

Вероятностные тематические модели

Лекция 1. Введение

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
<http://www.MachineLearning.ru/wiki>
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • 27 февраля 2020

- 1 Постановка задачи и элементарное решение**
 - Понятие темы в тематическом моделировании
 - Вероятностная модель порождения текста
 - Элементарное решение обратной задачи
- 2 Аддитивная регуляризация тематических моделей**
 - Максимизация и регуляризация правдоподобия
 - Классические модели PLSA и LDA
 - Мультимодальные тематические модели
- 3 Библиотека BigARTM**
 - Рациональный EM-алгоритм
 - Библиотека тематического моделирования BigARTM
 - Задания

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

Имея коллекцию текстовых документов, хотим узнать:

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

Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
 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 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
 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.

Пример 2. Биграммная модель научных конференций

Коллекция 1000 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
униграммы	биграммы	униграммы	биграммы
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

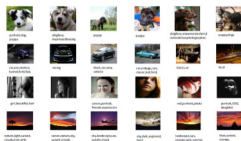
Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели // Магистерская диссертация, МФТИ, 2015.

Некоторые приложения тематического моделирования

разведочный поиск в
электронных библиотеках



мультимодальный поиск
текстов и изображений



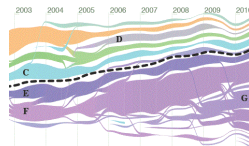
поиск тематического
контента в соцсетях



анализ банковских
транзакционных данных



детектирование и трекинг
новостных сюжетов

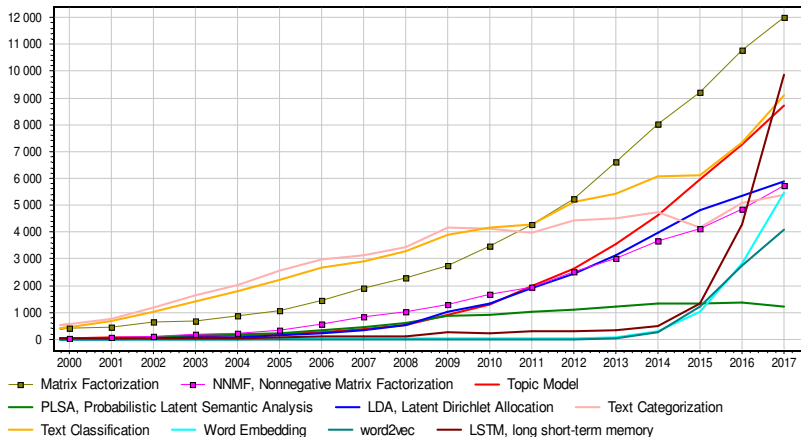


управление диалогом в
разговорном интеллекте



Тематическое моделирование и смежные области исследований

Динамика цитирования, по данным Google Scholar:



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:

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

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

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

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



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).

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

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

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

Вход: распределение $p(w|t)$ для каждой темы $t \in T$;
распределение $p(t|d)$ для каждого документа $d \in D$;

Выход: коллекция документов;

для всех $d \in D$

для всех позиций $i = 1, \dots, n_d$ в документе d

сгенерировать тему t_i из $p(t|d)$;

сгенерировать терм w_i из $p(w|t_i)$;

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

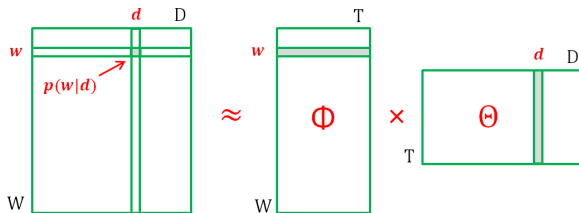
Дано: коллекция текстовых документов

- 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:



Система обозначений для частот — счётчиков числа термов

Ненаблюдаемые частоты, зависящие от t :

$n_{dwt} = \sum_{i=1}^n [d_i = d] [w_i = w] [t_i = t]$ — частота (d, w, t) в коллекции

$n_{wt} = \sum_d n_{dwt}$ — частота термина w в теме t

$n_{td} = \sum_w n_{dwt}$ — частота термов темы t в документе d

$n_t = \sum_{d,w} n_{dwt}$ — частота термов темы t в коллекции

Наблюдаемые частоты, не зависящие от t :

$n_{dw} = \sum_t n_{dwt}$ — частота термина w в документе d

$n_w = \sum_d n_{dw}$ — частота термина w в коллекции

$n_d = \sum_w n_{dw}$ — длина документа d

$n = \sum_{d,w} n_{dw}$ — длина коллекции

Элементарное решение обратной задачи

Выразим n_{dwt} через ϕ_{wt} , θ_{td} по формуле Байеса:

$$\frac{n_{dwt}}{n_{dw}} = p(t|d, w) = \frac{p(w, t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}.$$

Оценим условные вероятности ϕ_{wt} , θ_{td} частотами через n_{dwt} .

Получим систему уравнений относительно ϕ_{wt} , θ_{td} и n_{dwt} :

$$\left\{ \begin{array}{l} n_{dwt} = n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}, \quad d \in D, w \in W, t \in T; \\ \phi_{wt} \equiv \frac{n_{wt}}{n_t} = \frac{\sum_d n_{dwt}}{\sum_{d,w} n_{dwt}}, \quad w \in W, t \in T; \\ \theta_{td} \equiv \frac{n_{td}}{n_d} = \frac{\sum_w n_{dwt}}{\sum_{t,w} n_{dwt}}, \quad d \in D, t \in T. \end{array} \right.$$

Численное решение — методом простых итераций

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

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

при ограничениях неотрицательности и нормировки

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

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*,
если её решение

- существует,
- единственно,
- устойчиво.



Жак Саломон Адамар
(1865–1963)

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

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

Регуляризация — стандартный приём доопределения решения
с помощью дополнительных критериев.

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Условия вырожденности модели для тем и документов

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

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

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

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

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

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

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

Напоминания. Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \rightarrow \min_x; \\ g_i(x) \leq 0, & i = 1, \dots, m; \\ h_j(x) = 0, & j = 1, \dots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители $\mu_i, i = 1, \dots, m, \lambda_j, j = 1, \dots, k$:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} = 0, & \mathcal{L}(x; \mu, \lambda) = f(x) + \sum_{i=1}^m \mu_i g_i(x) + \sum_{j=1}^k \lambda_j h_j(x); \\ g_i(x) \leq 0; h_j(x) = 0; & \text{(исходные ограничения)} \\ \mu_i \geq 0; & \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; & \text{(условие дополняющей нежёсткости)} \end{cases}$$

Вывод системы уравнений из условий Каруша–Куна–Таккера

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

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

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

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

3. Если $\lambda_t \leq 0$, то тема t вырождена, $\phi_{wt} \equiv 0$ для всех w .

4. Если $\lambda_t > 0$, то либо $\phi_{wt} = 0$, либо $n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} > 0$:

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

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

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

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

Обобщение: произвольная функция потерь

Задача максимизации (при тех же ограничениях):

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ell \left(\sum_{t \in T} \phi_{wt} \theta_{td} \right) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \phi_{wt} \theta_{td} \ell' \left(\sum_s \phi_{ws} \theta_{sd} \right); \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum_{w \in d} n_{dw} p_{tdw} \end{cases} \end{cases}$$

Только при $\ell(z) = \ln z$ на E-шаге возникает формула Байеса.
 При $\ell(z) = z$ ставится задача $\sum_d n_d \langle \hat{p}(w|d), p(w|d) \rangle \rightarrow \max$.

Модель вероятностного латентного семантического анализа

PLSA — Probabilistic Latent Semantic Analysis:

- $R(\Phi, \Theta) = 0$ — нет никакой регуляризации.

Получаем то самое «элементарное решение обратной задачи».

EM-алгоритм: метод простой итерации для системы уравнений

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

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Модель латентного размещения Дирихле

LDA — latent Dirichlet allocation:

- распределения ϕ_t близки к заданному распределению β ;
- распределения θ_d близки к заданному распределению α .

Логарифм правдоподобия простой вероятностной модели:

$$R(\Phi, \Theta) = \sum_t \sum_w \beta_w \ln \phi_{wt} + \sum_d \sum_t \alpha_t \ln \theta_{td}.$$

M-шаг — *сглаженные* частотные оценки с параметрами β_w, α_t :

$$\phi_{wt} = \text{norm}_w(n_{wt} + \beta_w), \quad \theta_{td} = \text{norm}_t(n_{td} + \alpha_t).$$

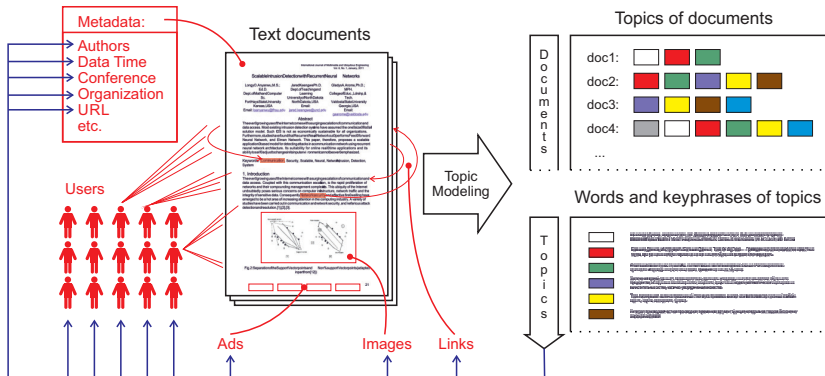
При $\beta_w < 0$ или $\alpha_t < 0$ — *разреживание* вместо сглаживания.

Причём здесь распределение Дирихле? — в следующей лекции...

Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:
 $p(\text{слово} | t)$, $p(n\text{-грамма} | t)$, $p(\text{слово-из-языка} | t)$, $p(\text{автор} | t)$,
 $p(\text{время} | t)$, $p(\text{ссылка} | t)$, $p(\text{пользователь} | t)$, $p(\text{баннер} | t), \dots$



Мультимодальная ARTM

W_m — словарь термов 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

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

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

Рациональный EM-алгоритм

Идея: E-шаг встраивается внутрь M-шага,
чтобы не хранить трёхмерный массив значений n_{dwt} .

Вход: коллекция D , число тем $|T|$, число итераций i_{\max} ;

Выход: матрицы термов тем Θ и тем документов Φ ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D, w \in W_m, m \in M, t \in T$;

для всех итераций $i = 1, \dots, i_{\max}$

$n_{wt}, n_{td} := 0$ для всех $d \in D, w \in W_m, m \in M, t \in T$;

для всех документов $d \in D$ и всех термов $w \in d$

$n_{tdw} := n_{dw} \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td})$ для всех $t \in T$;

$n_{wt} += n_{tdw}; n_{td} += n_{tdw}$ для всех $t \in T$;

$\phi_{wt} := \operatorname{norm}_{w \in W_m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ для всех $w \in W_m, m \in M, t \in T$;

$\theta_{td} := \operatorname{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$ для всех $d \in D, t \in T$;

Онлайновый EM-алгоритм (реализован в BigARTM)

Вход: коллекция D , число тем $|T|$, параметры j_{\max} , γ ;

Выход: матрицы термов тем Θ и термов документов Φ ;

инициализировать $n_{wt} := 0$; $n'_{wt} := 0$; $\phi_{wt} := \text{random}$;

для всех документов $d \in D$

инициализировать $\theta_{td} := \frac{1}{|T|}$;

для всех $j = 1, \dots, j_{\max}$ (итерации по документу)

$n_{tdw} := n_{dw} \text{norm}_{t \in T}(\phi_{wt} \theta_{td})$ для всех $w \in d$;

$\theta_{td} := \text{norm}_{t \in T} \left(\sum_w n_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$;

$n'_{wt} := n'_{wt} + n_{tdw}$ для всех $w \in d$;

если пора обновить матрицу Φ **то**

$n_{wt} := \gamma n_{wt} + n'_{wt}$; $n'_{wt} := 0$;

$\phi_{wt} := \text{norm}_{w \in W_m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$;

Сравнение оффлайн и онлайн алгоритмов

Оффлайн EM-алгоритм:

- 1 многократное итерирование по коллекции
- 2 однократный проход по документу
- 3 хранение матрицы Θ
- 4 обновление Φ в конце каждого прохода по коллекции
- 5 применяется при обработке небольших коллекций

Онлайн EM-алгоритм:

- 1 однократный проход по коллекции
- 2 многократное итерирование по каждому документу
- 3 нет необходимости хранить матрицу Θ
- 4 обновление Φ через заданное число пакетов
- 5 применяется при потоковой обработке больших коллекций

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>



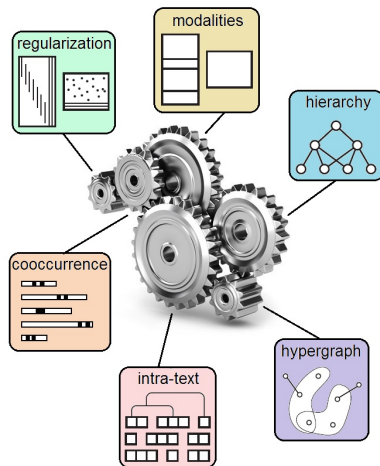
Лицензия и среда разработки:

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

Шесть ключевых механизмов BigARTM

- 1 ✓ регуляризация
- 2 ✓ модальности
- 3 иерархия тем
- 4 сочетаемость термов
- 5 внутритекстовые связи
- 6 гиперграфовые данные

(далее в курсе они будут изучаться подробно)



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.

Теоретическое домашнее задание

Два упражнения на принцип максимума правдоподобия:

- Униграммная модель документов: $p(w|d) = \xi_{dw}$
Найти параметры модели ξ_{dw} .
- Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d
Найти параметры модели ξ_w .

Творческое задание:

- Предложить модель, которая определяет роли слов в текстах и разделяет их на три группы:
 - слова общей лексики (фон)
 - тематические слова
 - специфичные слова документа (шум)
- Подсказка: можно использовать документную частоту слов.

Какими будут практические задания по курсу

Задача-минимум: научиться решать задачи NLP и строить тематические модели в BigARTM

- Устанавливать BigARTM и запускать примеры
- Делать предварительную обработку текстов
- Использовать «шесть механизмов» BigARTM
- Вычислять встроенные и собственные метрики качества
- Оптимизировать модель, подбирая регуляризаторы
- Оценивать интерпретируемость тем и улучшать её
- Визуализировать тематические структуры

Проблемы для исследования

Задача-максимум: решить открытую проблему

- Несбалансированность и семантическая однородность тем
- Агрегирование гетерогенных коллекций
- Создание новых тем при расширении коллекции
- В том числе, создание новых тем в иерархиях
- Прослеживание тем в новостных потоках
- Тематическая сегментация и посегментный поиск
- Предобученные тематические векторные представления слов
- Визуализация «карт знаний» (D3.js)

Датасеты для заданий по спецкурсу

- Научные статьи: arXiv, PubMed
- Научно-популярные статьи: ПостНаука, Элементы, Хабр
- Википедия
- Вики-227
- Новостной поток (RSS lenta.ru / нефильТРованный поток)
- Акты арбитражных судов РФ
- TechCrunch (английский)
- Открытые датасеты (английский): 20 newsgroups, NIPS, KOS
- Транзакции клиентов Sberbank DSD 2016

<http://bigartm.org>

<http://drive.google.com/drive/folders/1PPnw6aZOJAJolRYuwdGm437RssV-XQx0>

Несколько слов о предварительной обработке текста

- Удаление чисел, не-слов и «прочей грязи»
- Устранение переносов (когда текст был в pdf)
- Исправление опечаток (для пользовательских данных)
- Лемматизация (для русского языка)
- Стемминг (для английского языка)
- Удаление стоп-слов
- Удаление редких слов
- Автоматическое выделение терминов (ATE)
- Выделение именованных сущностей (NER)
- Разрешение кореференции (синонимии, анафоры)

Извлечение объектов и фактов из текстов в Яндексе. Лекция для Малого ШАДа, 2013. <https://habr.com/ru/company/yandex/blog/205198>

https://nlpub.ru/Обработка_текста

- Тематическое моделирование — это восстановление латентных тем в коллекции текстовых документов
- Это некорректно поставленная задача стохастического матричного разложения, её решение не единственно
- ARTM — комбинирование регуляризаторов для построения тематических моделей с требуемыми свойствами
- Базовый метод оптимизации — EM-алгоритм
- Рациональный EM-алгоритм со сложностью $O(n \cdot |T|)$
- Онлайнный EM-алгоритм: одного прохода может оказаться достаточно для большой коллекции текстов
- BigARTM — эффективная открытая реализация
- Что дальше:
 - применяем для решения практических задач
 - придумываем регуляризаторы, обобщаем модели
 - измеряем и оптимизируем качество моделей