0.868**	0.855
Recall@15	0.835	0.782	0.807	0.855	**0.890**	0.882
Recall@20	0.867	0.792	0.823	0.862	**0.919**	0.903

- Оптимальное число тем существенно зависит от коллекции

Влияние комбинаций регуляризаторов на качество поиска

Три регуляризатора: 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 качество поиска в явном виде

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

Математическая постановка задачи тематического моделирования

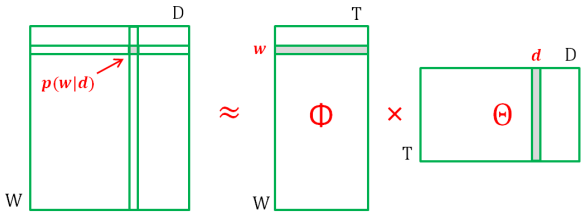
Дано: коллекция текстовых документов D, словарь слов или *термов* W

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

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

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

Это задача стохастического матричного разложения, T — заданное число тем:



PLSA — Probabilistic Latent Semantic Analysis [T.Hofmann, 1999]

Максимизация log-правдоподобия при $\phi_{wt} \geqslant 0$, $\theta_{td} \geqslant 0$, $\sum_w \phi_{wt} = 1$, $\sum_t \theta_{td} = 1$:

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

EM-алгоритм: метод простой итерации для системы уравнений

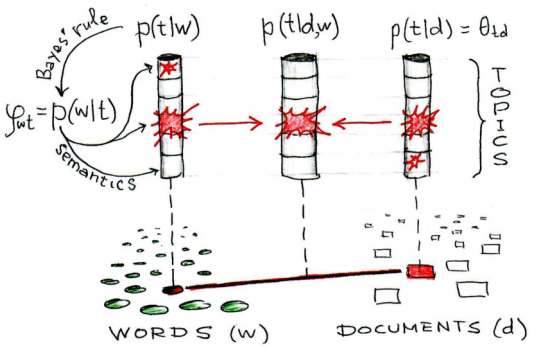
$$\text{E-шаг:} \quad p_{tdw} = p(t|d,w) = \underset{t\in T}{\text{norm}}\big(\phi_{wt}\theta_{td}\big)$$

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

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

Тематические векторные представления слов и документов

- Коллекция текстов — это двудольный граф с рёбрами (d, w)
- Слово w встречается в документе d потому, что у них есть общие темы t
- Темы интерпретируются благодаря распределению слов $p(w|t) = p(t|w)\frac{p(w)}{p(t)}$



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 решения системы уравнений

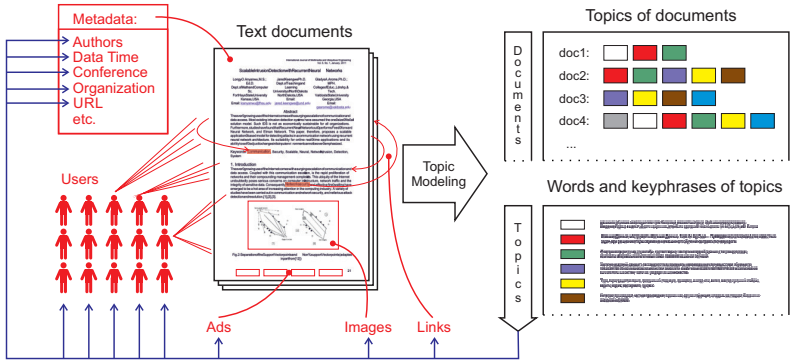
$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

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

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

Задачи мультимодального тематического моделирования

Темы определяют распределения термов различных *модальностей* $p(w|t)$:
$p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{категория}|t)$, $p(\text{класс}|t)$, $p(\text{тег}|t)$, $p(\text{ссылка}|t)$,
$p(\text{баннер}|t)$, $p(\text{элемент\_изображения}|t)$, $p(\text{пользователь}|t), \ldots$



Мультимодальная ARTM

Максимизация log-правдоподобий модальностей со словарями термов W^m, $m \in M$:

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

EM-алгоритм: метод простой итерации для решения системы уравнений

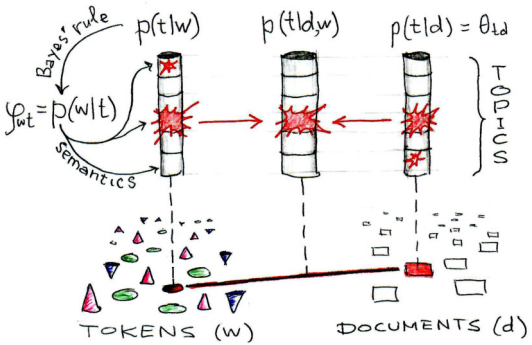
$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right)$$

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

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

Мультимодальные тематические векторные представления

- Документы содержат слова и термы других *модальностей*
- Примеры модальностей: авторы, время, теги, пользователи,...
- Через темы смыслы слов передаются другим модальностям



Пример. Модальность n-грамм улучшает качество тем

Коллекция 1000 статей конференций ММРО, ИОИ на русском языке

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

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

- 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

		$T = 50$		$T = 200$	
	проц.	минут	перплексия	минут	перплексия
BigARTM	1	42	5117	83	3347
BigARTM async	1	25	5131	53	3362
VowpalWabbit	1	50	5413	154	3960
Gensim	1	142	4945	637	3241
BigARTM	4	12	5216	26	3520
BigARTM async	4	7	5353	16	3634
Gensim	4	88	5311	315	3583
BigARTM	8	8	5648	15	3929
BigARTM async	8	5	6220	10	4309
Gensim	8	88	6344	288	4263

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

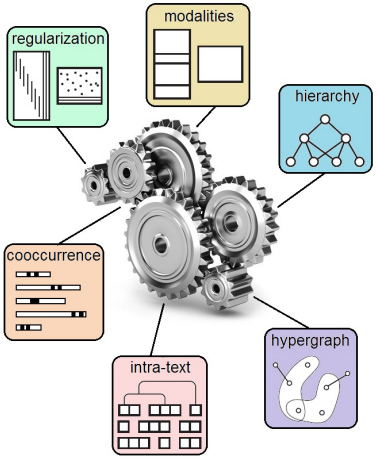
-- стандартизуемые этапы

Ключевые механизмы BigARTM

Благодаря ARTM, эти механизмы можно комбинировать в любых сочетаниях:

1. регуляризация
2. модальности
3. иерархия тем
4. парная встречаемость термов
5. гиперграфы транзакций
6. потекстовая векторная обработка

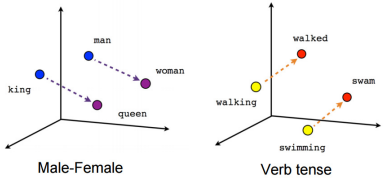
Новые механизмы позволяют учитывать порядок слов в обход гипотезы «мешка слов»



Дистрибутивная гипотеза и семантические векторные представления 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].

Задача: найти для каждого слова w вектор $x_w \in \mathbb{R}^T$ так, чтобы близкие по смыслу слова имели близкие векторы.



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. JAIR, 2010.

Модели векторных представлений для текстов и графов

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

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

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

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

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

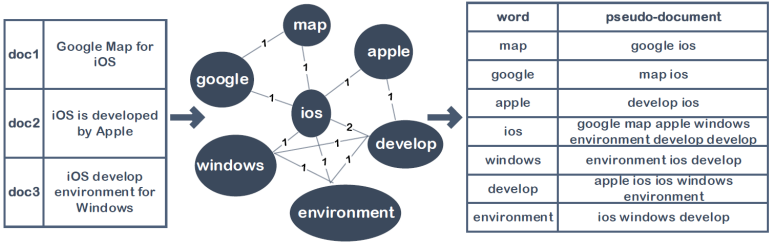
graph2vec: более общие векторные представления на графах
A.Narayanan et al. Graph2vec: learning distributed representations of graphs. 2017.

StarSpace: векторные представления 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

Формализация дистрибутивной гипотезы в тематическом моделировании

Модель сети слов WNTM моделирует 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.



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 (Word Network Topic Model) 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:

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

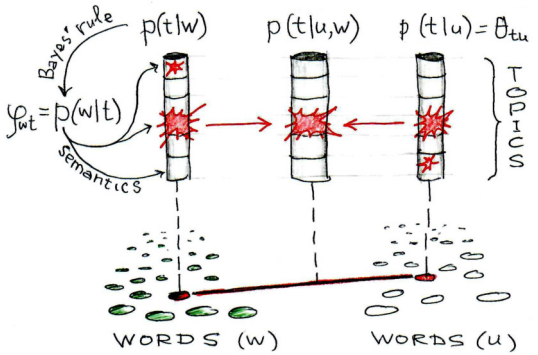
где n_{uw} — совстречаемость слов u, w (кстати, $n_{uw} = n_{wu}$).

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.

Интерпретируемые векторные представления на основе совстречаемости слов

- Идея *дистрибутивной семантики*: “Words that occur in the same contexts tend to have similar meanings” [Harris, 1954].
- Слово индуцирует псевдо-документ всех его контекстов



word2vec и ARTM на задачах аналогии слов

Два подхода к синтезу векторных представлений слов:

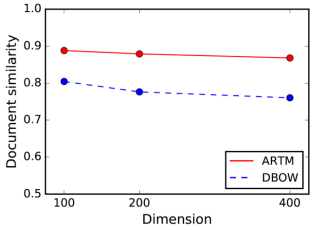
- **ARTM**: интерпретируемые разреженные компоненты
- **word2vec**: интерпретируемые векторные операции

Операция	Результат ARTM	Результат word2vec
king – boy + girl	*queen*, princess, lord, prince	*queen*, princess, regnant, kings
moscow – russia + spain	*madrid*, barcelona, aires, buenos	*madrid*, barcelona, valladolid, malaga
india – russia + ruble	*rupee*, birbhum, pradesh, madhaya	*rupee*, rupiah, devalued, debased
cars – car + computer	*computers*, software, servers, implementations	*computers*, software, hardware, microcomputers

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

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.

Транзакционные данные

Выборка может содержать не только пары (d, w), но также тройки, четвёрки, ..., n-ки элементов разных модальностей.

Примеры:

- **Данные социальной сети:**
  (d, u, w) — пользователь u записал слово w в блоге d
- **Данные сети интернет-рекламы:**
  (u, d, b) — пользователь u кликнул баннер b на странице d
- **Данные финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g

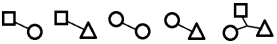
Задача: по выборке рёбер гиперграфа выявить латентные темы его вершин.

Тематическая модель гиперграфа: определения и обозначения

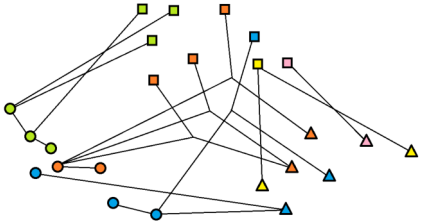
$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:



T — множество тем:



X^k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x) состоит из вершины-контейнера $d \in V$ и множества вершин $x \subset V$,
n_{dx} — число вхождений ребра (d, x) в выборку X^k
$p(d, x)$ — неизвестное распределение на рёбрах типа k

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k:

$$p(x|d) = \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{vt},$$

$\theta_{td} = p(t|d)$ — тематика контейнера не зависит от типа ребра k
$\phi_{vt} = p(v|t)$ — распределение термов модальности v в теме t

Задача максимизации log-правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in X^k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{vt} \to \max_{\Phi, \Theta},$$

$$\phi_{vt} \geqslant 0, \quad \sum_{v \in V^m} \phi_{vt} = 1; \qquad \theta_{td} \geqslant 0, \quad \sum_{t \in T} \theta_{td} = 1;$$

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного log-правдоподобия:

$$\sum_{k\in K}\tau_k \sum_{(d,x)\in X^k} n_{dx} \ln \sum_{t\in T}\theta_{td}\prod_{v\in x}\phi_{vt} \;+\; R(\Phi,\Theta) \to \max_{\Phi,\Theta}.$$

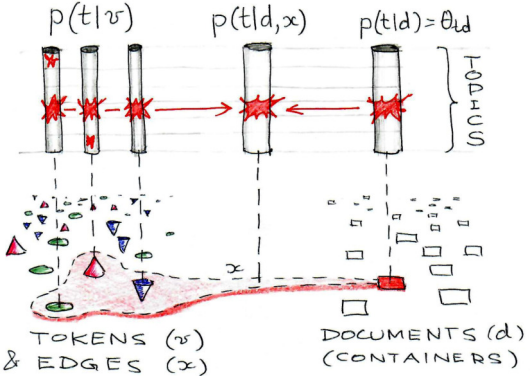
EM-алгоритм: метод простой итерации для решения системы уравнений

$$\text{E-шаг:}\quad p_{tdx} = \underset{t\in T}{\operatorname{norm}}\Big(\theta_{td}\prod_{v\in x}\phi_{vt}\Big)$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{vt} = \underset{v\in V^m}{\operatorname{norm}}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)} \big[v\in x\big] n_{dx} p_{tdx} + \phi_{vt}\frac{\partial R}{\partial \phi_{vt}}\Big) \\ \theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)} n_{dx} p_{tdx} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

Интерпретируемые эмбедингы транзакционных данных

- *Гиперграф* — это система подмножеств вершин-термов
- Транзакция = подмножество термов = ребро гиперграфа
- Транзакция тем более вероятна, чем больше общих тем имеют её термы



Модели предложений и коротких текстов TwitterLDA, senLDA

S_d — множество предложений документа d

n_{sw} — сколько раз терм w встречается в предложении s

Тематическая модель предложения s:

$$p(s|d) = \sum_{t\in T} p(t|d) \prod_{w\in s} p(w|t)^{n_{sw}} = \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}}$$

Максимизация регуляризованного log-правдоподобия

$$\sum_{d\in D} \sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

это частный случай гиперграфовой модели, предложения являются гипер-рёбрами.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim et al. Comparing Twitter and traditional media using topic models. ECIR 2011.

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Гиперграфовые тематические модели языка

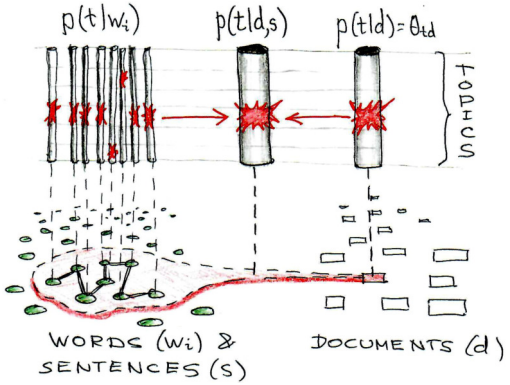
Что ещё может быть ребром гиперграфа?

Любое подмножество связанных по смыслу термов, порождаемых общей темой.

- предложение
- синтагма, ветка синтаксического дерева
- именная группа
- факт «объект, субъект, действие»
- пары термов в одном или соседних предложениях, связанных тезаурусными отношениями: синонимы, гипоним–гипероним, мероним–холоним
- лексическая цепочка
- текст сообщения и его автор
- финансовая транзакция с текстом платёжного поручения

Интерпретируемые эмбедингы предложений

- Предложение — это наиболее семантически однородная единица языка
- Предложение = подмножество слов = ребро гиперграфа
- Предложение тем более вероятно, чем больше общих тем имеют его слова



Сегментная структура текста и пост-обработка E-шага

Документ $d = \{w_1, \ldots, w_{n_d}\}$, n_d — длина документа d

Матрица *тематики слов в документах* $p(t|d, w_i)$ размера $T \times n_d$:

Регуляризация E-шага

Трёхмерная матрица $\Pi = \left(p_{tdw} = p(t|d,w)\right)_{T\times D\times W}$

Максимизация log правдоподобия с регуляризаторами R и $\tilde{R}$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Pi(\Phi,\Theta)) + \tilde{R}(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}.$$

EM-алгоритм: метод простой итерации для системы уравнений

E-шаг:
$$\begin{cases} p_{tdw} = \underset{t\in T}{\operatorname{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right) \end{cases}$$

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

Пример. Тематическая модель сегментированного текста

S_d — множество микро-сегментов документа d
n_{sw} — число вхождений терма w в сегмент s длины n_s
Тематика сегмента $s \in S_d$ — среднее по всем его термам:

$$p_{tds} \equiv p(t|d,s) = \frac{1}{n_s} \sum_{w \in s} n_{sw} p_{tdw}.$$

Кросс-энтропийный регуляризатор разреживания $p(t|d,s)$:

$$R(\Pi) = -\sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T} \ln \sum_{w \in s} n_{sw} p_{tdw} \to \max.$$

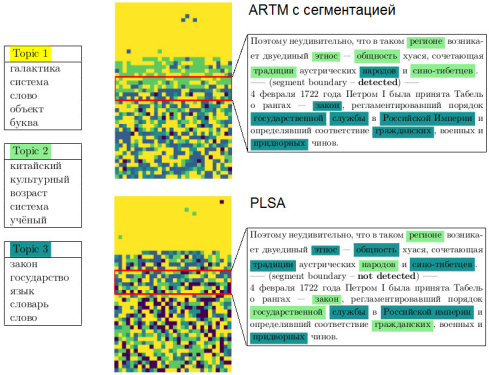
Формула регуляризованного E-шага:

$$\tilde{p}_{tdw} = p_{tdw} \left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s} \left(\frac{1}{p_{tds}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zds}}\right)\right).$$

Интерпретация: если $p_{tds} < \frac{1}{|T|}$, то p_{tdw} уменьшатся $\forall\, w \in s$.
Тематика сегмента концентрируется в небольшом числе тем.

Пример. Эксперимент на полусинтетической коллекции

Сегментация текстов, склеенных из фрагментов монотематичных статей научно-просветительского портала postnauka.ru



N.Skachkov, K.Vorontsov. Improving topic models with segmental structure of texts. Dialogue, 2018.

Резюме

- Тематическое моделирование — ключевой механизм разведочного поиска
- Стандартное тематическое моделирование (PLSA, LDA):
  матричное разложение + гипотеза «мешка слов»
- Аддитивная регуляризация (ARTM) эксплуатирует неединственность решения, чтобы строить модели с заданными свойствами
- BigARTM — эффективная реализация этого подхода
- Новые механизмы ARTM, выходящие за рамки «мешка слов»:
  — автоматическое выделение терминов (устойчивых выражений),
  — тематические модели дистрибутивной семантики,
  — гиперграфовые модели лексических и семантических связей,
  — обработка текста как последовательности тематических векторов термов

Литература

- *К.В.Воронцов.* Обзор вероятностных тематических моделей. 2018. – 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.