

Тематическое моделирование текстовых коллекций и транзакционные данные

Константин Воронцов, МФТИ

8 апреля 2021

1 Теория

- Задача тематического моделирования
- ARTM: больше, чем LDA
- BigARTM: быстрее, выше, сильнее

2 Тематизация текстовых данных

- Тематический поиск
- Мониторинг новостных потоков
- Рубрикация интенгов и сегментация диалогов

3 Тематизация транзакционных данных

- Транзакции физических лиц
- Транзакции юридических лиц

Что такое «тема» в коллекции текстовых документов?

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

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

Имея коллекцию текстовых документов, хотим узнать:

- из каких тем состоит коллекция,
 $p(t)$ — вероятность (доля) темы t в коллекции;
- из каких тем состоит каждый документ,
 $p(t|d)$ — вероятность (доля) темы t в документе d ;
- из каких слов или терминов состоит каждая тема,
 $p(w|t)$ — вероятность (доля) слова w в теме t .

Пример. Мультиязычная модель Википедии

216 175 русско-английских пар статей.

Первые 10 слов и их частоты $p(w|t)$ 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

Ассессор 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 модель Википедии

216 175 русско-английских пар статей.

Первые 10 слов и их частоты $p(w|t)$ 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

Ассессор 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.

Некоторые приложения тематического моделирования

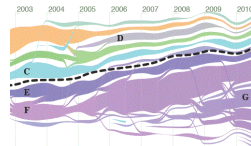
разведочный поиск в
электронных библиотеках



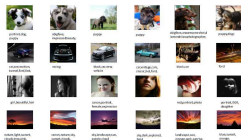
поиск тематического
контента в соцсетях



выявление и отслеживание
цепочек новостей



мультимодальный поиск
текстов и изображений



анализ банковских
транзакционных данных



управлением диалогом в
разговорном интеллекте



Пусть

- W — конечное множество *термов* (слов, терминов)
- D — конечное множество текстовых документов
- T — конечное множество тем
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- каждый терм w в документе d связан с некоторой темой t
- $D \times W \times T$ — дискретное вероятностное пространство
- 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:

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

Задача восстановления $p(w|t)$ и $p(t|d)$ по коллекции текстов

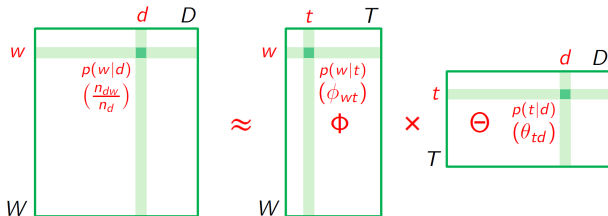
Дано: коллекция текстовых документов

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

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

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

Это задача стохастического матричного разложения:



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

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

эквивалентна максимизации функционала

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

при ограничениях неотрицательности и нормировки

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

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Задача максимизации логарифма правдоподобия:

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

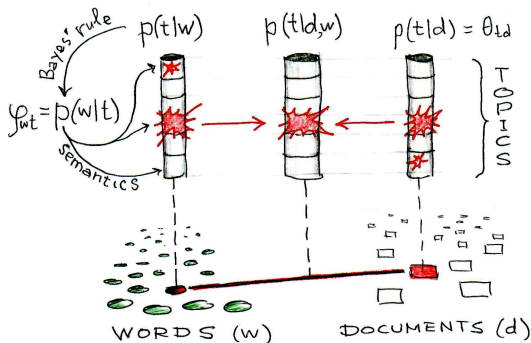
EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdw} = p(t|d, w)$:

$$\begin{array}{l} \text{E-шаг:} \\ \text{M-шаг:} \end{array} \left\{ \begin{array}{l} p_{tdw} = \mathop{\text{norm}}_{t \in T}(\phi_{wt} \theta_{td}) \\ \phi_{wt} = \mathop{\text{norm}}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} \right) \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} \right) \end{array} \right.$$

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

Интерпретируемые эмбединги слов и документов

- Коллекция текстов — двудольный граф с рёбрами (d, w)
- Слово w встречается в d , когда у них есть общие темы
- Интерпретируемость тем возникает благодаря $p(w|t)$



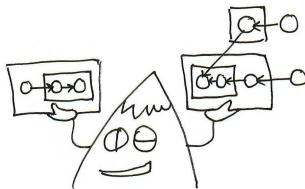
Обобщение №1

Проблема

В PLSA слишком много параметров, возможно переобучение. Надо наложить ограничения на столбцы матриц Φ и Θ . Желательно так, чтобы они стали более разреженными.

Решение

Модель латентного размещения Дирихле (2003).



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

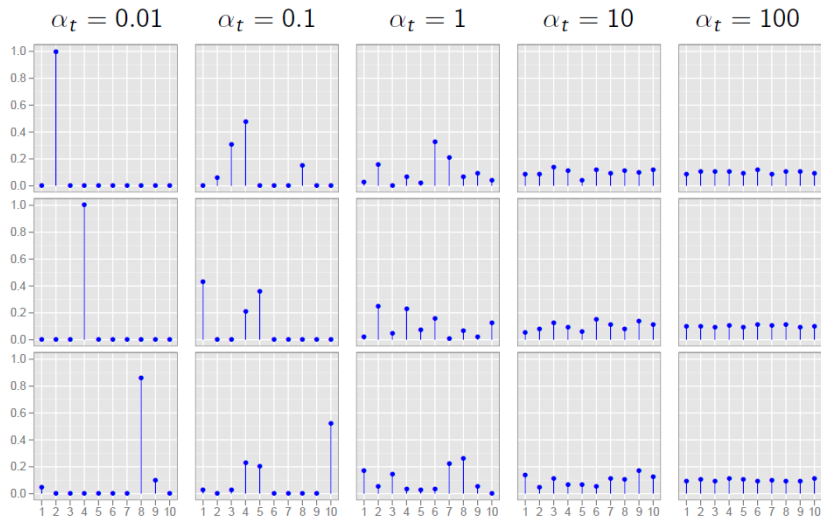
Регуляризованный EM-алгоритм: модель LDA

Задача максимизации апостериорной вероятности:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\text{ln правдоподобия } \mathcal{L}(\Phi, \Theta)} + \underbrace{\sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}}_{\text{критерий регуляризации } R(\Phi, \Theta)} \rightarrow \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \mathop{\text{norm}}_{t \in T} (\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \mathop{\text{norm}}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \beta_w - 1 \right) \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} + \alpha_t - 1 \right) \end{cases} \end{cases}$$

Пример. Выборки из трёх 10-мерных векторов $\theta \sim \text{Dir}(\theta|\alpha)$ 

Обобщение №2

Проблема

LDA — слишком простой и слабый регуляризатор.

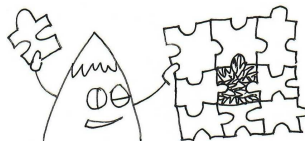
LDA не позволяет комбинировать разные регуляризаторы.

Решение

Ввести произвольный регуляризатор $R(\Phi, \Theta)$

или сумму регуляризаторов $R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$.

Аддитивность → модульный подход к моделированию.



ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация \log правдоподобия с регуляризатором R :

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг: } \begin{cases} p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}) \end{cases}$$

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

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

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

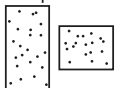
Регуляризаторы для улучшения интерпретируемости тем

background

Сглаживание фоновых тем $B \subset T$:

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

sparse

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

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

decorrelated



Декоррелирование для повышения различности тем:

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

interpretable

Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

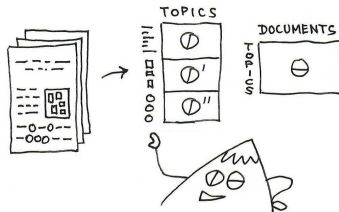
Обобщение №3

Проблема

Есть много задач, в которых документы содержат не только слова, но и элементы других модальностей

Решение

Ввести для каждой модальности свою матрицу Φ и максимизировать свой критерий лог-правдоподобия



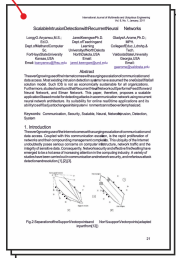
Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,

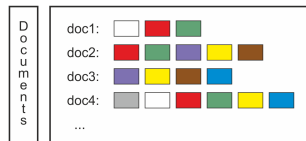
Metadata:
Authors
Data Time
Conference
Organization
URL
etc.

Text documents



Topic
Modeling

Topics of documents



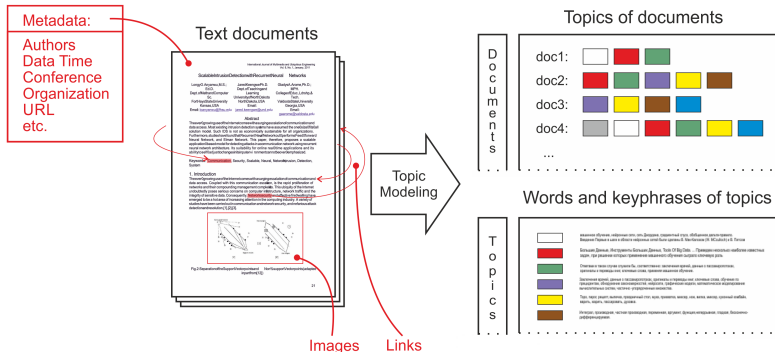
Words and keyphrases of topics



Мультимодальная тематическая модель

Тема может порождать термины различных модальностей:

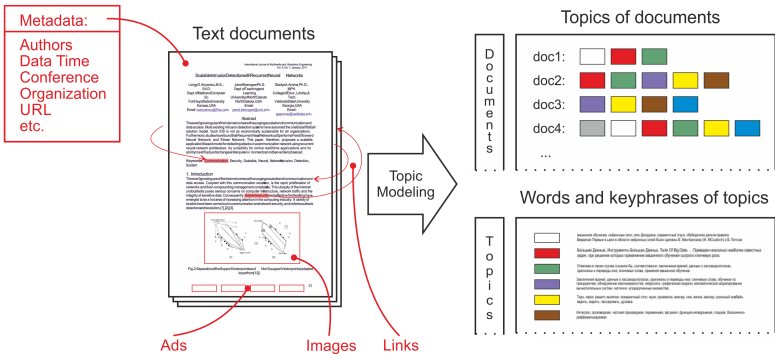
$p(\text{слово} | t)$, $p(\text{n-грамма} | t)$, $p(\text{автор} | t)$, $p(\text{время} | t)$, $p(\text{источник} | t)$,
 $p(\text{объект} | t)$, $p(\text{ссылка} | t)$,



Мультимодальная тематическая модель

Тема может порождать термины различных модальностей:

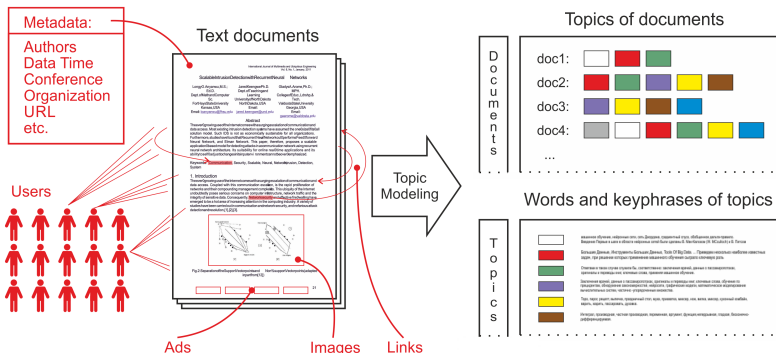
$p(\text{слово} | t)$, $p(n\text{-грамма} | t)$, $p(\text{автор} | t)$, $p(\text{время} | t)$, $p(\text{источник} | t)$,
 $p(\text{объект} | t)$, $p(\text{ссылка} | t)$, $p(\text{баннер} | t)$,



Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

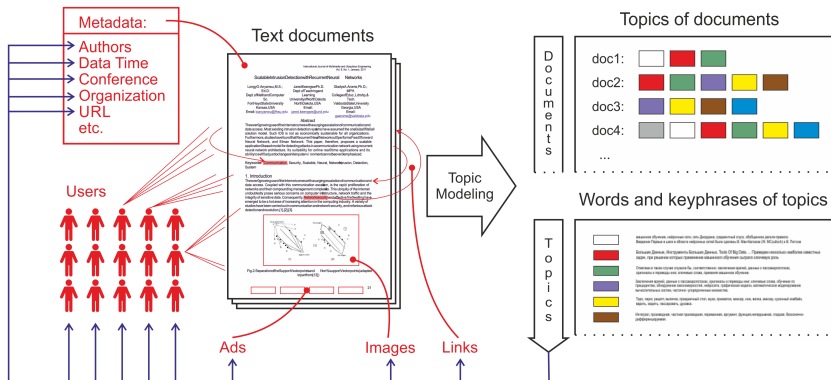
$p(\text{слово} | t)$, $p(n\text{-грамма} | t)$, $p(\text{автор} | t)$, $p(\text{время} | t)$, $p(\text{источник} | t)$,
 $p(\text{объект} | t)$, $p(\text{ссылка} | t)$, $p(\text{баннер} | t)$, $p(\text{пользователь} | t)$



Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

$p(\text{слово} | t)$, $p(n\text{-грамма} | t)$, $p(\text{автор} | t)$, $p(\text{время} | t)$, $p(\text{источник} | t)$,
 $p(\text{объект} | t)$, $p(\text{ссылка} | t)$, $p(\text{баннер} | t)$, $p(\text{пользователь} | t)$



Мультимодальная ARTM

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

Максимизация суммы \log правдоподобий с регуляризацией:

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

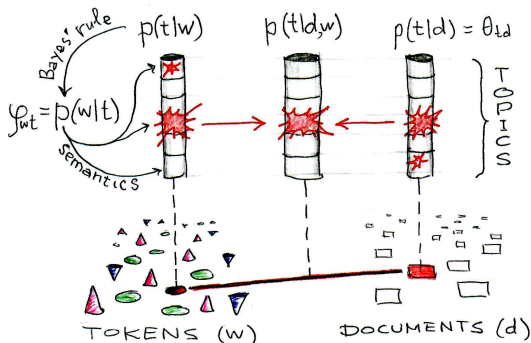
EM-алгоритм: метод простой итерации для системы уравнений

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

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

Интерпретируемые эмбединги мультимодальных документов

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



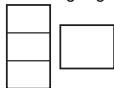
Регуляризаторы для мультимодальных тематических моделей

supervised



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

multilanguage

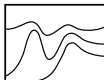


Модальность языков и регуляризация со словарём

$\pi_{uwt} = p(u|w, t)$ переводов с языка k на ℓ :

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

temporal



Темпоральные модели с модальностью времени i :

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

geospatial



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

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

Биграммы радикально улучшают интерпретируемость тем

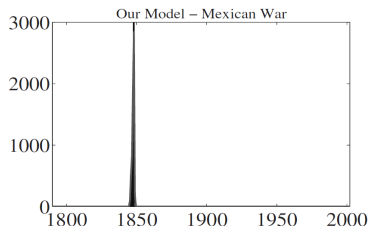
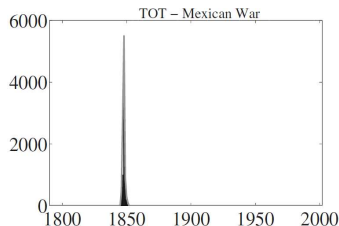
Коллекция 1000 статей конференций MMPO, ИОИ на русском

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.

Совмещение темпоральной и n -граммной модели

По коллекции выступлений президентов США



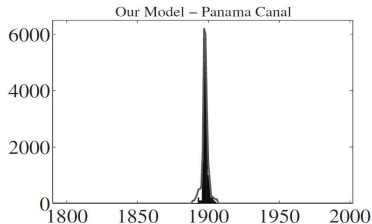
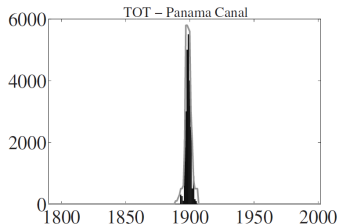
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

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

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

Совмещение темпоральной и n -граммной модели

По коллекции выступлений президентов США



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

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

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

Эксперимент по Automatic Term Extraction (ATE)

- Коллекция $|D| = 3200$ аннотаций статей NIPS (Neural Information Processing Systems), $n = 500\,000$ слов
- Ручная разметка небольшого *случайного* подмножества (2000 n -грамм) на термины / не-термины
- Train : Test = 1000 : 1000
- 7 статистических признаков из TopMine
- 2 синтаксических признака из SyntaxNet
- 3 тематических признака из BigARTM, 30 тем
- две модели классификации:
логистическая регрессия, градиентный бустинг

Владимир Полушин. Тематические модели для ранжирования рекомендаций текстового контента. Бакалаврская диссертация, ВМК МГУ, 2017.

Сравнение методов автоматического отбора терминов

Найти *как можно больше терминов* — полнота важнее точности

Группа признаков			Линейная модель			Градиентный бустинг		
Синт	Стат	Тем	AUC	Точность	Полнота	AUC	Точность	Полнота
+			0.83	0.20	0.91	0.83	0.20	0.91
	+		0.71	0.09	0.94	0.73	0.11	0.90
		+	0.92	0.32	1.00	0.95	0.32	1.00
+	+		0.88	0.22	0.91	0.88	0.24	0.91
+		+	0.91	0.36	0.91	0.95	0.34	0.99
	+	+	0.93	0.29	0.94	0.98	0.34	1.00
+	+	+	0.95	0.38	0.91	0.97	0.41	0.99

$$\boxed{\text{Стат}} < \boxed{\text{Син}} < \boxed{\text{Син+Стат}} < \boxed{\text{Тем}} < \begin{matrix} \boxed{\text{Стат+Тем}} \\ \boxed{\text{Син+Тем}} \end{matrix} < \boxed{\text{Стат+Син+Тем}}$$

- Тематические признаки существенно повышают качество
- Синтаксические признаки можно не использовать

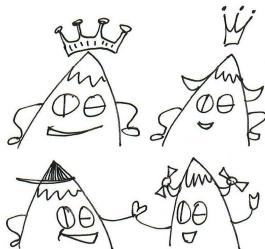
Обобщение №4

Проблема

Тематические модели формируют векторные представления (эмбединги) слов, но почему-то они не способны решать задачи семантической близости слов, как word2vec.

Решение

Понять, что такого есть в word2vec, и ввести это в ТМ.



Модели векторных представлений для текстов и графов

word2vec: эмбединги (векторные представления) 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.

BERT: эмбединги фраз и предложений от Google AI Language

J. Devlin et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018.

Недостаток: координаты векторов не интерпретируемы

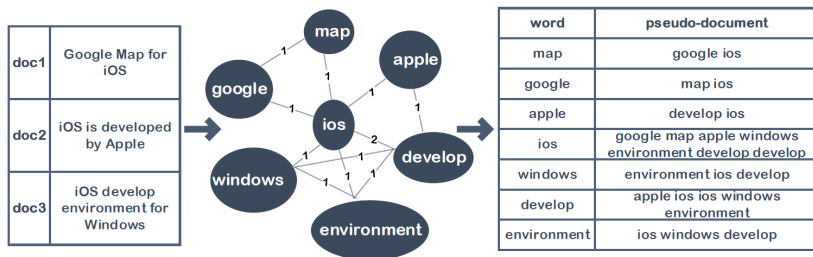
Модель сети слов WNTM для коротких текстов

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.



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:

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

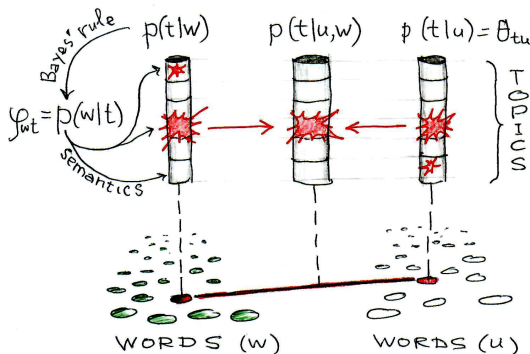
где n_{uw} — сочетаемость слов u, w (кстати, $n_{uw} = n_{wu}$).

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 эмбединги сочетаемости 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

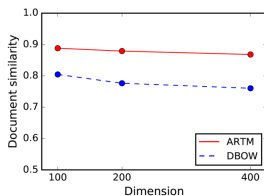
Операция	Результат ARTM	Результат word2vec
king – boy + girl	<i>queen, princess, lord, prince</i>	<i>queen, princess, regnant, kings</i>
moscow – russia + spain	<i>madrid, barcelona, aires, buenos</i>	<i>madrid, barcelona, valladolid, malaga</i>
india – russia + ruble	<i>rupee, birbhum, pradesh, madhaya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

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: 20К троек статей:

⟨ статья А, схожая статья В, непохожая статья С ⟩



- обучение по 1М текстов статей ArXiv
- тестирование на триплетях 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.

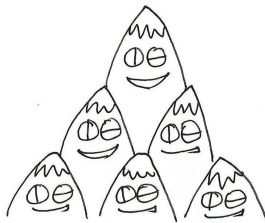
Обобщение №5

Проблема

Невозможно определить оптимальное число тем.
Хотелось бы разделять темы на подтемы иерархически.
Придумано много иерархических моделей, но они либо ограниченные, либо тормозные, либо замороженные.

Решение

Придумать что-то радикально простое



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:

$$\sum_{t \in T} n_t \text{KL}_w \left(p(w|t) \parallel \sum_{s \in S} p(w|s)p(s|t) \right) \rightarrow \min_{\Phi, \Psi}$$

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

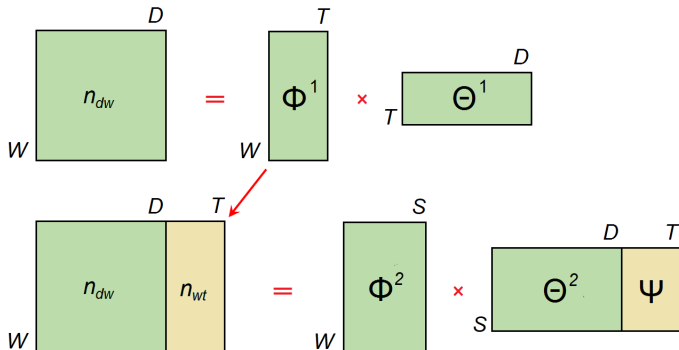
Родительская $\Phi^P \approx \Phi\Psi$, отсюда регуляризатор матрицы Φ :

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

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

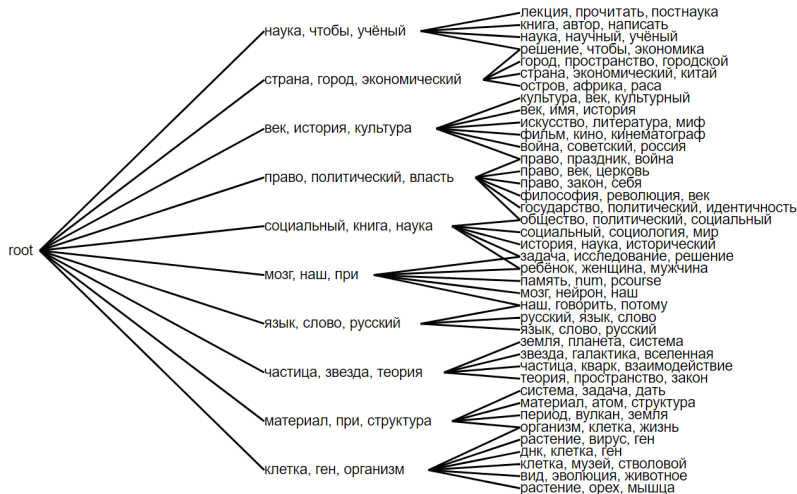
Построение второго уровня иерархии с подтемами S

В коллекцию добавляются $|T|$ псевдодокументов родительских тем с частотами термов $n_{wt} = \tau n_t \phi_{wt}$, $t \in T$



Матрица связей тем с подтемами $\Psi = (p(s|t))$ образуется в столбцах матрицы Θ , соответствующих псевдодокументам.

Пример тематической иерархии (коллекция postnauka.ru)



Дмитрий Федоряка. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация, МФТИ, 2017.

Обобщение №6

Проблема

Исходные данные могут быть сложнее, чем парные взаимодействия (транзакции) между объектами

Решение

Тематическая модель должна описывать транзакции, состоящие из любых подмножеств объектов



Транзакционные данные

Выборка может содержать не только пары (d, w) , но также тройки, четвёрки, \dots , 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 рекомендательной системы:**
 (u, f, s) — пользователь u оценил фильм f в ситуации s
- **Данные финансовых организаций:**
 (b, s, g) — покупатель u купил у продавца s товар g

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

Тематическая модель гиперграфа: определения и обозначения

$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.

$V = V^1 \sqcup \dots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

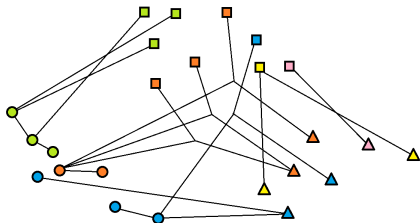
□ ○ △

K — множество типов рёбер:

□○ □△ ○○ ○△ ○△ □△

T — множество тем:

● ● ● ● ●



X^k — наблюдаемая выборка транзакций — рёбер типа k

ребро (d, x) : вершина-контейнер $d \in V$ и вершины $x \subset V$,

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:

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

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

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

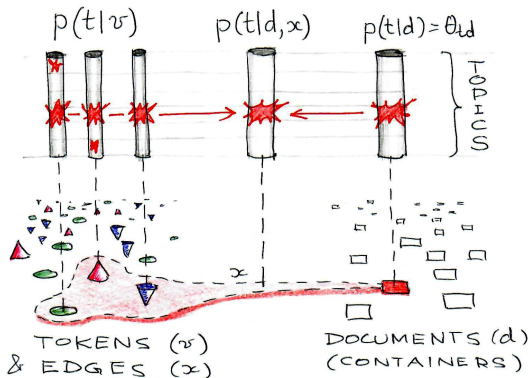
$$\sum_{k \in K} \tau_k \sum_{(d,x) \in X^k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in X} \phi_{vt} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdx} = p(t|d, x)$:

$$\begin{cases} \text{E-шаг:} & p_{tdx} = \operatorname{norm}_{t \in T} \left(\theta_{td} \prod_{v \in X} \phi_{vt} \right) \\ \text{M-шаг:} & \begin{cases} \phi_{vt} = \operatorname{norm}_{v \in V^m} \left(\sum_{k \in K} \tau_k \sum_{(d,x)} [v \in X] n_{dx} p_{tdx} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right) \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(\sum_{k \in K} \tau_k \sum_{(d,x)} n_{dx} p_{tdx} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases} \end{cases}$$

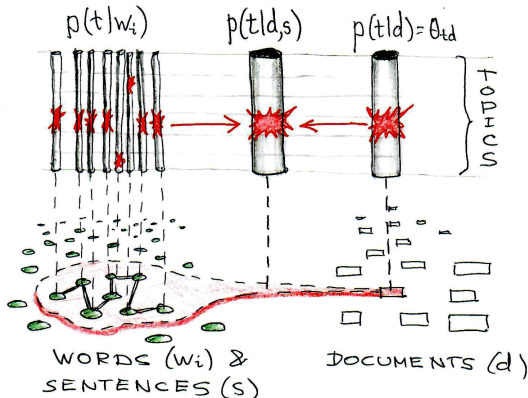
Интерпретируемые эмбединги транзакционных данных

- *Гиперграф* — множество подмножеств вершин-токенов
- Транзакция = подмножество токенов = ребро гиперграфа
- Транзакция происходит, когда токены имеют общие темы



Интерпретируемые эмбединги предложений

- Предложение — семантически однородная единица языка
- Предложение образуется из слов, имеющих общие темы
- Предложение = подмножество слов = ребро гиперграфа



Обобщение №7

Проблема

Гипотеза «мешка слов» — самое часто критикуемое допущение тематического моделирования.

Как строить модели, учитывающие порядок слов?

Решение

Преобразовать матрицу тем $p(t|d, w_i)$ с учётом локальных контекстов слов w_{i-1}, w_i, w_{i+1}



Сегментная структура текста и пост-обработка E-шага

Документ $d = \{w_1, \dots, w_{n_d}\}$, 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 $\Pi = (p_{tdw} = p(t|d, w))_{T \times D \times W}$

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

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

EM-алгоритм: метод простой итерации для системы уравнений

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

Регуляризаторы для моделирования последовательного текста

n-gram



Модели с модальностями n -грамм, коллокаций, именованных сущностей (например, TopMine)

syntax



Модели, учитывающие результаты автоматического синтаксического разбора (например, SyntaxNet)

coherence



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

sentence



Тематические модели, учитывающие границы предложений, абзацев и секций документов

segmentation



Тематические модели сегментации с автоматическим определением границ сегментов

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Онлайн-параллельный мультимодальный ARTM
- Встроенная библиотека регуляризаторов и мер качества

Сообщество:

- Открытый код <https://github.com/bigartm>
(discussion group, issue tracker, pull requests)
- Документация <http://bigartm.org>



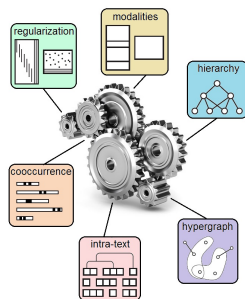
Лицензия и среда разработки:

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

Ключевые возможности библиотек BigARTM и TopicNet

BigARTM

- библиотека регуляризаторов
- мультимодальные модели
- иерархические модели
- гиперграфовые модели
- модели связности текста



TopicNet

- Перебор сценариев регуляризации для выбора моделей
- Автоматическое протоколирование экспериментов
- Построение «банка тем» из множества моделей
- Визуализация тематических моделей

V. Bulatov, E. Egorov, E. Veselova, D. Polyudova, V. Alekseev, A. Goncharov, K. Vorontsov.
TopicNet: making additive regularisation for topic modelling accessible. LREC-2020

Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

3.7М статей Википедии, 100К 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 модель порождения данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Единый регуляризованный EM-алгоритм для любых моделей и их композиций	
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

Модульность аддитивной регуляризации

Мешок регуляризаторов под каждую прикладную задачу

Выявления этнорелевантного дискурса в социальных сетях:

$$\mathcal{L} \left(\begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array} \right) + R \left(\begin{array}{c} \text{interpretable} \\ \text{[Bar chart]} \quad \text{[Scatter plot]} \end{array} \right) + R \left(\begin{array}{c} \text{n-gram} \\ \text{[Grid of boxes]} \end{array} \right) + R \left(\begin{array}{c} \text{seed words} \\ \text{[Bar chart]} \quad \text{[Box]} \end{array} \right) \rightarrow \max$$

Тематический поиск научных и научно-популярных статей:

$$\mathcal{L} \left(\begin{array}{c} \text{multimodal} \\ \text{[Stacked boxes]} \quad \text{[Box]} \end{array} \right) + R \left(\begin{array}{c} \text{interpretable} \\ \text{[Bar chart]} \quad \text{[Scatter plot]} \end{array} \right) + R \left(\begin{array}{c} \text{n-gram} \\ \text{[Grid of boxes]} \end{array} \right) + R \left(\begin{array}{c} \text{hierarchy} \\ \text{[Tree diagram]} \end{array} \right) \rightarrow \max$$

Выявление и прослеживание событий в новостном потоке:

$$\mathcal{L} \left(\begin{array}{c} \text{multimodal} \\ \text{[Stacked boxes]} \quad \text{[Box]} \end{array} \right) + R \left(\begin{array}{c} \text{interpretable} \\ \text{[Bar chart]} \quad \text{[Scatter plot]} \end{array} \right) + R \left(\begin{array}{c} \text{temporal} \\ \text{[Line graph]} \end{array} \right) + R \left(\begin{array}{c} \text{sentiment} \\ \text{[Sentiment diagram]} \end{array} \right) \rightarrow \max$$

Общий взгляд на байесовское обучение, MAP и ARTM

Байесовский вывод апостериорного распределения $p(\Omega|X)$ (обычно приближённый) ради получения точечной оценки Ω :

$$\text{Posterior}(\Omega|X, \gamma) \propto p(X|\Omega) \text{Prior}(\Omega|\gamma)$$

$$\Omega := \arg \max_{\Omega} \text{Posterior}(\Omega|X, \gamma)$$

Максимизация апостериорной вероятности (MAP)

даёт точечную оценку Ω напрямую, без вывода Posterior:

$$\Omega := \arg \max_{\Omega} (\ln p(X|\Omega) + \underbrace{\ln \text{Prior}(\Omega|\gamma)}_{R(\Omega)})$$

Многокритериальная аддитивная регуляризация (ARTM)

обобщает MAP на любые регуляризаторы и их комбинации:

$$\Omega := \arg \max_{\Omega} (\ln p(X|\Omega) + \sum_{i=1} \tau_i R_i(\Omega))$$

Две коллекции новостей про технологии

Nabrahbr.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 r morphology2

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 ассессорам

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

Поиск MapReduce

Поиск MapReduce – программа поиска (библиотека) вычислений распределенных вычислений для больших объемов данных в рамках параллельных шардов, представляющих собой набор Java-классов и исполнимых утилит для создания и обработки данных на параллельной обработке.

Основные возможности Поиск MapReduce можно сформулировать как:

- обработка вычислений больших объемов данных;
- масштабируемость;
- автоматическое распределение заданий;
- работа на неидеальных оборудовании;
- автоматическая обработка отказов вычислений заданий.

Поиск – популярная программная платформа (библиотека, библиотека) построения распределенных приложений для высокопараллельной обработки (задачи, задачи, решение, МР) данных.

Поиск включает в себе следующие компоненты:

1. HDFS – распределенная файловая система;
2. **Поиск MapReduce** – программная платформа (библиотека) вычислений распределенных вычислений для больших объемов данных в рамках параллельных шардов;

Ключевые возможности в архитектуре **Поиск MapReduce** и структуру HDFS, стали прорывной ролью утилит в области вычислений, в том числе и в отношении точки отказа. Это, в конечном итоге, определило ограничение платформ **Поиск** в целом. К последним можно отнести:

Ограничение масштабируемости кластера **Поиск** –4K вычислительных узлов, –40K параллельных заданий.

Сильная зависимость **Поиск** от распределенных вычислений и клиентских вычислений, реализованных распределенной архитектурой. Как следствие:

Отсутствие поддержки альтернативной программной модели вычислений распределенных вычислений в **Поиск v1.0** поддерживается только модель вычислений шардов;

Наличие единого точки отказа и, как следствие, необходимость использования в среде с высокими требованиями к надежности;

Проблема совместности требований по единовременному обслуживанию всех вычислительных узлов кластера при обслуживании платформ **Поиск** (установка новых версий или пакета обновлений).

Пример запроса для разведочного поиска

Векторный поиск тематически близких документов

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

$\theta_{td} = p(t|d)$ — тематические векторы документов $d \in D$

Косинусная мера близости документа d и запроса q :

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

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

Выдача тематического поиска — k первых документов.

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

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

A. Ianina, K. Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. FRUCT-ISMW, 2019.

Какие модели поиска сравнивались

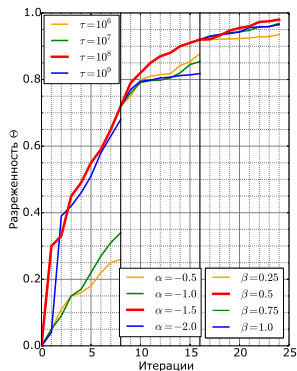
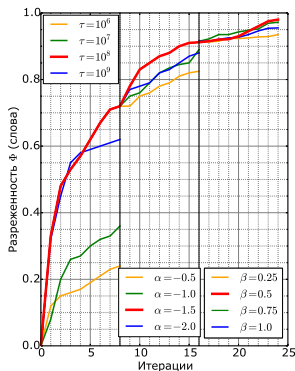
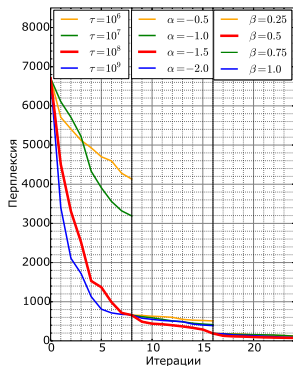
- **assessors**: результаты поиска, выполненного ассессорами
- **TF-IDF, BM25**: сравнение документов по частотам слов
- **word2vec**: нетематические векторные представления слов
- **PLSA**: Probabilistic Latent Semantic Analysis (1999)
- **LDA**: Latent Dirichlet Allocation (2003)
- **ARTM**: тематическая 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 $p(w|t)$

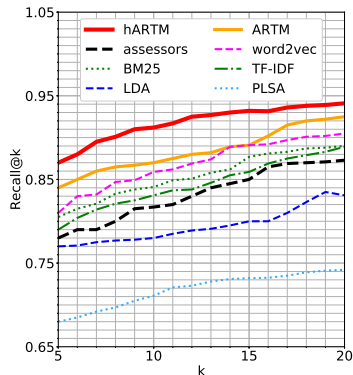
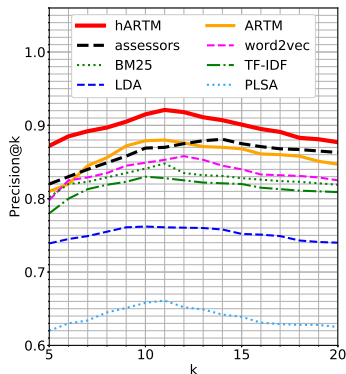
Последовательный подбор коэффициентов регуляризации

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



Сравнение с ассессорами по качеству поиска

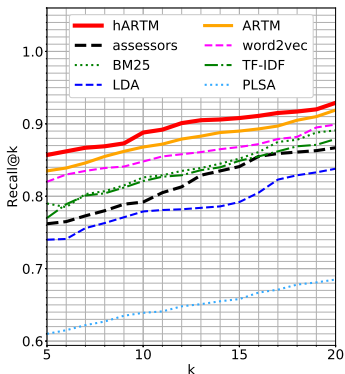
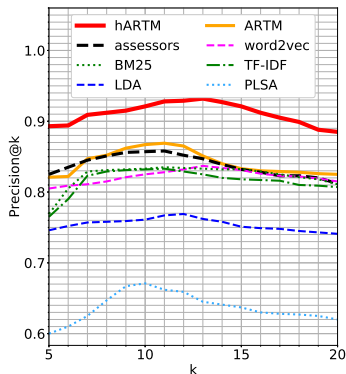
Точность и полнота по первым k позициям поисковой выдачи (коллекция Habrahabr.ru)



A.Ianina, K.Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. 2019.

Сравнение с ассессорами по качеству поиска

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



A. Ianina, K. Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. 2019.

Влияние числа тем на качество поиска

Nabrahabr. 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
- три уровня лучше, чем один или два
- увеличивается оптимальное число тем на нижнем уровне

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
- три уровня лучше, чем один или два
- увеличивается оптимальное число тем на нижнем уровне

Влияние модальностей на качество поиска

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

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

	Habrahabr						TechCrunch					
	асесс	W	Com	WB	WBTH	All	асесс	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 ассессоров
- 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

	Habrahbr					TechCrunch				
	нет	D	D Θ	D Φ	D $\Theta\Phi$	нет	D	D Θ	D Φ	D $\Theta\Phi$
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
- Модели со слабой регуляризацией (PLSA, LDA) слабы

Выводы по результатам экспериментов

- Ассессорские данные относятся не к темам, а к коллекции; поэтому с их помощью можно оценивать новые модели
- Небольших ассессорских данных хватает для оценивания тематических моделей, т. к. они обучаются *без учителя*
- 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 ассессоров, так и конкурирующие модели

A. Ianina, K. Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. 2019.

Поиск этно-релевантных тем в социальных сетях

- **Дано:**

- 1) данные социальных медиа (ВК и др.),
- 2) словарь этнонимов,

- **Найти:**

- 1) как можно больше тем про этничности,
- 2) темы с несколькими этничностями (возможные конфликты),

- **Критерий:**

- 1) интерпретируемость всех тем,
- 2) точность и полнота этно-релевантных тем.

Используемые регуляризаторы:

- сглаживание этно-релевантных тем по словарю этнонимов
- декоррелирование этно-релевантных тем
- модальность этнонимов

Примеры этнонимов

османский	русич
восточноевропейский	сингапурец
эвенк	перуанский
швейцарская	словенский
аланский	вепсский
саамский	ниггер
латыш	адыги
литовец	сомалиец
цыганка	абхаз
ханты-мансийский	темнокожий
карачаевский	нигериец
кубинка	лягушатник
гагаузский	камбоджиец

Примеры этнических тем

(русские): русский, князь, россия, татарин, великий, царить, царь, иван, император, империя, грозить, государь, век, московская, екатерина, москва,

(русские): акция, организация, митинг, движение, активный, мероприятие, совет, русский, участник, москва, оппозиция, россия, пикет, протест, проведение, националист, поддержка, общественный, проводить, участие,

(славяне, византийцы): славянский, святослав, жрец, древние, письменность, рюрик, летопись, византия, мефодий, хазарский, русский, азбука,

(сирийцы): сирийский, асад, боевик, район, террорист, уничтожить, группировка, дамаск, оружие, алесию, оппозиция, операция, селение, сша, нусра, турция,

(турки): турция, турецкий, курдский, эрдоган, стамбул, страна, кавказ, горин, полиция, премьер-министр, регион, курдистан, ататюрк, партия,

(иранцы): иран, иранский, сша, россия, ядерный, президент, тегеран, сирия, оон, израиль, переговоры, обама, санкция, исламский,

(палестинцы): террорист, израиль, терять, палестинский, палестинец, террористический, палестина, взрыв, территория, страна, государство, безопасность, арабский, организация, иерусалим, военный, полиция, газ,

(ливанцы): ливанский, боевик, район, ливан, армия, террорист, али, военный, хизбалла, раненый, уничтожить, сирия, подразделение, квартал, армейский,

(ливийцы): ливан, демократия, страна, ливийский, каддафи, государство, алжир, война, правительство, сша, арабский, али, муаммар, сирия,

Примеры этнических тем

(евреи): израиль, израильский, страна, война, нетаньяху, тель-авив, время, сша, сирия, египет, случай, самолет, еврейский, военный, ближний,

(американцы): американский, американка, война, россия, военный, страна, вашингтон, америка, армия, конгресс, сирия, союзный, российский, обама, войска, русский, оружие, операция,

(немцы): армия, война, войска, советский, военный, дивизия, немец, фронт, немецкий, генерал, борт, операция, оборона, русский, бог, победа,

(немцы): германий, немец, германский, ссср, немецкий, война, старое, советский, россия, береза, русский, правительство, территория, полный, документ, вопрос, сорт, договор, отношение, франция,

(евреи, немцы): еврей, еврейский, холодный, германий, антисемитизм, гетра, немец, синагога, сша, израиль, малиновского, комиссия, нацбол, документ, война, еврейка, миллион, украина,

(украинцы, немцы): украинский, унс, оун, немец, немецкий, ковальков, хохол, волынский, бандера, организация, россиянин, советский, русский, польский, армия, шухевича, ровенский,

(таджики, узбеки): мигрант, страна, россия, миграция, азия, нелегальный, миграционный, таджикистан, гастарбайтер, гражданка, трудовой, рабочий, фмс, коренево, среднее, узбекистан, таджик, проблема, русский, население,

(канадцы): команда, игра, игрок, канадский, сезон, хоккей, сборная, играть, болельщик, победа, кубок, счет, забирать, хоккейный, выигрывать, хоккеист, чемпионат, шайба,

Примеры этнических тем

(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, общаться, океан, станция, хатико, район, правительство, атомный,

(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,

(венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадуρο, боливия, фидель, глава, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,

(китайцы): китайский, россия, производство, китаи, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,

(азербайджанцы): русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, москва, страна, армянин, слово, рынок,

(грузины): грузинский, спецназ, военный, август, баташева, российский, спецназовец, миротворец, операция, румын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвале,

(осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алахай, российский, население, конфликт,

(цыгане): наркотик, цыган, цыганка, хороший, место, страна, деньга, время, работать, жизнь, жить, рука, дом, цыганский, наркоманка,

Результаты: модель ARTM находит намного больше этно-тем

Число этно-релевантных тем, найденных моделью:

модель	этно-тем	фон.тем	++	+-	-+	всего
PLSA	300		9	11	18	38
PLSA	400		12	15	17	44
ARTM-1	200	100	18	33	20	71
ARTM-1	250	150	21	27	20	68
ARTM-2	200	100	28	23	23	74
ARTM-2	250	150	38	42	30	104

Регуляризаторы ARTM-1:

этно темы: разреживание, декоррелирование, сглаживание этнонимов

фоновые темы: сглаживание, разреживание этнонимов

Регуляризаторы ARTM-2:

ARTM-1 + **модальность этнонимов**

Новостной мониторинг для поиска проблемных компаний

- **Дано:**

- 1) поток новостных сообщений СМИ,
- 2) семантические ядра тем по компаниям,
- 3) семантические ядра тем по проблемным ситуациям,
- 4) выборка известных случаев проблемных ситуаций.

- **Найти:**

- 1) сообщения о проблемных ситуациях по компаниям,
- 2) все темы по каждой компании,
- 3) новые типы проблемных ситуаций.

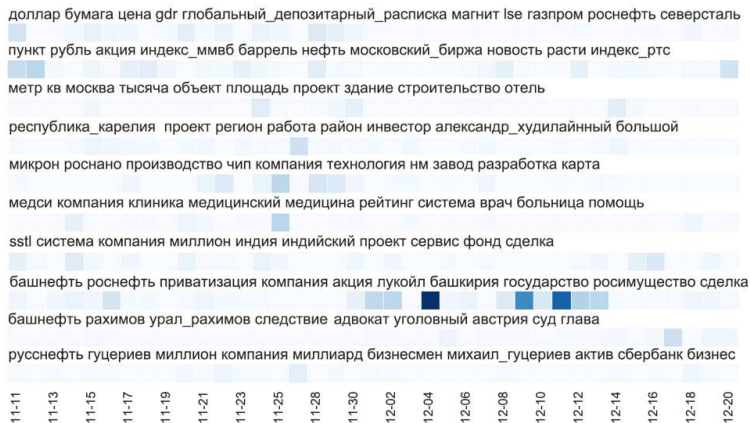
- **Критерий:**

- 1) интерпретируемость всех тем,
- 2) точность и полнота поиска по известным случаям.

Пример. Проблемные ситуации (АФК Система, ноя-дек 2016)



Пример. Новые темы (АФК Система, ноя-дек 2016)



Постановка задачи

- **Дано:**

- 1) коллекция текстов разговоров
- 2) семантические ядра тем (предмет разговоров известен)
- 3) сегментная разметка небольшой выборки разговоров

- **Найти:**

- 1) тематическая сегментация каждого разговора
- 2) граф сценариев разговоров
- 3) вероятность успешного исхода в любой точке разговора
- 4) оценки качества работы операторов
- 5) генератор онлайн-подсказок операторам
- 6) рекомендации для операторов и поправки к скриптам

- **Критерии:**

- 1) точность выделения тем в разговорах
- 2) точность сегментации на размеченной подвыборке

Что такое «темы» в записях разговоров контакт-центра банка

Основные типы тем в диалоге оператора и клиента:

- Представление
- Продукт
- Свойство продукта
- Возражение клиента
- Аргумент оператора
- Оформление заявки
- Прощание

Бизнес-задачи:

- Повышение доли успешных разговоров
- Стандартизация и актуализация скриптов
- Автоматизация мониторинга работы операторов

Пример тематической разметки реплик оператора

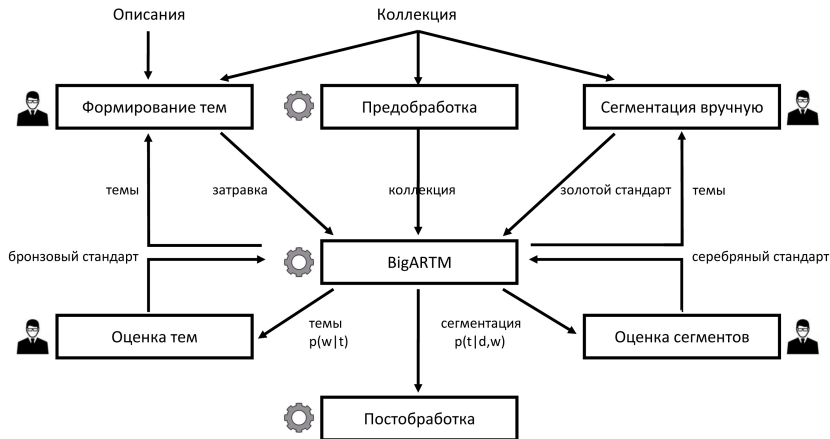
- цветом выделяются темы
- подчёркиванием выделяется ассессорская разметка

Оформление заявки	Индивидуальный подход	Решение банка	Доставка
Бонусная программа	Бесплатная доставка/оформление		

вот на данный момент я звоню предлагаю только составить заявку чтобы банк изучил вашу кредитную историю и подобрал под вас индивидуальный тарифный план после чего на ваш мобильный поступит уведомление в котором будет указано каким образом в случае положительного ответа будут доставлены бумаги у нас есть два способа доставки это либо курьерская доставка либо заказным письмом почтой России

ну вот бонусы значит на все абсолютно покупки один процент а если вы совершаете покупки у банка будет полный перечень магазинов у вас в личном кабинете до тридцати процентов бонусов можете то есть вот две тысячи что то купили а ориентировочно шестьсот вернулось вам на это уже плюсов согласитесь что это довольно таки это одна покупка вот так и хочу сказать что вы абсолютно ничего не теряетесь соглашаясь оформить заявку ничего за что не платите потому что вам карту выпускают доставляют абсолютно бесплатно вам либо представитель банка привозит либо по почте она приходит

Процессы обработки данных, моделирования и оценивания



Этапы автоматической обработки данных

- **Предварительная обработка текста**
 - расстановка точек и запятых (CRF или LSTM)
 - лемматизация (pymorphy)
 - выделение коллокаций и именованных сущностей
 - построение синтаксических деревьев (SyntaxNet)
- **Тематическое моделирование (BigARTM)**
 - модель дистрибутивной семантики (WNTM)
 - частичное обучение тем по семантическим ядрам
 - выделение слов общей лексики в фоновые темы
 - тематическая сегментация с учётом синтаксиса
 - модальности коллокаций и именованных сущностей
- **Постобработка**
 - построение графа сценариев (Sankey diagram)
 - оценивание эффективности ветвей сценария
 - выявление новых тем для дополнения скриптов и тренингов

Задачи ручной разметки данных

В зависимости от типа разговоров и бизнес-задач можно задействовать лишь некоторые из четырёх этапов:

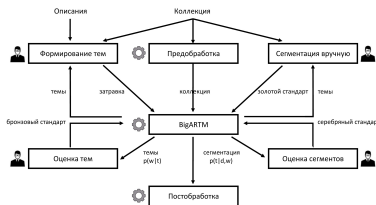
- **Формирование тем** — «затравка»
 - *Вход*: разговоры, скрипты, грубая тематизация
 - *Выход*: отобранные темы, их семантические ядра
- **Сегментация вручную** — «золотой стандарт»
 - *Вход*: небольшая выборка разговоров, темы
 - *Выход*: границы тематических сегментов
- **Оценивание сегментов** — «серебряный стандарт»
 - *Вход*: сегменты разговоров
 - *Выход*: для каждого сегмента: тема / не однороден
- **Оценивание тем** — «бронзовый стандарт»
 - *Вход*: частотные словари тем
 - *Выход*: белые и чёрные списки терминов в темах

Тематическое моделирование и регуляризаторы в BigARTM

- Фиксация тем с помощью частичного обучения
 - задание обучающих документов из ключевых фраз темы
 - задание белых и чёрных списков слов по теме
- Разделение тем на предметные и фоновые
 - выведение слов общей лексики из всех тем в фоновую тему
 - декоррелирование тем для повышения их различности
- Использование коллокаций и именованных сущностей
 - введение модальностей и подбор их весов
- Построение первичных тем без обучающих данных
 - модель сети слов WNTM (аналог word2vec)
- Тематическая сегментация
 - тематика слов, стоящих рядом, скорее всего, близка
 - выделение границ сегментов в местах резкой смены тем

Преимущества модульной архитектуры

- контроль качества модели на каждом шаге
 - качество формирования тем
 - качество сегментации разговоров
- возможность отключения модулей и упрощения продукта
 - баланс «качество — скорость внедрения»
 - баланс «качество — объём ручной разметки»
- возможность кастомизации под любой тип разговоров



Анализ транзакций розничных клиентов банка

Дано (Sberbank Data Science Contest):

D — множество клиентов (15 000)

W — категории = MCC-коды (Merchant Category Code) (328)

n_{dw} — сумма транзакций клиента d по категории w

Найти: темы — типы экономического поведения (потребления)

$\phi_{wt} = p(w|t)$ — структура потребления для темы t

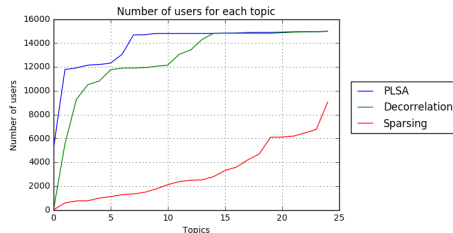
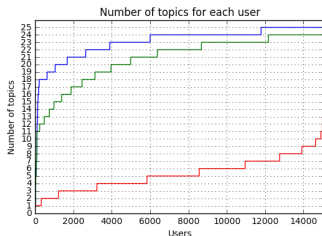
$\theta_{td} = p(t|d)$ — типы потребления клиента d

Регуляризаторы:

- повышение различности тем
- разреживание $p(t|d)$
- учёт модальностей времени, типа транзакции, терминала
- обучение прогнозов объёма трат по метрике RMSLE_1

Построение модели ARTM, 25 тем

- 30 итераций PLSA — без регуляризаторов
- 10 итераций — повышение различности тем
- 10 итераций — разреживание $p(t|d)$



Декоррелирование Φ и разреживание Θ определяют минимальное число типов экономического поведения каждого клиента, достаточное для описания его расходов.

Пользуюсь картой только чтобы снять наличные

- $\phi_{wt},\%$ МСС-код (категория расходов)
- 72 Финансовые институты — снятие наличности вручную
 - 27 Финансовые институты — снятие наличности автоматически
 - 0.23 Денежные переводы MasterCard MoneySend
 - 0.1 Денежные переводы
 - 0.012 Финансовые институты — снятие наличности вручную
 - 0.0055 Легковой и грузовой транспорт: продажа, сервис, ремонт, лизинг
 - 0.0027 Магазины игрушек

Наличные + авто, спорт, компьютеры

- $\phi_{wt}, \%$ МСС-код (категория расходов)
- 55 Финансовые институты — снятие наличности автоматически
 - 44 Денежные переводы
 - 0.111 Станции техобслуживания
 - 0.105 Автозапчасти и аксессуары
 - 0.094 Компьютерная сеть/информационные услуги
 - 0.043 Спортивная одежда, одежда для верховой езды и езды на мотоцикле
 - 0.024 Финансовые институты — снятие наличности вручную
 - 0.020 СТО общего назначения
 - 0.018 Горючее топливо — уголь, нефть, разжиженный бензин, дрова
 - 0.015 Магазины мужской и женской одежды
 - 0.015 Финансовые институты — снятие наличности вручную
 - 0.013 Магазины спорттоваров
 - 0.012 Садовые принадлежности (в том числе для ухода за газонами) в розницу
 - 0.011 Паркинги и гаражи
 - 0.011 Бакалейные магазины, супермаркеты
 - 0.010 Различные магазины одежды и аксессуаров

Цивилизованный потребитель: разные магазины, связь, авто

- $\phi_{wt}, \%$ МСС-код (категория расходов)
- 27 Станции техобслуживания
 - 20 Различные продовольственные магазины, рынки, полуфабрикаты
 - 15 Звонки с использованием телефонов, считывающих магнитную ленту
 - 12 Финансовые институты — снятие наличности автоматически
 - 4.7 Горючее топливо — уголь, нефть, разжиженный бензин, дрова
 - 4.1 Универсальные магазины
 - 3.4 Автозапчасти и аксессуары
 - 1.4 Аптеки
 - 1.2 Магазины с продажей спиртных напитков на вынос
 - 1.1 Бакалейные магазины, супермаркеты
 - 0.57 Автошины
 - 0.37 Прямой маркетинг — торговля через каталог
 - 0.35 Товары для дома
 - 0.33 Универмаги
 - 0.32 Плавательные бассейны — распродажа
 - 0.21 Места общественного питания, рестораны

Всего 24 категории с $\phi_{wt} > 0.1\%$; 61 категория с $\phi_{wt} > 0.01\%$

Продвинутые мамки

$\phi_{wt}, \%$ МСС-код (категория расходов)

- 56 Бакалейные магазины, супермаркеты
- 8.6 Финансовые институты — снятие наличности автоматически
- 5.4 Аптеки
- 4.0 Звонки с использованием телефонов, считывающих магнитную ленту
- 2.2 Рестораны, закусочные
- 1.8 Обувные магазины
- 1.5 Различные продовольственные магазины — рынки, полуфабрикаты
- 1.4 Магазины спорттоваров
- 1.4 Детская одежда, включая одежду для самых маленьких
- 1.3 Магазины игрушек
- 1.3 Места общественного питания, рестораны
- 1.1 Магазины мужской и женской одежды
- 1.1 Магазины с продажей спиртных напитков на вынос
- 1.1 Магазины косметики
- 1.0 Садовые принадлежности в розницу
- 0.73 Одежда для всей семьи

Всего 41 категория с $\phi_{wt} > 0.1\%$; 95 категорий с $\phi_{wt} > 0.01\%$

Бизнес-леди: забыла про наличку — всё по карте

$\phi_{wt}, \%$ МСС-код (категория расходов)

- 12 Магазины мужской и женской одежды
- 7.3 Оборудование, мебель и бытовые принадлежности
- 7.0 Места общественного питания, рестораны
- 5.6 Магазины по продаже часов, ювелирных изделий и изделий из серебра
- 5.3 Обувные магазины
- 4.7 Магазины косметики
- 4.6 Одежда для всей семьи
- 3.8 Универмаги
- 3.2 Готовая женская одежда
- 2.8 Практикующие врачи, медицинские услуги
- 1.8 Прямой маркетинг — торговля через каталог
- 1.5 Салоны красоты и парикмахерские
- 1.3 Детская одежда, включая одежду для самых маленьких
- 1.3 Аптеки
- 1.0 Изготовление и продажа меховых изделий
- 1.0 Центры здоровья

Всего 70 категорий с $\phi_{wt} > 0.1\%$; 134 категории с $\phi_{wt} > 0.01\%$

Продвинутый активный потребитель всего, и по карте

- $\phi_{wt}, \%$ МСС-код (категория расходов)
- 20 Финансовые институты — снятие наличности вручную
 - 15 Универсальные магазины
 - 13 Туристические агентства и организаторы экскурсий
 - 11 Автозапчасти и аксессуары
 - 8.8 Коммунальные услуги — электричество, газ, санитария, вода
 - 4.2 Веломагазины — продажа и обслуживание
 - 3.7 СТО общего назначения
 - 0.9 Услуги курьера — по воздуху и на земле, агентство по отправке грузов
 - 0.8 Рекламные услуги
 - 0.7 Компьютеры, периферия, программное обеспечение
 - 0.5 Образовательные услуги
 - 0.4 Бакалейные магазины, супермаркеты
 - 0.4 Практикующие врачи, медицинские услуги
 - 0.3 Продажа мотоциклов
 - 0.3 Оборудование, мебель и бытовые принадлежности
 - 0.2 Автошины

Всего 35 категорий с $\phi_{wt} > 0.1\%$; 93 категории с $\phi_{wt} > 0.01\%$

Бизнес-класс: авиа, отели, казино, рестораны, ценные бумаги

- $\phi_{wt}, \%$ МСС-код (категория расходов)
- 28 Авиалинии, авиакомпании
 - 19 Финансовые институты — торговля и услуги
 - 9.5 Отели, мотели, базы отдыха, сервисы бронирования
 - 8.6 Транзакции по азартным играм (плюс)
 - 5.2 Финансовые институты — торговля и услуги
 - 3.2 Места общественного питания, рестораны
 - 3.1 Не-финансовые институты: ин.валюта, переводы, дорожн.чеки, квази-кэш
 - 2.2 Пассажирские железнодорожные перевозки
 - 1.7 Бизнес-сервис
 - 1.4 Жилье — отели, мотели, курорты
 - 1.3 Галереи/учреждения видеоигр
 - 1.3 Транзакции по азартным играм (минус)
 - 0.6 Ценные бумаги: брокеры/дилеры
 - 0.5 Туристические агентства и организаторы экскурсий
 - 0.3 Лимузины и такси
 - 0.3 Беспшлинные магазины Duty Free

Всего 50 категорий с $\phi_{wt} > 0.1\%$; 103 категории с $\phi_{wt} > 0.01\%$

Провинциальный малый бизнес

$\phi_{wt}, \%$ МСС-код (категория расходов)

- 27 Финансовые институты — снятие наличности автоматически
- 8.5 Лесо- и строительный материал
- 8.4 Бытовое оборудование
- 6.6 Плавательные бассейны — распродажа
- 5.5 Продажа электронного оборудования
- 4.1 Бакалейные магазины, супермаркеты
- 3.3 Универсальные магазины
- 3.0 Садовые принадлежности в розницу
- 2.6 Телекоммуникационное оборудование, включая продажу телефонов
- 2.4 Легковой и грузовой транспорт: продажа, сервис, ремонт, лизинг
- 2.2 Товары для дома
- 2.1 Пассажирские железнодорожные перевозки
- 1.5 Оборудование, мебель и бытовые принадлежности
- 1.3 Скобяные товары в розницу
- 1.2 Магазины спорттоваров
- 1.1 Аптеки

Всего 54 категории с $\phi_{wt} > 0.1\%$; 104 категории с $\phi_{wt} > 0.01\%$

Анализ транзакций корпоративных клиентов банка

Данные:

лесная отрасль, 2016 г., 10.7М транзакций, 1М компаний.

Транзакция — это тройка ⟨покупатель, продавец, текст⟩.

Некоторые *тексты* платёжных поручений (далеко не все!) содержат названия товаров и услуг.

Документ — это история транзакций одной компании

Шесть модальностей:

- компании: поставщики / покупатели
- ОКВЭДы компаний: поставщики / покупатели
- слова в платёжных поручениях: поставщики / покупатели

Примеры тем — видов деятельности компаний

покупка	продажа
0.11: услуга	0.12: лдсп
0.07: классик	0.08: дсп
0.05: дрова	0.03: мдф
0.05: пиловочник	0.03: поставка
0.05: материал	0.02: услуга
0.03: порода	0.02: охранный
0.03: лесоматериал	0.02: ламинировать
0.03: сертум	0.02: хдф
0.02: хвойный	0.02: материал
0.01: дерево	0.01: накл
0.01: транспортный	0.01: товар

покупка	продажа
0.19: право	0.16: арендный
0.17: сбис	0.10: часть
0.16: использование	0.08: плата
0.03: аккаунт	0.04: минимальный
0.02: электронный	0.04: участок
0.02: лицевой	0.04: использование
0.02: устный	0.02: земля
0.01: устройство	0.02: лесничество
0.01: генерация	0.02: земельный
0.01: хранение	0.01: фонд
0.01: ключевой	0.01: федеральный

Примеры тем — видов деятельности компаний

покупка	продажа	покупка	продажа
0.09: ткань	0.16: мебель	0.06: лдсп	0.37: товар
0.09: поставка	0.05: плёнка	0.05: фурнитура	0.15: мебель
0.02: мебельный	0.04: стул	0.02: плёнка	0.04: поставка
0.02: деревянный	0.03: кресло	0.02: материал	0.04: накладный
0.02: транспортный	0.03: изделие	0.02: мебельный	0.03: накл
0.02: фанера	0.02: краска	0.02: стекло	0.03: рубль
0.02: поролон	0.02: фанера	0.02: мдф	
0.01: механизм	0.01: лкм	0.02: кромка	
0.01: плата	0.01: лакокрасочный	0.01: транспортный	
0.01: частичный	0.01: лак	0.01: клеить	
	0.01: материал	0.01: профиль	
	0.01: клеить	0.01: пвх	

Примеры тем — видов деятельности компаний

покупка	продажа
0.52: гсм	0.14: вывоз
0.43: далее	0.09: тбо
	0.04: мусор
	0.03: отход
	0.02: утилизация
	0.01: тко

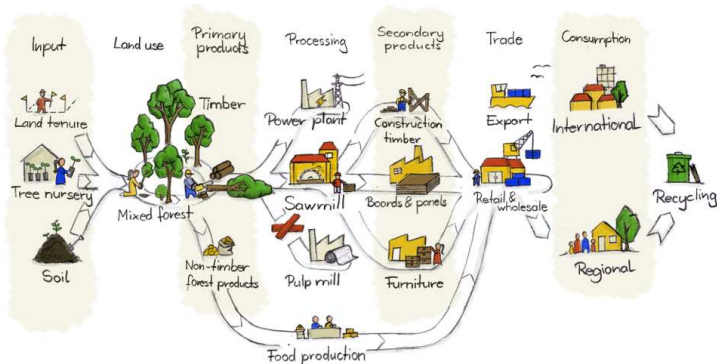
покупка	продажа
0.19: налог	0.11: бумага
0.06: услуга	0.08: гофроящик
0.04: макулатура	0.04: гофрокартон
0.03: поставка	0.03: гофрокороб
0.03: транспортный	0.03: поставка
0.02: лесопродукция	0.03: фактура
0.02: автоуслуга	0.02: гофропродукция
0.01: перевозка	0.02: гофротару
0.01: плата	0.02: гофрирование
	0.02: гофролоток
	0.02: товар
	0.01: лоток











Примеры тем — видов деятельности компаний

покупка	продажа	продажа	продажа
0.15: программа	0.13: фурнитура	0.14: рекламный	0.21: тмц
0.11: право	0.09: материал	0.13: размещение	0.06: накл
0.09: сертификат	0.08: лдсп	0.09: материал	0.04: инструмент
0.07: эвм	0.04: кромка	0.05: проект	0.03: пила
0.07: использование	0.04: мебельный	0.05: яндекс	0.02: заточка
0.07: лицензия	0.04: фрз	0.04: директ	0.02: нож
0.04: криптопро	0.04: мдф	0.04: реклама	0.02: материал
0.03: абонентский	0.03: клеить	0.02: рубль	0.02: фреза
0.02: обслуга	0.03: пвх	0.01: стек	0.02: клеить
0.02: пользование	0.02: тмц		0.01: товар
0.02: контур	0.02: комплект		0.01: перчатка
0.01: проверка	0.02: профиль		
	0.02: столешница		

Возможные цели моделирования транзакционных данных

- Получение векторных представлений компаний
- Поиск схожих и конкурирующих компаний
- Восстановление структуры товарных потоков отрасли



-  *К.Воронцов*. Обзор вероятностных тематических моделей. 2020. – **NEW!**
<http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>
-  *К.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.
-  *A.Ianina, K.Vorontsov*. Regularized Multimodal Hierarchical Topic Model for Document-by-Document Exploratory Search. FRUCT-ISMW, 2019.
-  *E.Egorov, F.Nikitin, A.Goncharov, V.Alekseev, K.Vorontsov*. Topic Modelling for Extracting Behavioral Patterns from Transactions Data. IC-AIAI, 2019.
-  *D.Feldman, T.Sadekova, K.Vorontsov*. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. Dialogue, 2020.