• Вероятностные языковые модели • Лекция 1. Оптимизация и регуляризация языковых моделей

Константин Вячеславович Воронцов
k.vorontsov@iai.msu.ru

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

ВМК МГУ • 2 марта 2026

Вероятностные языковые модели (PLM): о чём этот курс

В основном про (и это прежнее название курса)
вероятностные тематические модели (Probabilistic Topic Model)

Про прикладные задачи и методы анализа текстов для

- систематизации научных знаний
- социогуманитарных исследований

Про математику — про то, как

- ставить задачи анализа текстов в терминах оптимизации
- строить и упрощать прикладные математические теории
- придумывать модели для всё более сложных задач

Пререквизиты (какие знания потребуются)

- теория вероятности (в основном на конечных множествах)
- линейная алгебра, методы оптимизации (самые азы)
- машинное обучение (база, методология, эмбединги)
- язык Python

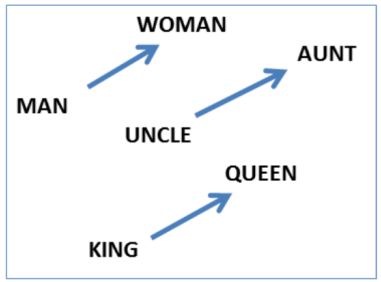
Содержание

Эволюция подходов машинного обучения в анализе текстов

Как решали задачи 10 лет назад: пирамида NLP

- морфологический анализ, лемматизация, опечатки
- синтаксический анализ, выделение терминов, NER
- семантический анализ, выделение фактов, тем

Контекстно независимые эмбединги слов в вероятностных моделях языка на основе матричных разложений



- модели дистрибутивной семантики
  word2vec [Mikolov, 2013], FastText [Bojanowski, 2016]
- тематические модели LDA [Blei, 2003], ARTM [2014]

Контекстно зависимые нейросетевые эмбединги

- рекуррентные нейронные сети LSTM [1997]
- модели внимания и трансформеры
  BERT [2018], GPT-3 [2020], GPT-4 [2023]
- NEW! тематические модели внимания

$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d}}\right) V$$

Простейшая частотная вероятностная языковая модель

Дано:

$(w_1, \ldots, w_n)$ — текст, состоящий из слов w_i словаря W, либо

$n_w = \sum_{i=1}^{n} [w_i = w]$ — частоты слов (*гипотеза «мешка слов»*)

Найти:

$p(w) = \xi_w$ — вероятностную языковую модель (в.п. W)

Критерий: максимум логарифма правдоподобия:

$$\ln \prod_{i=1}^{n} p(w_i) = \sum_{i=1}^{n} \ln \xi_{w_i} = \sum_{w \in W} n_w \ln \xi_w \to \max_{\{\xi_w\}}$$

при ограничениях $\sum_{w \in W} \xi_w = 1, \quad \xi_w \geqslant 0, \quad w \in W$

Решение (из условий Каруша–Куна–Таккера):

$\xi_w = \frac{n_w}{n}$ — частотная оценка вероятности встретить слово w

Вероятностная языковая модель коллекции документов

Дано:

$d = (w_{d1}, \ldots, w_{dn_d})$ — тексты документов, $d \in D$, либо

$n_{dw} = \sum_{i=1}^{n_d} [w_{di} = w]$ — частоты слов (*гипотеза «мешка слов»*)

Найти:

$p(w|d) = \xi_{wd}$ — вероятностную языковую модель (в.п. $D \times W$)

Критерий: максимум логарифма правдоподобия:

$$\ln \prod_{d \in D} \prod_{i=1}^{n_d} p(w_{di}|d) = \sum_{d \in D} \sum_{i=1}^{n_d} \ln \xi_{w_{di}d} = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln \xi_{wd} \to \max_{\{\xi_{wd}\}}$$

при ограничениях $\sum_{w \in W} \xi_{wd} = 1, \quad \xi_{wd} \geqslant 0, \quad w \in W, \quad d \in D$

Решение (из условий Каруша–Куна–Таккера):

$\xi_{wd} = \frac{n_{dw}}{n_d}$ — частотная оценка условной вероятности $\left(\neq \frac{n_w}{n} \right)$

Напоминание. Условия Каруша–Куна–Таккера

Задача математического программирования:

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

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

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

Принцип максимума правдоподобия + условия ККТ

Униграммная модель коллекции $p(w) = \xi_w$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_w \to \max_{\xi}, \quad \sum_{w\in W} \xi_w = 1, \quad \xi_w \geqslant 0.$$

Лагранжиан: $\mathscr{L} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_w - \lambda\Big(\sum_{w\in W} \xi_w - 1\Big)$;

$$\frac{\partial \mathscr{L}}{\partial \xi_w} = n_w \frac{1}{\xi_w} - \lambda = 0 \quad \Rightarrow \quad \lambda = n, \quad \xi_w = \frac{n_w}{n} \equiv \hat{p}(w).$$

Униграммная модель документов $p(w|d) = \xi_{wd}$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \xi_{wd} \to \max_{\xi}, \quad \sum_{w\in W} \xi_{wd} = 1, \quad \xi_{wd} \geqslant 0.$$

Лагранжиан: $\mathscr{L} = \sum_{d\in D}\Big(\sum_{w\in d} n_{dw} \ln \xi_{wd} - \lambda_d\Big(\sum_{w\in W} \xi_{wd} - 1\Big)\Big)$;

$$\frac{\partial \mathscr{L}}{\partial \xi_{wd}} = n_{dw} \frac{1}{\xi_{wd}} - \lambda_d = 0 \quad \Rightarrow \quad \lambda_d = n_d, \quad \xi_{wd} = \frac{n_{dw}}{n_d} \equiv \hat{p}(w|d).$$

Вероятностная тематическая модель коллекции документов

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

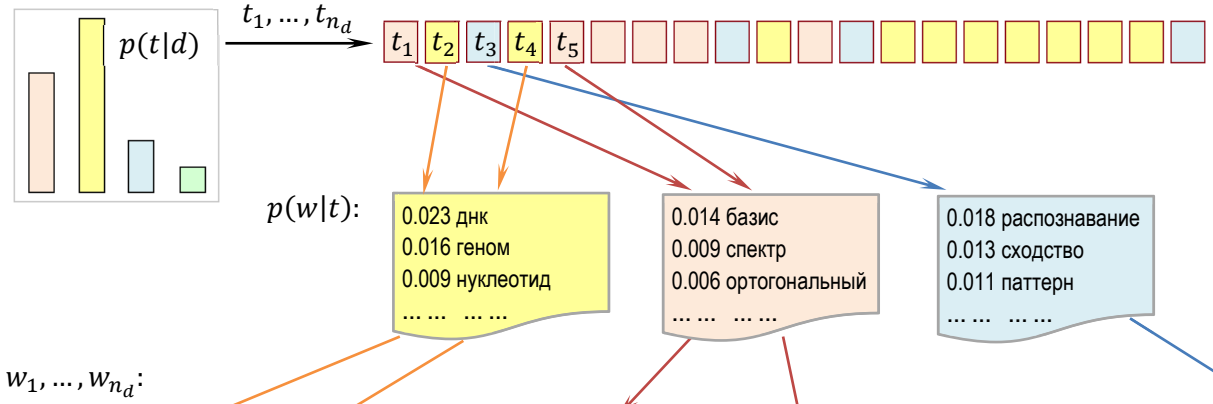
Тематическая модель, по формуле полной вероятности:

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

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

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

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



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 термов w по темам t в документах d:

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

Алгоритм порождения данных (generative story)

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 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 как «мешков-слов»

- n_{dw} — частота слова (терма) $w \in W$ в документе $d \in D$
- $|T|$ — сколько тем хотим определить в коллекции D

Найти: вероятностную тематическую языковую модель

- $p(w|d) = \sum\limits_{t \in T} p(w|t)\, p(t|d) = \sum\limits_{t \in T} \phi_{wt} \theta_{td}$
- $p(w|t) = \phi_{wt}$ — какие термы w образуют каждую тему t
- $p(t|d) = \theta_{td}$ — из каких тем t состоит каждый документ d

Критерий: максимум log-правдоподобия языковой модели:

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

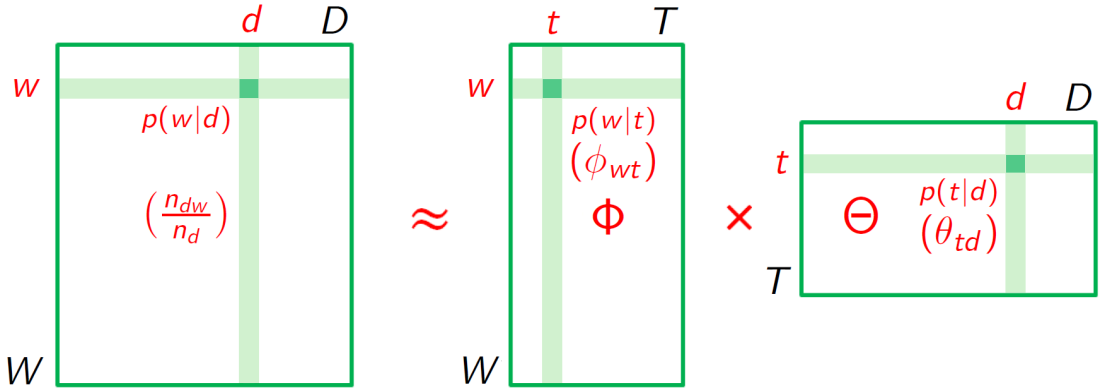
$$\phi_{wt} \geqslant 0; \quad \textstyle\sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \textstyle\sum_t \theta_{td} = 1$$

Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

Три интерпретации задачи тематического моделирования

1. **Мягкая би-кластеризация** документов и слов по темам

2. **Матричное разложение** — низкоранговое, стохастическое:



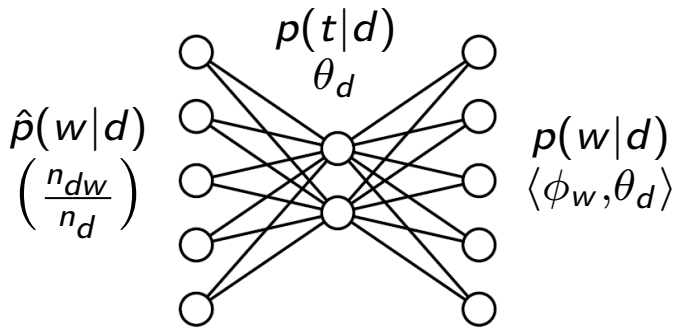
3. **Автокодировщик** документов в тематические эмбединги:

— кодировщик $f_\Phi : \frac{n_{dw}}{n_d} \to \theta_d$

— декодировщик $g_\Phi : \theta_d \to \Phi\theta_d$

задача реконструкции:

$$\sum_d n_d \sum_w \hat{p}(w|d) \ln p(w|d) \to \min_{\Phi,\Theta}$$



Пример 1. 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

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.

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

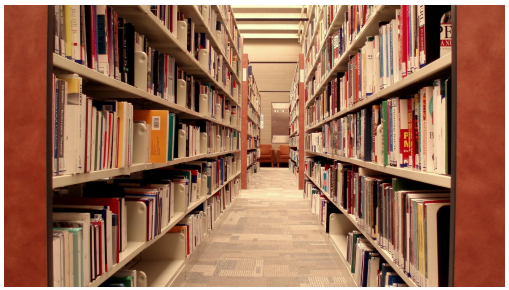
Коллекция 1000 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
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.

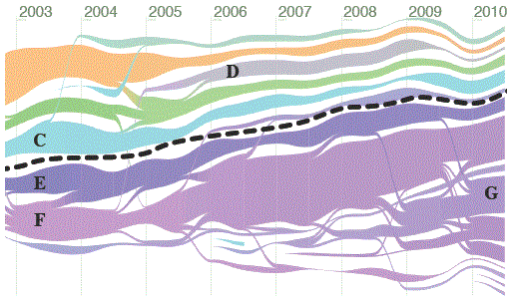
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

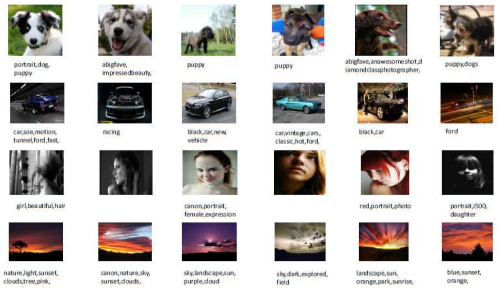


поиск тематических сообществ в соцсетях

выявление и отслеживание цепочек новостей

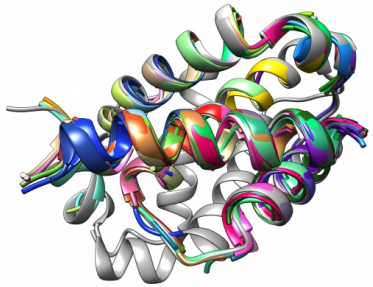


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

поиск паттернов в задачах биоинформатики



J.Boyd-Graber, Yuening Hu, D.Mimno. Applications of Topic Models. 2017.

H.Jelodar et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. 2019.

Цели и не-цели тематического моделирования

Цели:

- Выявлять кластерную тематическую структуру текстовой коллекции, сколько в ней тем и о чём они
- Получать *интерпретируемые* тематические векторные представления (эмбединги) слов $p(t|w)$, $p(t|d,w)$, документов $p(t|d)$, фрагментов $p(t|s)$, объектов $p(t|x)$
- Решать задачи поиска, категоризации, сегментации, суммаризации с помощью тематических эмбедингов

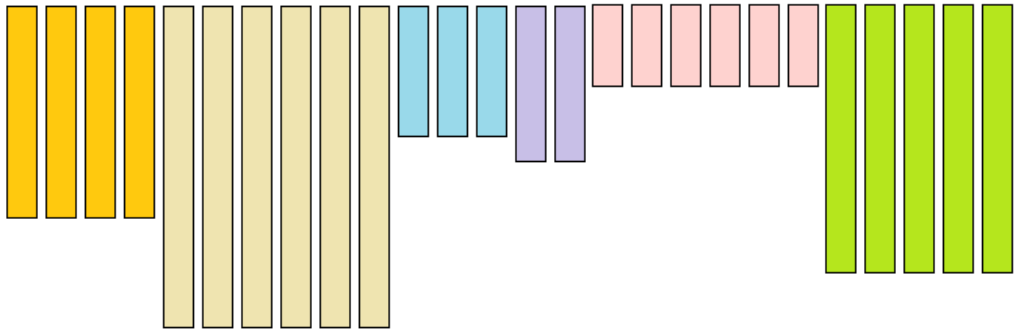
Не-цели:

- угадывать слова по контексту (это слабая модель языка)
- генерировать связный текст (слабые эмбединги)
- понимать смысл текста (тем не достаточно для этого)

Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$$\Omega = \left(\quad \right)$$



Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Vorontsov K. V. Rethinking Probabilistic Topic Modeling from the Point of View of Classical Non-Bayesian Regularization. 2023.

Необходимые условия экстремума и метод простых итераций

Операция нормировки вектора: $p_i = \underset{i \in I}{\text{norm}}(x_i) = \dfrac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω. Если ω_j — вектор локального экстремума задачи $f(\Omega) \to \max$ и $\exists i$: $\omega_{ij} \dfrac{\partial f}{\partial \omega_{ij}} > 0$, то ω_j удовлетворяет системе уравнений

$$\omega_{ij} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right).$$

- Численное решение системы — методом простых итераций
- Векторы $\omega_j = 0$ отбрасываются как вырожденные решения
- Итерации похожи на градиентную оптимизацию:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}},$$

но учитывают ограничения и не требуют подбора шага η

Доказательство леммы о максимизации на симплексах

Задача: $f(\Omega) \to \max\limits_{\Omega}$; $\sum\limits_{i \in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J.$

Функция Лагранжа:

$$\mathscr{L}(\Omega; \mu, \lambda) = -f(\Omega) + \sum_{j \in J} \lambda_j \Big(\sum_{i \in I_j} \omega_{ij} - 1 \Big) - \sum_{j \in J} \sum_{i \in I_j} \mu_{ij} \omega_{ij}.$$

Условия Каруша–Куна–Таккера для вектора ω_j:

$$\frac{\partial f(\Omega)}{\partial \omega_{ij}} = \lambda_j - \mu_{ij}, \quad \mu_{ij} \omega_{ij} = 0, \quad \mu_{ij} \geqslant 0.$$

Умножим обе части первого равенства на ω_{ij}:

$$A_{ij} \equiv \omega_{ij} \frac{\partial f(\Omega)}{\partial \omega_{ij}} = \omega_{ij} \lambda_j.$$

Согласно условию леммы $\exists i$: $A_{ij} > 0$. Значит, $\lambda_j > 0$.

Если $\frac{\partial f(\Omega)}{\partial \omega_{ij}} < 0$ для некоторого i, то $\mu_{ij} > 0 \;\Rightarrow\; \omega_{ij} = 0$.

Тогда $\omega_{ij} \lambda_j = (A_{ij})_+$; $\lambda_j = \sum\limits_i (A_{ij})_+ \;\Rightarrow\; \omega_{ij} = \operatorname{norm}\limits_i (A_{ij})$. ■

Теорема о сходимости итерационного процесса

$$\omega_{ij}^{t+1} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^{t}} \right)$$

Теорема. Пусть $f(\Omega)$ — ограниченная сверху, непрерывно дифференцируемая функция, и все Ω^t, начиная с некоторой итерации t^0 обладают свойствами:

- $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^{t} = 0 \to \omega_{ij}^{t+1} = 0$ (сохранение нулей)
- $\exists \varepsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^{t} \notin (0, \varepsilon)$ (отделимость от нуля)
- $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ (невырожденность)
- $\exists \lambda > 0 \;\; f(\Omega^{t+1}) - f(\Omega^t) \geqslant \lambda H(\Omega^t)$ (монотонный рост f)

Тогда $\left|\omega_{ij}^{t+1} - \omega_{ij}^{t}\right| \to 0$ при $t \to \infty$.

Ирхин И. А., Воронцов К. В. Сходимость алгоритма аддитивной регуляризации тематических моделей. Труды Института математики и механики УрО РАН. 2020.

Открытая проблема: неудобное четвёртое условие

Определение. $H(\Omega^t)$ есть линейное приближение приращения функции f в окрестности точки Ω^t:

$$f(\Omega^{t+1}) - f(\Omega^t) = H(\Omega^t) + o(\Delta\Omega^t)$$

Лемма. Квадратичное представление функции $H(\Omega)$:

$$H(\Omega) = \frac{1}{2}\sum_{j\in J}\sum_{i,k\in I_j}\left(\frac{\partial f(\Omega)}{\partial \omega_{ij}} - \frac{\partial f(\Omega)}{\partial \omega_{kj}}\right)^2 \omega_{ij}\omega_{kj}$$

Следовательно, $H(\Omega^t) \geqslant 0$.

$f(\Omega^{t+1}) - f(\Omega^t) \approx H(\Omega^t)$ — согласно определению;

$f(\Omega^{t+1}) - f(\Omega^t) \geqslant \lambda H(\Omega^t)$, начиная с некоторой итерации t при некотором $\lambda > 0$ — хотелось бы получить это как результат, а не вводить как предположение. Доказать это пока не удалось.

A.M.Ostrowski. Solution of equations and systems of equations. New York, 1966.

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена по Адамару*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар
(1865–1963)

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

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, $\operatorname{rank} S = |T|$
- $f(\Phi', \Theta') \approx f(\Phi, \Theta)$

Регуляризация — доопределение решения путём добавления критерия $+\tau R(\Phi, \Theta)$

Скаляризация критериев: $+\sum_i \tau_i R_i(\Phi, \Theta)$

А.Н.Тихонов
(1906–1993)

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

Теорема (необходимое условие экстремума). Точка локального экстремума (Φ,Θ) удовлетворяет системе уравнений

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

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

EM-алгоритм — решение этой системы методом простой итерации

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

Доказательство (по лемме о максимизации на симплексах)

Применим лемму к log-правдоподобию с регуляризатором:

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

$$\phi_{wt} = \operatorname*{norm}_{w\in W}\left(\phi_{wt}\frac{\partial f}{\partial \phi_{wt}}\right) = \operatorname*{norm}_{w\in W}\left(\phi_{wt}\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) =$$

$$= \operatorname*{norm}_{w\in W}\left(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right);$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\theta_{td}\frac{\partial f}{\partial \theta_{td}}\right) = \operatorname*{norm}_{t\in T}\left(\theta_{td}\sum_{w\in W} n_{dw}\frac{\phi_{wt}}{p(w|d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) =$$

$$= \operatorname*{norm}_{t\in T}\left(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right);$$

где определения вспомогательных переменных $p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{p(w|d)}$ выделяются в отдельные уравнения, и в итерационном процессе образуют E-шаг. ■

Свойства алгоритма EM (Expectation–Maximization) для ARTM

E-шаг — это формула Байеса:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}} = p_{tdw}$$

M-шаг — это оценки частот n_{wt}, n_{td} и условных вероятностей: при отсутствии регуляризатора $\phi_{wt} = \frac{n_{wt}}{n_t}$ и $\theta_{td} = \frac{n_{td}}{n_d}$, где

$n_{dwt} = n_{dw} p_{dwt}$ — частота тройки (d,w,t) в коллекции
$n_{wt} = \sum_d n_{dwt}$ — частота терма 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

Тема t вырождена и исключается из модели (topic selection), если $n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} \leqslant 0$ для всех термов $w \in W$

Документ d вырожден и его темы не определяются моделью, если $n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \leqslant 0$ для всех тем $t \in T$

Каверзный вопрос про частотные оценки вероятностей

Имеем ли мы формальное право записывать такие равенства:

- $p(w|d) = \frac{n_{dw}}{n_d}$ — распределение термов в документе d
- $p(t|d) = \frac{n_{td}}{n_d}$ — искомое распределение тем в документе d
- $p(w|t) = \frac{n_{wt}}{n_t}$ — искомое распределение термов в теме t

ДА, но только в ограниченной вероятностной модели текста, при предположении, что $(d_i, w_i, t_i)_{i=1}^n$ — фиксированная последовательность элементарных событий с вероятностями $\frac{1}{n}$

При общем предположении $(d_i, w_i, t_i) \overset{\text{i.i.d.}}{\sim} p(d, w, t)$ это лишь *приближённые частотные оценки условных вероятностей* (i.i.d. — independent identically distributed)

Рациональный EM-алгоритм

Вход: коллекция D, число тем $|T|$, число итераций $i_{\max}$;
Выход: матрицы термов тем Φ и тем документов Θ;
инициализация ϕ_{wt}, θ_{td} для всех $d \in D$, $w \in W$, $t \in T$;
для всех итераций $i = 1, \ldots, i_{\max}$
  $n_{wt} := 0$ для всех $w \in W$, $t \in T$;
  для всех документов $d \in D$
    $n_{td} := 0$ для всех $t \in T$;
    для всех термов $w \in d$
      $n_{tdw} := n_{dw} \operatorname*{norm}_{t \in T}(\phi_{wt}\theta_{td})$ для всех $t \in T$;
      $n_{wt} \mathrel{+}= n_{tdw}$; $n_{td} \mathrel{+}= n_{tdw}$ для всех $t \in T$;
    $\theta_{td} := \operatorname*{norm}_{t \in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$ для всех $t \in T$;
  $\phi_{wt} := \operatorname*{norm}_{w \in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$ для всех $w \in W$, $t \in T$;

K.V.Vorontsov, A.A.Potapenko. Additive Regularization of Topic Models. MLJ 2015.

PLSA, LDA: первые и самые известные тематические модели

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

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

Thomas Hofmann

M-шаг — частотные оценки условных вероятностей:

$$\phi_{wt} = \underset{w}{\text{norm}}\big(n_{wt}\big), \qquad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td}\big)$$

LDA: latent Dirichlet allocation (латентное размещение Дирихле):

$$R(\Phi, \Theta) = \sum_{t,w} \underset{\beta_w > 0}{(\beta_w - 1)} \ln \phi_{wt} + \sum_{d,t} \underset{\alpha_t > 0}{(\alpha_t - 1)} \ln \theta_{td}$$

David Blei

M-шаг — частотные оценки со смещением β_w, α_t:

$$\phi_{wt} = \underset{w}{\text{norm}}\big(n_{wt} + \beta_w - 1\big), \quad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td} + \alpha_t - 1\big)$$

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. NIPS-2001. JMLR 2003.

Обобщённая модель LDA (без ограничений на параметры)

Сглаживание ($\beta_{wt} > 0$, $\alpha_{td} > 0$) и разреживание ($\beta_{wt} < 0$, $\alpha_{td} < 0$):

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td}$$

Сглаживание фоновой темы t_Φ с общей лексикой языка:

- $\beta_{wt_\Phi} = \beta_0 p_\Phi(w)$ — тема t_Φ похожа на заданное $p_\Phi(w)$
- $\alpha_{t_\Phi d} = \alpha_0$ — общая лексика есть в каждом документе d

Сглаживание по «белым спискам» (seed words):

- $\beta_{wt} = \beta_0 \left[w \in W_t \right]$ — термы из W_t должны быть в t
- $\alpha_{td} = \alpha_0 \left[t \in T_d \right]$ — темы из T_d должны быть в d

Разреживание по «чёрным спискам»:

- $\beta_{wt} = -\beta_0 \left[w \in W_t \right]$ — термов из W_t не должно быть в t
- $\alpha_{td} = -\alpha_0 \left[t \in T_d \right]$ — тем из T_d не должно быть в d

Библиотека BigARTM

Ключевые возможности:

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

Сообщество:

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

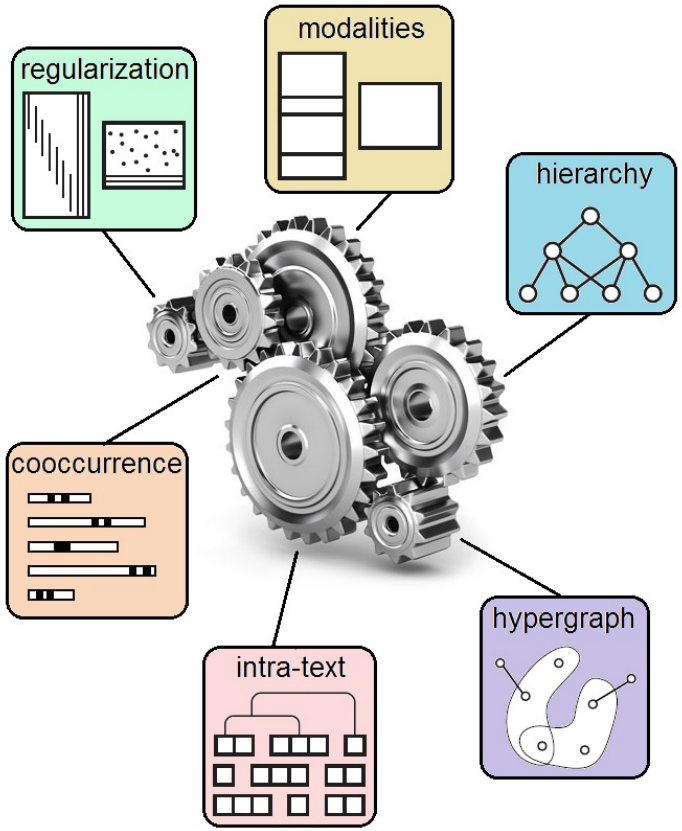
Лицензия и среда разработки:

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

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: open source library for regularized multimodal topic modeling of large collections. 2015.

Шесть ключевых механизмов BigARTM (спойлер)

1. регуляризация
2. модальности:
   n-граммы, термины, теги,
   категории, пользователи,
   время, авторы, источники,
   ссылки, языки
3. иерархия тем
4. парная сочетаемость термов
5. структуры и связи внутри текста
6. транзакции, гиперграфовые данны



Далее в курсе эти механизмы будут изучаться подробно

Этапы исследования при решении практических задач

1. Понимание задачи, «что нужно заказчику»
2. Выбор и настройка инструментария (BigARTM или др.)
3. Получение коллекции, перевод в удобный формат
4. Предварительная обработка (токенизация) текстов
5. Реализация базовой модели (обычно PLSA)
6. Измерение качества тематической модели
7. Интерпретация и визуализация тем
8. Добавление данных, регуляризаторов, модальностей
9. Оптимизация коэффициентов регуляризации
10. Оптимизация весов модальностей
11. Оптимизация числа тем
12. Постобработка и анализ результатов моделирования

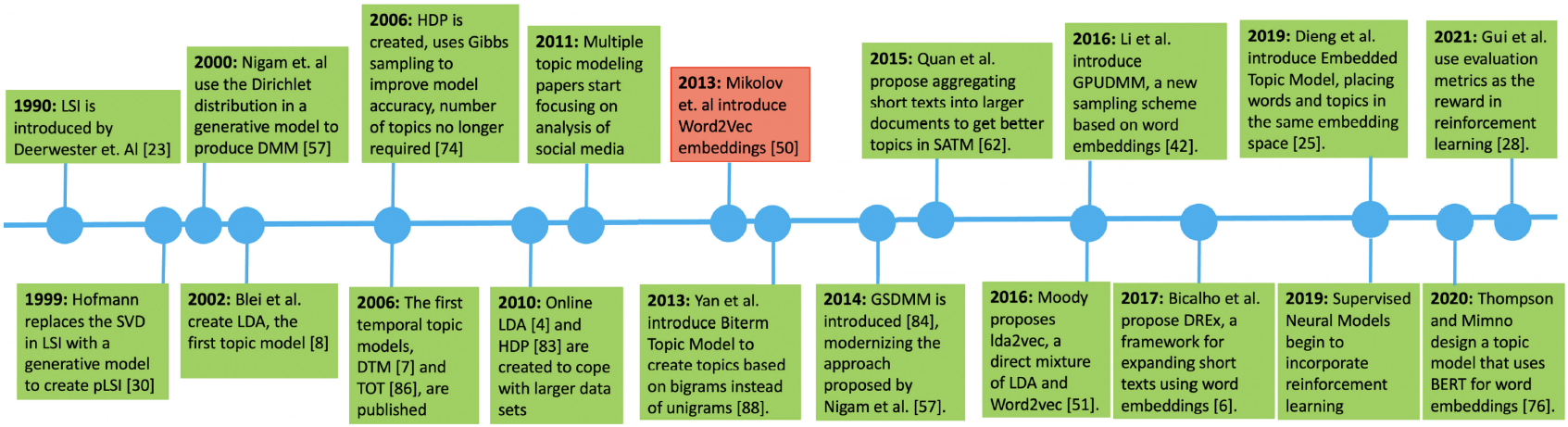
Методы предварительной обработки текста

- 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 слишком редких слов (если «мешок слов»)
- Удаление стоп-слов (если не строить фоновые темы)
- Автоматическое выделение терминов (ATE)
- Выделение именованных сущностей (NER)
- Сокращение словаря (Vocabulary Reduction)

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`

Neural Topic Models — эволюция PTM в сторону LLM



Как «объединить лучшее от двух миров»?

- **Neural:** общность, качество, предобучение, генерация
- **Topics:** интерпретируемость, полнота, простота, скорость

Что объединяет PTM и LLM, и что их разобщает:

⊕ обе — вероятностные языковые модели,

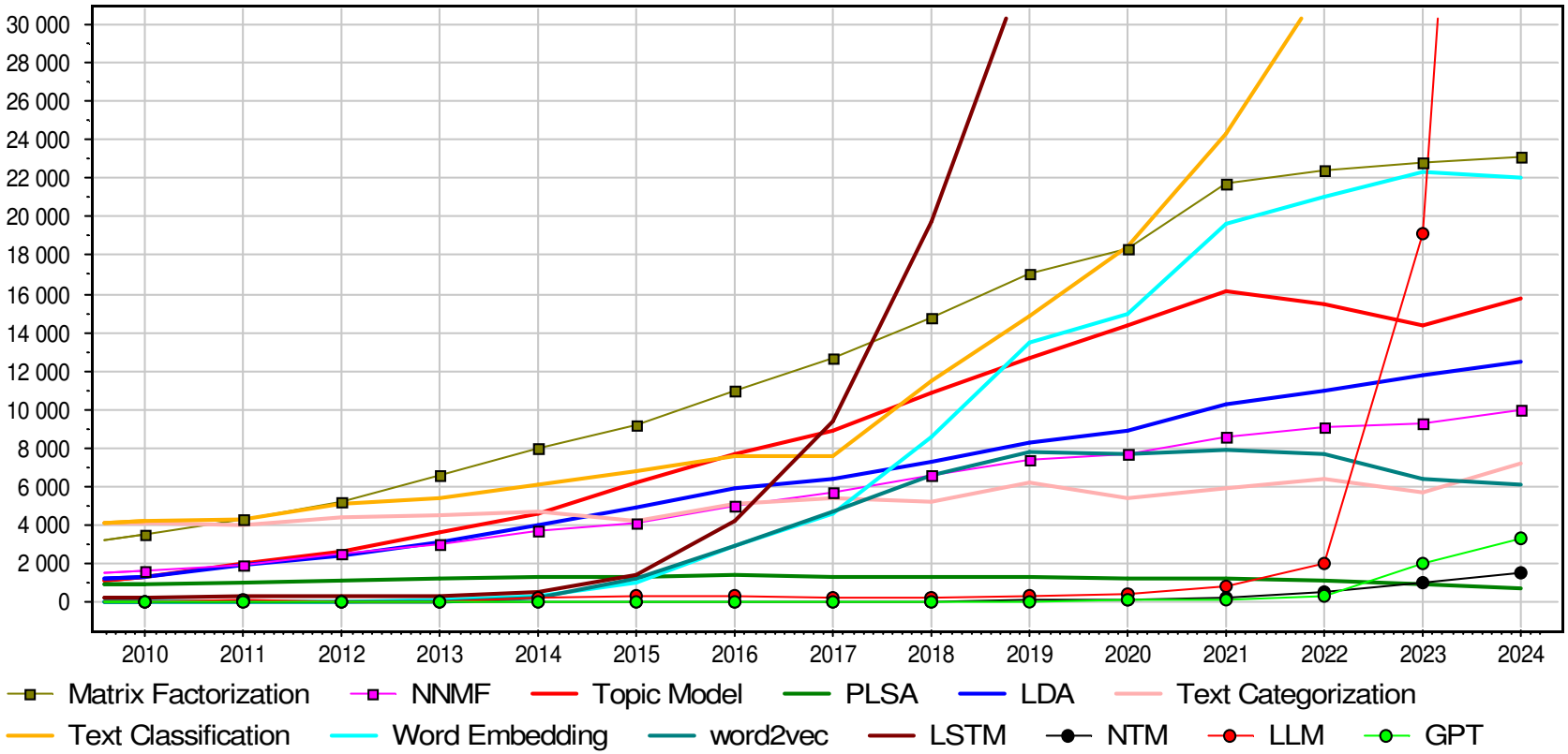
⊕ обе — автокодировщики, векторные представления текста

⊖ PTM: байесовское обучение, архитектура MF, мешок слов

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.

Научные тренды: PTM, LLM и смежные с ними

Динамика цитирования (по данным Google Scholar):
Topic Modeling и смежные области исследований:



Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.
He Zhao et al. Topic Modelling Meets Deep Neural Networks: A Survey. 2021

Резюме

- *Вероятностные модели языка* предсказывают слова в тексте
- *Основная лемма* о максимизации на единичных симплексах
- *Вероятностная тематическая модель* (PTM) — это:
  — мягкая кластеризация документов по кластерам-темам
  — стохастическое матричное разложение
  — вероятностные эмбединги текстов и слов
- Задача некорректно поставлена, её решение не единственно
- ARTM — построение моделей с заданными свойствами
- BigARTM — открытая реализация http://bigartm.org
- Что дальше в этом курсе:
  — изучаем много разных моделей и регуляризаторов
  — изучаем языковые модели, и не только тематические
  — избавляемся от гипотезы «мешка слов»
  — «объединяем лучшее от двух миров»
  — применяем для решения практических задач

Книга в серии «Классический учебник МГУ»

Задача тематического моделирования: «разложить по полочкам» коллекцию текстовых документов, не читая их

Теория аддитивной регуляризации: простая, изящная и выразительная альтернатива байесовскому обучению

Практика: библиотека BigARTM с открытым кодом и модульным подходом к комбинированию моделей

Приложения: научный поиск, анализ социальных сетей и новостных потоков, социогуманитарные исследования

Пререквизиты: математический анализ, линейная алгебра, теория вероятности, машинное обучение, язык Python

---

Воронцов К. В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. Москва, издательство URSS. 2025. ISBN 978-5-9710-9933-8.

Задания по курсу

Задача-минимум: научиться решать задачи анализа текстов с использованием тематического моделирования

Задача-максимум: получить новый научный результат

виды деятельности	оценка
теоретическая задача	X
решение прикладной задачи	$10X$
обзор по последним PTM/NTM	$10X$
участие в проекте	$20X$
работа над открытой проблемой	$25X$

где X — оценка за вид деятельности по 5-балльной шкале.
score — суммарная оценка по всем видам деятельности.

Итоговая оценка: $\min\big(5, \lfloor \text{score}/20 \rfloor\big)$ по 5-балльной шкале.

Задания к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Биграммная модель коллекции: $p(w|v) = \xi_{wv}$,
где v — слово, идущее в тексте перед w.
Найти параметры модели ξ_{wv}.

2. Биграммная модель документов: $p(w|v,d) = \xi_{dwv}$.
Найти параметры модели ξ_{dwv}.

Подсказка: применить условия ККТ или основную лемму.

3. Творческое задание (возможны разные решения)
Предложите модель, разделяющую роли слов в текстах:
— тематические слова
— специфичные слова документа (шум)
— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т}, \text{ш}, \text{ф}\}$.
Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.
Подсказка 3: можно использовать документную частоту слов.

Примеры датасетов для практических заданий по курсу

1. Открытые датасеты (английский): 20NG, NIPS, KOS
2. Ранжированные результаты поиска научных статей (по данным eLibrary, arXiv, PubMed)
3. Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
4. Техноблоги: Хабр (русский), TechCrunch (английский)
5. Данные социальных сетей: VK, Twitter, Telegram,...
6. Статьи по Complexity Sciences (для хронокарты науки)

- Википедия
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

---

http://bigartm.org
http://drive.google.com/drive/folders/1PPnw6aZOJAJoLRYuwdGm437RssV-XQx0

Проекты

- «Тематизатор» для социо-гуманитарных исследований:
  — пользователь задаёт грубый фильтр текстового потока;
  — задача: «классифицировать иголки в стоге сена»,
  — разделив темы на информативные и мусорные,
  — выделив аспекты и тональности в каждой теме;
  — конечная цель: q&q аналитика проблемной среды,
  — реализация данного сценария как модуля в среде Orange

- «Мастерская знаний» для научного поиска:
  — пользователь строит тематические подборки статей,
  — поисковая выдача формируется моделью SciRus;
  — задача: показать пользователю тематику подборки;
  — понадобится: автоматическое выделение терминов,
  — выделение тематических фраз из документов,
  — автоматическое именование и суммаризация тем;
  — конечная цель: помочь в понимании предметной области

Открытые проблемы тематического моделирования

1. Тематические модели внимания последовательного текста
2. Проблема несбалансированности тем в коллекции
3. Измерение интерпретируемости тем (когерентность)
4. Обеспечение 100%-й интерпретируемости тем
5. Автоматическое именование и суммаризация тем
6. Калибровка моделей тематической фильтрации
7. Согласование тем с предобученными эбмедингами LLM
8. Статистические оценки состоятельности тем
9. Обнаружение новых тем или трендов в потоке текстов
10. Обеспечение устойчивости и полноты множества тем
11. Автоматический подбор гиперпараметров, AutoML
12. Гиперграфовые тематические модели для RecSys