

Матричные разложения в задачах семантического анализа текстов

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

Семинар отдела
«Математическое моделирование экономических систем»
ВЦ РАН ФИЦ ИУ РАН
26 октября 2016

- 1 Задачи семантического анализа текстов**
 - Вероятностное тематическое моделирование
 - Простые тематические модели PLSA и LDA
 - Некорректность по Адамару
- 2 Задачи математического программирования**
 - Аддитивная регуляризация тематических моделей
 - Мультимодальные и гиперграфовые обобщения
 - Тематические модели последовательного текста
- 3 Реализация, эксперименты, приложения**
 - Программная реализация
 - Эксперименты
 - Приложения

Что такое «тема» в коллекциях текстовых документов?

Неформально,

- *тема* — семантически однородный кластер текстов
- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся терминов
- тем много меньше, чем терминов и чем документов

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 .

Тематическая модель оценивает вероятности $p(w|t)$ и $p(t|d)$ по наблюдаемым частотам $p(w|d)$ слов w в документах d .

Вероятностная порождающая модель

Основные предположения:

- каждый термин $w \in W$ в документе $d \in D$ имеет тему $t \in T$
- $D \times W \times T$ — 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: $p(w|d, t) = p(w|t)$

Вероятностная модель порождения документа d :

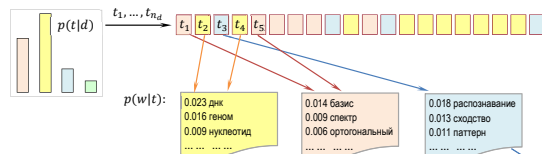
$$p(w|d) = \sum_{t \in T} p(w|t) p(t|d)$$

- $\phi_{wt} \equiv p(w|t)$ — распределение терминов в темах $t \in T$;
- $\theta_{td} \equiv p(t|d)$ — распределение тем в документах $d \in D$.

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

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

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



w_1, \dots, w_{n_d} :

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

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

Дано: W — словарь терминов

D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$

n_{dw} = сколько раз термин w встречается в документе d

Найти: параметры модели $p(w|d) = \sum_t \phi_{wt} \theta_{td}$ при ограничениях

$$\phi_{wt} \geq 0, \sum_w \phi_{wt} = 1, \quad \theta_{td} \geq 0, \sum_t \theta_{td} = 1;$$

$\phi_{wt} = p(w|t)$ — вероятности терминов w в теме t ,

$\theta_{td} = p(t|d)$ — вероятности тем t в документе d .

Это задача *стохастического матричного разложения*.

Задача некорректно поставлена — её решение не единственно:

$$\Phi \Theta = (\Phi S)(S^{-1} \Theta) = \Phi' \Theta'$$

для невырожденных $S_{T \times T}$ таких, что Φ', Θ' — стохастические.

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 $p_{tdw} = p(t|d, w)$:

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

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

LDA — Latent Dirichlet Allocation [Blei, Ng, Jordan, 2003]

Максимизация апостериорной вероятности (Dirichlet prior):

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td}}_{\text{ln правдоподобия } \mathcal{L}(\Phi, \Theta)} + \underbrace{\sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}}_{\text{регуляризатор с параметрами } \beta_w > 0, \alpha_t > 0} \rightarrow \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

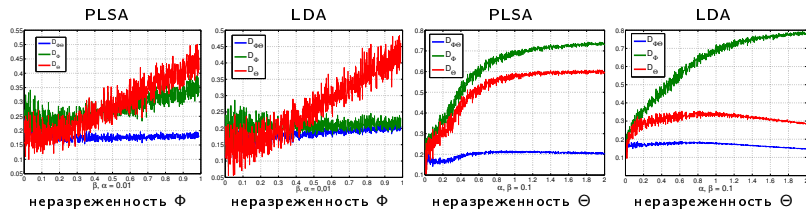
$$\begin{cases} \text{E-шаг:} & p_{tdw} = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \beta_w - 1 \right) \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(\sum_{w \in W} n_{dw}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

Задача тематического моделирования некорректно поставлена

Неединственность стохастического матричного разложения:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

Эксперимент. Произведение $\Phi\Theta$ восстанавливается устойчиво,
матрица Φ и матрица Θ — только когда сильно разрежены:



Вывод 1: нужны дополнительные ограничения на модель.

Вывод 2: регуляризатор Дирихле в LDA — слишком слабый.

ARTM — Аддитивная регуляризация тематических моделей

Максимизация \log правдоподобия с регуляризатором R :

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Комбинирование регуляризованных тематических моделей

Максимизация \log правдоподобия с n регуляризаторами R_i :

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + \sum_{i=1}^n \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta},$$

где τ_i — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

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

Примеры регуляризаторов

- 1 декоррелирование тем как столбцов Φ :

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

- 2 учёт *битермов* — пар слов, близко стоящих n_{uw} раз:

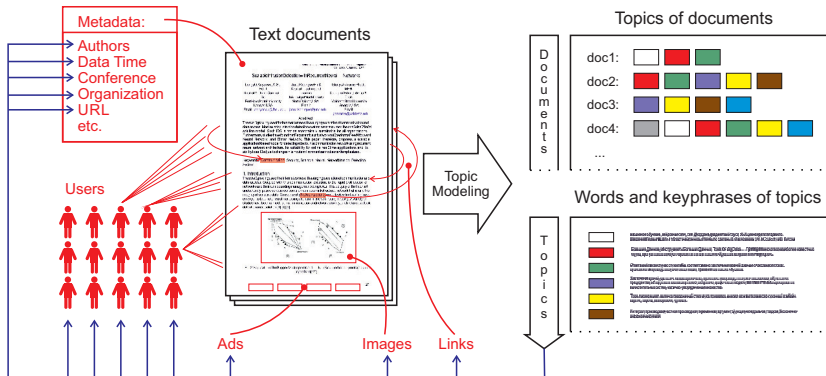
$$R(\Phi) = \tau \sum_{u, w \in W} n_{uw} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{wt} \rightarrow \max$$

- 3 удаление неинформативных тем:

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

ARTM легко обобщается на мультимодальные задачи

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



Мультимодальная ARTM [Vorontsov et al, 2015]

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

$W = W^1 \sqcup \dots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы \log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

Порождающая модальность и трёх-матричные разложения

Основные предположения:

- C — порождающая модальность (категории, авторы, ...)
- $D \times W \times T \times C$ — дискретное вероятностное пространство
- коллекция — i.i.d. выборка $(d_i, w_i, t_i, c_i)_{i=1}^n \sim p(d, w, t, c)$
- два предположения об условной независимости:
 $p(w|d, t) = p(w|t)$, $p(t|c, d) = p(t|c)$

Вероятностная модель порождения документа d :

$$p(w|d) = \sum_{t \in T} p(w|t) \sum_{c \in C} p(t|c) p(c|d) = \sum_{t \in T} \phi_{wt} \sum_{c \in C} \psi_{tc} \pi_{cd}$$

- $\phi_{wt} \equiv p(w|t)$ — распределение терминов в темах
- $\psi_{tc} \equiv p(t|c)$ — распределение тем в категориях
- $\pi_{cd} \equiv p(c|d)$ — распределение категорий в документах

ARTM для трёх-матричных разложений $\Phi\Psi\Pi$

Максимизация \log правдоподобия с регуляризатором R :

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \sum_{c \in C} \phi_{wt} \psi_{tc} \pi_{cd} + R(\Phi, \Psi, \Pi) \rightarrow \max_{\Phi, \Psi, \Pi};$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\left\{ \begin{array}{l} \text{E-шаг:} \\ \text{M-шаг:} \end{array} \right. \left\{ \begin{array}{l} p_{tcdw} \equiv p(t, c|d, w) = \mathop{\text{norm}}_{(t,c) \in T \times C} (\phi_{wt} \psi_{tc} \pi_{cd}); \\ \phi_{wt} = \mathop{\text{norm}}_{w \in W^m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right); \quad n_{wt} = \sum_{d,c} n_{dw} p_{tcdw} \\ \psi_{tc} = \mathop{\text{norm}}_{t \in T} \left(n_{tc} + \psi_{tc} \frac{\partial R}{\partial \psi_{tc}} \right); \quad n_{tc} = \sum_{d,w} n_{dw} p_{tcdw} \\ \pi_{cd} = \mathop{\text{norm}}_{c \in C} \left(n_{cd} + \pi_{cd} \frac{\partial R}{\partial \pi_{cd}} \right); \quad n_{cd} = \sum_{w,t} n_{dw} p_{tcdw} \end{array} \right.$$

Гиперграфовые модели для транзакционных данных

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

Примеры:

- **Данные социальной сети:**
 (d, u, w) — в блоге d пользователь u записал слово w
- **Данные сети интернет-рекламы:**
 (u, d, b) — пользователь u кликнул рекламное объявление b на веб-странице d
- **Данные рекомендательной системы:**
 (u, f, s) — пользователь u оценил фильм f в ситуативном контексте s

Хотим объяснить наблюдаемую выборку рёбер гиперграфа латентными векторами тематики 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:

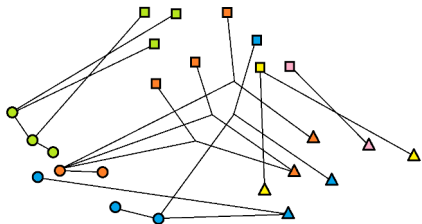
□ ○ △

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 ,

$p_k(d, x)$ — неизвестное распределение на рёбрах типа k .

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k :

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

$\theta_{td} = p(t|d)$ — тематика контейнера не зависит от типа ребра k

$\phi_{kvt} = p_k(v|t)$ — для модальности v в теме t на рёбрах типа k

Задача максимизации суммы \log правдоподобий с весами τ_k :

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

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

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

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

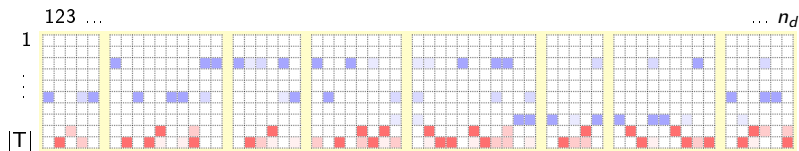
EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{ktdx} = p_k(t|d, x)$:

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

Тематическое моделирование последовательного текста

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

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



Предположения разреженности и непрерывности тематики:

- каждое предложение относится к 1–2 предметным темам
- соседние предложения часто имеют одинаковые темы
- слова общей лексики не влияют на тематику предложений
- между абзацами вероятность смены темы выше
- между секциями она ещё выше

EM-алгоритм с регуляризацией E-шага

Позиционный регуляризатор R_{di} зависит от позиции слова i в документе d и от параметров Φ, Θ через $p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$,

$$\mathcal{L}(\Phi, \Theta) + \sum_{d \in D} \sum_{i=1}^{n_d} R_{di}(p_{1dw_i}, \dots, p_{Tdw_i}) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}.$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\tilde{p}_{tdw} = p_{tdw} \frac{1}{n_{dw}} \sum_{\substack{i=1 \\ w_i=w}}^{n_d} \left(1 + \frac{\partial R_{di}}{\partial p_{tdw}} - \sum_{s \in T} p_{sdw} \frac{\partial R_{di}}{\partial p_{sdw}} \right);$$

$$\phi_{wt} = \text{norm}_w \left(\sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right);$$

$$\theta_{td} = \text{norm}_t \left(\sum_{w \in d} n_{dw} \tilde{p}_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right).$$

BigARTM: библиотека тематического моделирования

Ключевые возможности:

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

Сообщество:

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



Лицензия и среда разработки:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

ARTM — альтернатива байесовскому подходу


ARTM унифицирует разработку моделей с заданными свойствами


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	

 -- нестандартизируемые этапы, уникальная разработка для каждой задачи

 -- стандартизируемые этапы

Подзадачи, решаемые при построении тематических моделей

- Предобработка текста (лемматизация и не только)
- Инициализация EM-алгоритма
- Выбор стратегии регуляризации
- Подбор коэффициентов регуляризации
- Выбор критериев для контроля качества модели
- Подбор числа тем
- Построение тематических иерархий

На будущее:

- Автоматическая адаптация коэффициентов регуляризации
- Обеспечение полноты и устойчивости тем

Тестирование производительности

- 3.7М статей английской Вики, 100К уникальных слов

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

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

Мультиязычная модель Википедии

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

Тема 68				Тема 79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Дударенко М. А. Регуляризация многоязычных тематических моделей.
 Вычислительные методы и программирование. 2015. Т. 16. С. 26–36.

Мультиязычная модель Википедии

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

Тема 88				Тема 251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

Биграммы радикально улучшают интерпретируемость тем

Коллекция 850 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

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

Поиск этно-релевантных тем в социальных сетях

Основные задачи проекта:

- Разведочный поиск этнических тем в социальных медиа
- Мониторинг этих тем во времени и по регионам
- Сентимент-анализ и оценивание конфликтности

Примеры этнонимов:

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

Примеры этно-релевантных 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, уничтожать, 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, уничтожать, сирия, подразделение, квартал, армейский,
(ливийцы): ливан, демократия, страна, ливийский, каддафи, государство, алжир, война, правительство, сша, арабский, али, муаммар, сирия,

Примеры этно-релевантных 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 этно-релевантных тем

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

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

(венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадура, 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,

Результат: 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**

Приложения

- Сфокусированный тематический поиск в социальных сетях
- Разведочный поиск научно-технической информации
- Тематизация коллекции научных статей ММРО/ИОИ
- Тематизация новостных потоков для медиапланирования
- Сценарный анализ записей разговоров колл-центра
- Кросс-язычный разведочный поиск arXiv.org+Википедия
- Тематизация картин британского музея и их описаний
- Классификация авторефератов по областям знаний
- Информационный анализ электрокардиосигналов
- Поиск мотивов в задачах биоинформатики



<http://bigartm.org>

-  *K. Vorontsov*. Additive regularization for topic models of text collections. 2014.
-  *K. Vorontsov, A. Potapenko*. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. AIST 2014.
-  *K. Vorontsov, A. Potapenko*. Additive regularization of topic models. Machine Learning, 2015.
-  *K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Suvorova, A. Yanina*. Non-bayesian additive regularization for multimodal topic modeling of large collections. 2015.
-  *K. Vorontsov, A. Potapenko, A. Plavin*. Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.
-  *O. Frei, M. Apishev*. Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016. (в печати)
-  *M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko, K. Vorontsov*. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI 2016. (в печати)
-  *А.О.Янина, К.В.Воронцов*. Мультимодальные тематические модели для разведочного поиска в коллективном блоге. ИОИ 2016. (на рецензии)