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 • 15 февраля 2018

Содержание

Что такое «тема» в коллекции текстовых документов?

Неформально, *тема* — это:

- семантически однородное множество текстов
- специальная терминология предметной области
- набор часто совместно встречающихся терминов

Более формально:

- *тема* — частотный словарь терминов,
  $p(w|t)$ — частота термина w в теме t
- *тематика* документа — его распределение по темам,
  $p(t|d)$ — частота терминов темы t в документе d

Цели тематического моделирования:

- выявить латентные темы текстовой коллекции
- выявить тематику каждого документа

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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример 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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример 2. Биграммная модель научных конференций

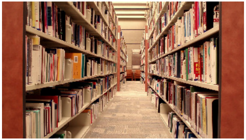
Коллекция 1000 статей конференций ММРО, ИОИ на русском

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

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

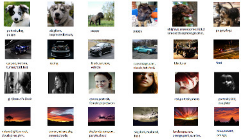
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

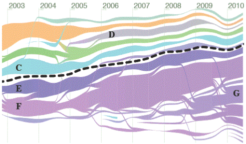


персонализированный поиск в соцсетях

мультимодальный поиск текстов и изображений

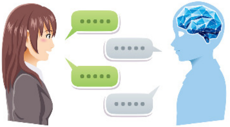


детектирование и трекинг новостных сюжетов



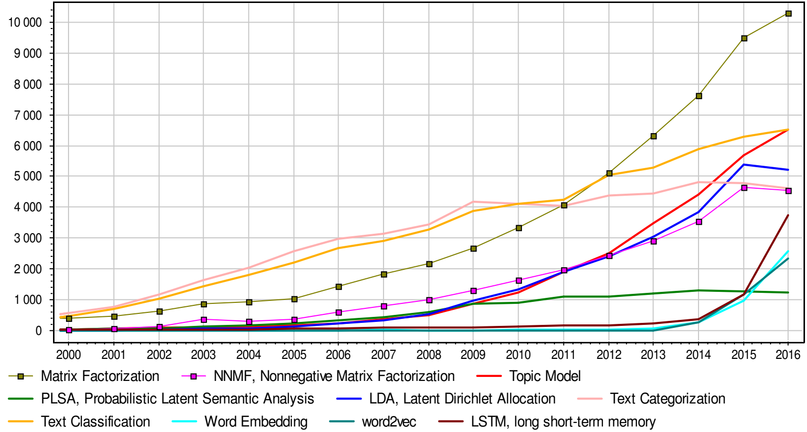
навигация по большим текстовым коллекциям

управлением диалогом в разговорном интеллекте



Тематическое моделирование и смежные области исследований

Динамика цитирования, по данным Google Scholar:



Пусть

- W — конечное множество слов (терминов, токенов)
- D — конечное множество текстовых документов
- T — конечное множество тем
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- каждое слово w в документе d связано 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: $p(w|d, t) = p(w|t)$

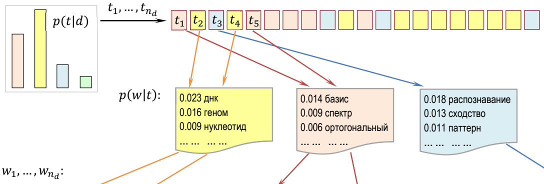
Тематическая модель, по формуле полной вероятности:

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

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

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

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



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

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

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

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

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
  - выбрать тему t_i из $p(t|d)$;
  - выбрать термин w_i из $p(w|t_i)$;

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

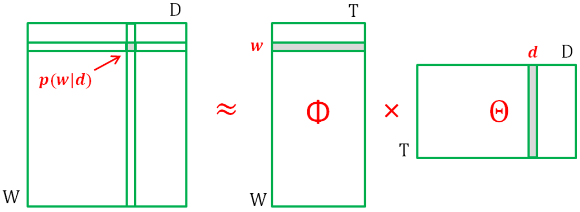
Дано: коллекция текстовых документов

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

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

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

Это задача стохастического матричного разложения:



Система обозначений для частот — счётчиков числа терминов

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 терминов темы t в документе d

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

Наблюдаемые частоты, не зависящие от t:

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

$n_w = \sum_{d,t} n_{dwt}$ — частота термина w в коллекции

$n_d = \sum_{w,t} n_{dwt}$ — длина документа d

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

Упрощённая вероятностная модель текста

- Пусть коллекция $(d_i, w_i, t_i)_{i=1}^n$ — это последовательность равновероятных элементарных событий.
- Тогда условные вероятности определяются через частоты:

  $p(w|d) = \dfrac{n_{dw}}{n_d}$ — распределение терминов в документе d,

  $p(t|d) = \dfrac{n_{td}}{n_d}$ — искомое распределение тем в документе d,

  $p(w|t) = \dfrac{n_{wt}}{n_t}$ — искомое распределение терминов в теме t.
- **Гипотеза условной независимости:**
  «вероятность термина в теме не зависит от документа»,

$$p(w|d,t) = p(w|t)$$
$$\frac{n_{dwt}}{n_{td}} = \frac{n_{wt}}{n_t}$$

Элементарное решение обратной задачи

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

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

Получим систему уравнений относительно параметров модели ϕ_{wt}, θ_{td} и вспомогательных переменных n_{dwt}:

$$\begin{cases} n_{dwt} = n_{dw}\dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}, & d \in D,\ w \in W,\ t \in T; \\ \phi_{wt} \equiv \dfrac{n_{wt}}{n_t} = \dfrac{\sum_d n_{dwt}}{\sum_{d,w} n_{dwt}}, & w \in W,\ t \in T; \\ \theta_{td} \equiv \dfrac{n_{td}}{n_d} = \dfrac{\sum_w n_{dwt}}{\sum_{t,w} n_{dwt}}, & d \in D,\ t \in T. \end{cases}$$

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

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

эквивалентна максимизации функционала

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

при ограничениях неотрицательности и нормировки

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

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар (1865–1963)

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

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

Регуляризация — стандартный приём доопределения решения с помощью дополнительных критериев.

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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 терминов $w \in W$

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

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

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

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

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

Напоминания. Условия Каруша–Куна–Таккера

Задача математического программирования:

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

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

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

Вывод системы уравнений из условий Каруша–Куна–Таккера

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

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

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

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

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

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

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

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

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

Классические модели PLSA и LDA

PLSA: probabilistic latent semantic analysis [Hofmann, 1999] (вероятностный латентный семантический анализ):

$$R(\Phi, \Theta) = 0.$$

M-шаг — частотные оценки условных вероятностей:

$$\phi_{wt} = \operatorname*{norm}_{w}\bigl(n_{wt}\bigr), \qquad \theta_{td} = \operatorname*{norm}_{t}\bigl(n_{td}\bigr).$$

LDA: latent Dirichlet allocation (латентное размещение Дирихле):

$$R(\Phi, \Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}.$$

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

$$\phi_{wt} = \operatorname*{norm}_{w}\bigl(n_{wt} + \beta_w\bigr), \qquad \theta_{td} = \operatorname*{norm}_{t}\bigl(n_{td} + \alpha_t\bigr).$$

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

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 терминов тем Θ и тем документов Φ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D$, $w \in W$, $t \in T$;
для всех итераций $i = 1, \dots, i_{\max}$
- $n_{wt}, n_{td} := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
- **для всех** документов $d \in D$ и всех слов $w \in d$
  - $n_{tdw} := n_{dw} \operatorname*{norm}_{t \in T} \bigl(\phi_{wt}\theta_{td}\bigr)$ для всех $t \in T$;
  - $n_{wt} \mathrel{+}= n_{tdw}$; $n_{td} \mathrel{+}= n_{tdw}$ для всех $t \in T$;
- $\phi_{wt} := \operatorname*{norm}_{w \in W} \Bigl(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\Bigr)$ для всех $w \in W$, $t \in T$;
- $\theta_{td} := \operatorname*{norm}_{t \in T} \Bigl(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\Bigr)$ 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 терминов тем Θ и тем 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)
  - $n_{tdw} := n_{dw} \operatorname*{norm}_{t \in T} \bigl(\phi_{wt}\theta_{td}\bigr)$ для всех $w \in d$;
  - $\theta_{td} := \operatorname*{norm}_{t \in T} \Bigl(\sum_w n_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\Bigr)$;
- $n'_{wt} := n'_{wt} + n_{tdw}$;
- **если** пора обновить матрицу Φ **то**
  - $n_{wt} := \gamma n_{wt} + n'_{wt}$; $n'_{wt} := 0$;
  - $\phi_{wt} := \operatorname*{norm}_{w \in W} \Bigl(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\Bigr)$;

BigARTM: библиотека тематического моделирования

Ключевые возможности:

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

Сообщество:

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

Лицензия и среда разработки:

- 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

BigARTM упрощает разработку тематических моделей

Для построения сложных моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».

Этапы моделирования	**Bayesian TM**	**ARTM**	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная порождающая модель данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Общий регуляризованный EM-алгоритм для любых моделей	
Реализация:	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)	
Оценивание:	Исследовательские метрики, исследовательский код	Стандартные метрики	Свои метрики
	Внедрение	Внедрение	

-- нестандартизуемые этапы, уникальная разработка для каждой задачи

-- стандартизуемые этапы

Теоретические задания

Два упражнения на принцип максимума правдоподобия:

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

Творческое задание:

1. Предложить модель, которая сама определяет роли слов в текстах и разделяет их на три группы:
   — слова общей лексики (фон),
   — тематические слова,
   — специфичные слова документа (шум)

Практические задания

Цель — научиться использовать BigARTM в нетривиальных прикладных проектах по анализу текстов

1. Установить BigARTM и научиться запускать примеры
2. Освоить методы предварительной обработки текстов
3. Построить «плоскую» тематическую модель
4. Построить иерархическую тематическую модель
5. Научиться делать тематический поиск
6. Научиться добавлять документы, не разрушая иерархию

Возможные коллекции текстов:

1. коллекции выпускных квалификационных работ
2. новостного потока
3. научно-популярного контента
4. архива научных статей

Резюме

- Тематическое моделирование — это восстановление латентных тем в коллекции текстовых документов
- Цели — поиск, систематизация, классификация текстов
- Базовая задача — стохастическое матричное разложение
- Базовый метод оптимизации — EM-алгоритм
- Рациональный EM-алгоритм со сложностью $O(n \cdot |T|)$
- Онлайновый EM-алгоритм: одного прохода может оказаться достаточно на большой коллекции текстов
- ARTM — многокритериальная аддитивная регуляризация
- BigARTM — эффективная открытая реализация