Towards bridging the gap between Probabilistic Topic Models and Large Language Models

Konstantin Vorontsov, Ilya Irkhin, Ilya Dyakov

Dr.Sc.(phys.-math.), professor of RAS,
head of Machine Learning & Semantic Analysis laboratory at Institute for Artificial Intelligence, Moscow State University;
head of department "Mathematical Methods of Forecasting", MSU;
head of department "Machine Learning & Digital Humanities", MIPT
(k.vorontsov@iai.msu.ru)

MathAI 2025 • Int. Conf. Mathematics in Artificial Intelligence
Sirius International Mathematical Center • Sochi • 24–28 March 2025

Содержание

Тематическое моделирование: «о чём все эти тексты?»

Дано: коллекция текстовых документов как «мешков-слов»

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

Найти: тематическую языковую модель

- $p(w|d) = \sum\limits_{t \in T} p(w\,|\,\cancel{d}, t)\, p(t|d) = \sum\limits_{t \in T} \phi_{wt}\theta_{td}$
- $p(w|t) = \phi_{wt}$ — из каких слов w состоит каждая тема $t \in T$
- $p(t|d) = \theta_{td}$ — из каких тем t состоит каждый документ d

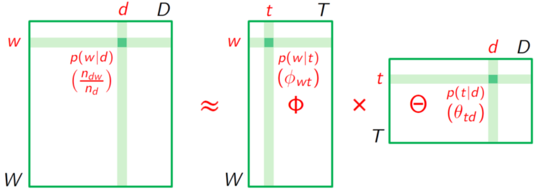
Критерий: правдоподобие предсказания слов w в документах d с дополнительными критериями-регуляризаторами $R_i(\Phi, \Theta)$:

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

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

Три интерпретации задачи тематического моделирования

1. **Мягкая кластеризация** документов по кластерам-темам
2. **Матричное разложение** — низкоранговое, стохастическое:



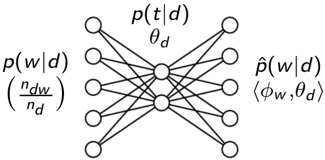
3. **Автокодировщик** документов в тематические эмбединги:

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

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

задача реконструкции:

$$\sum_{d,w} n_{dw} \ln\langle\phi_w, \theta_d\rangle \;\to\; \min_{\Phi,\Theta}$$



Цели и не-цели тематического моделирования

Цели:

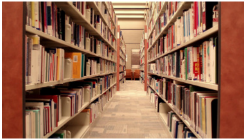
- выявлять тематическую кластерную структуру текстовой коллекции (сколько в ней тем и о чём они), представлять результат в удобной для человека форме
- получать *интерпретируемые* тематические векторы (эмбединги) слов $p(t|w)$, слов-в-контексте $p(t|d,w)$, документов $p(t|d)$, фрагментов $p(t|s)$, объектов $p(t|x)$
- решать с их помощью задачи поиска, фильтрации, категоризации, сегментации, суммаризации текстов

Не-цели:

- угадывать слова по контексту (это слабые модели языка)
- понимать смысл текста
- генерировать связный текст

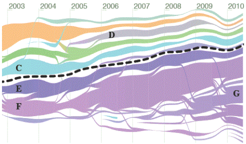
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

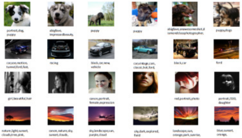


поиск тематических сообществ в соцсетях

выявление и отслеживание цепочек новостей

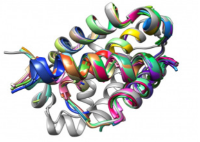


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

поиск паттернов в задачах биоинформатики

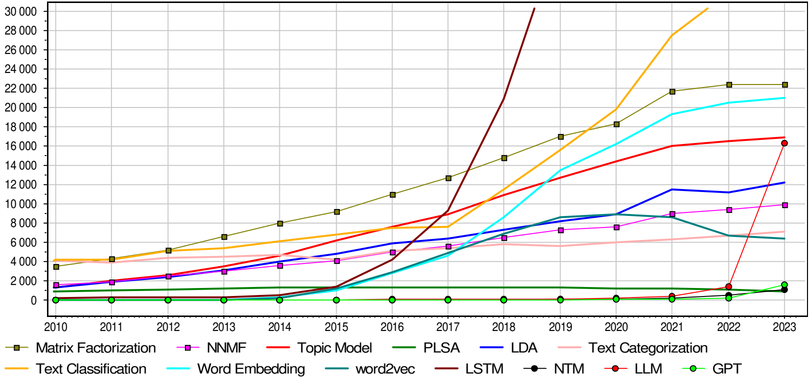


J.Boyd-Graber, Yuening Hu, D.Mimno. Applications of Topic Models. 2017.

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

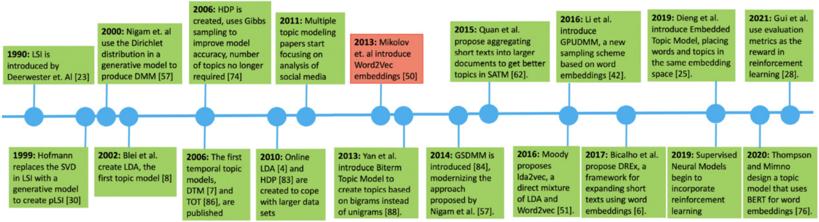
Научные тренды: PTM, LLM и смежные с ними

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

Эволюция тематического моделирования



Neural Topic Models — поток публикаций начиная с 2016

Как «объединить лучшее от двух миров»?

- **Neural:** качество, универсальность, генеративность
- **Topic:** скорость, интерпретируемость, простота

Что объединяет: векторизация, оптимизация, регуляризация, гомогенизация, локализация (контекст и внимание)

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

Сходства и отличия от LLM

PTM и LLM — что общего

- языковая модель, которая предсказывает слова в тексте
- автокодировщик, который переводит текст в эмбединги
- мультимодальность, мультиязычность данных
- многозадачность, многокритериальность обучения

PTM — принципиальные отличия от LLM

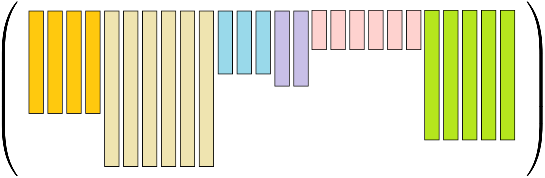
- намного более слабая языковая модель
- эмбединги вероятностные, разреженные, интерпретируемые
- простота и скорость матричного разложения

PTM — дальнейшее развитие навстречу LLM

- отказ от байесовского обучения $\to$ совместимость с SG
- отказ от мешка слов $\to$ тематическая модель внимания
- данные на гиперграфе $\to$ гомогенизация эмбедингов

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 $|I_j|$:

$\Omega =$ 

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

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

Необходимые условия экстремума и метод простых итераций

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

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

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

- 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

Напоминание. 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, $j = 1, \dots, k$:

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

Доказательство леммы о максимизации на симплексах

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

Функция Лагранжа:

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

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

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

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.

Промежуточные итоги и направления исследований

- Метод похож на обычную градиентную оптимизацию, но не требует подбора градиентного шага η
- Ограничения неотрицательности и нормировки могут накладываться не на все векторы, а лишь на некоторые
- Операция norm может приводить к обнулению части координат, следовательно, к разреживанию векторов ω_j
- **Приложения:**
  — вероятностное тематическое моделирование
  — неотрицательные матричные разложения
  — монотонные нейронные сети
  — сети для аппроксимации функций распределения
- **Открытая проблема:** упростить четвёртое условие в теореме сходимости (оно представляется избыточным)
- **Открытая проблема:** оценить скорость сходимости

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

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta), \;\; \mathrm{rank} S = |T|$
- $L(\Phi', \Theta') \approx L(\Phi, \Theta)$

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:

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

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

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:

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

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

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

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

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

■

PLSA и LDA — две самые известные тематические модели

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

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

М-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):

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

М-шаг — частотные оценки с поправками $\beta_w > -1$, $\alpha_t > -1$:

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

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

Байесовская и классическая регуляризация

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

$$\mathrm{Posterior}(\Omega|X,\gamma) \;\propto\; p(X|\Omega)\,\mathrm{Prior}(\Omega|\gamma)$$
$$\Omega := \arg\max_{\Omega} \mathrm{Posterior}(\Omega|X,\gamma)$$

Максимизация апостериорной вероятности (MAP) даёт точечную оценку Ω напрямую, без вывода Posterior:

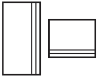
$$\Omega := \arg\max_{\Omega}\bigl(\ln p(X|\Omega) + \ln \mathrm{Prior}(\Omega|\gamma)\bigr)$$

Многокритериальная аддитивная регуляризация (ARTM) обобщает MAP на любые регуляризаторы и их комбинации:

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

Регуляризаторы для улучшения интерпретируемости тем

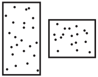
background



Сглаживание фоновых тем $B \subset T$:

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

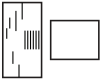
sparse



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

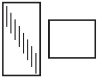
$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d} \sum_{t \in S} \alpha_t \ln \theta_{td}$$

seed words



Сглаживание для выделения релевантных тем
с помощью словаря «затравочных» ключевых слов

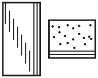
decorrelated



Декоррелирование для повышения различности тем:

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

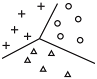
interpretable



Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

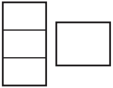
Регуляризаторы для мультимодальных тематических моделей

supervised



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

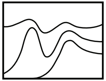
multilanguage



Модальность языков и регуляризация со словарём $\pi_{uwt} = p(u|w,t)$ переводов с языка k на ℓ:

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

temporal



Темпоральные модели с модальностью времени i:

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

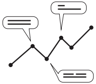
geospatial

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

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

Регуляризаторы для учёта дополнительной информации

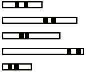
regression



Линейная модель регрессии $\hat{y}_d = \langle v, \theta_d \rangle$ документов:

$$R(\Theta, v) = -\tau \sum_{d \in D} \Big(y_d - \sum_{t \in T} v_t \theta_{td} \Big)^2$$

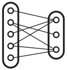
biterm



Связи сочетаемости слов (n_{uv} — частота битерма):

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt}$$

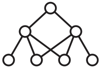
relational



Связи или ссылки между документами:

$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \sum_{t \in T} \theta_{td} \theta_{tc}$$

hierarchy



Связи родительских тем t с дочерними подтемами s:

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

Модульный подход к синтезу моделей с заданными свойствами

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

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

-- *нестандартизуемые этапы, уникальная разработка для каждой задачи*

-- *стандартизуемые этапы*

BigARTM: библиотека тематического моделирования

Ключевые возможности:

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

Сообщество:

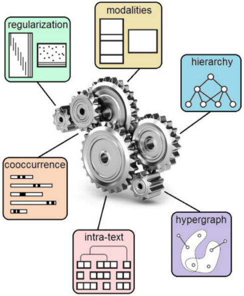
- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

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

Ключевые возможности библиотек BigARTM и TopicNet

BigARTM



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

TopicNet

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

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

Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

3.7M статей Википедии, 100K слов: время min (перплексия)

проц.	$\vert T\vert$	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.

Транзакционные данные

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

- **Данные социальной сети:**
  (d, u, w) — пользователь u записал слово w в блоге d
- **Данные сети интернет-рекламы:**
  (u, d, b) — пользователь u кликнул баннер b на странице d
- **Данные рекомендательной системы:**
  (u, f, s) — пользователь u оценил фильм f в ситуации s
- **Данные финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g
- **Данные о пассажирских авиаперелётах:**
  (u, a, b, c) — перелёт клиента u из a в b авиакомпанией c

Задача: по наблюдаемой выборке рёбер гиперграфа найти латентные тематические векторные представления его вершин.

Гиперграфовая транзакционная ARTM

n_{kdx} — частота транзакции (d, x), $x \subset W$ типа k в выборке E_k

Максимизация суммы log-правдоподобий с регуляризацией:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{kdx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{vt} \; + \; R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

ЕМ-алгоритм: метод простой итерации для системы уравнений

Е-шаг:
$$p_{tdx} = \operatorname*{norm}_{t \in T} \Big(\theta_{td} \prod_{v \in x} \phi_{vt} \Big)$$

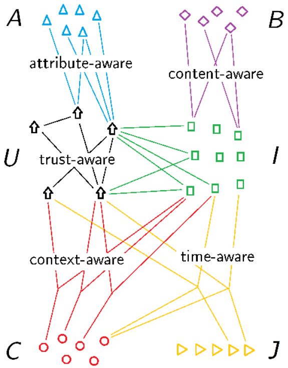
М-шаг:
$$\phi_{vt} = \operatorname*{norm}_{v \in W^m} \Big(\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} [v \in x] \, n_{kdx} p_{tdx} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \Big)$$

$$\theta_{td} = \operatorname*{norm}_{t \in T} \Big(\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{kdx} p_{tdx} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \Big)$$

K. Vorontsov. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization // Springer Optimization and Its Applications. 2023

Транзакционные данные в рекомендательных системах

U — конечное множество (словарь) клиентов (users)
I — конечное множество (словарь) объектов (items)
A — словарь атрибутов клиентов (соцдем, регион, хобби...)
B — словарь свойств объектов (слова в текстовых объектах)
C — словарь ситуативных контекстов
J — словарь интервалов времени



Возможные виды данных:

n_{ui} — клиент u выбрал объект i
n_{ua} — клиент u имеет атрибут a
n_{ib} — объект i имеет свойство b
n_{uv} — клиент u доверяет клиенту v
n_{uib} — клиент u отметил i тэгом b
n_{uic} — клиент u выбрал i в контексте c
n_{uicj} — u выбрал i в c в интервале j

Мультимодальная (гипер)графовая тематическая модель

Общая идея графовых эмбедингов и тематических моделей: наблюдаемые рёбра объясняются скрытыми векторами вершин

$$
\begin{aligned}
&\sum_{u,i} n_{ui} \ln \sum_{t\in T} p(t|u)\, p(t|i)\, p^{-1}(t) \\
+\,& \tau_1 \sum_{i,b} n_{ib} \ln \sum_{t\in T} p(t|i)\, p(t|b)\, p^{-1}(t) \\
+\,& \tau_2 \sum_{u,a} n_{ua} \ln \sum_{t\in T} p(t|u)\, p(t|a)\, p^{-1}(t) \\
+\,& \tau_3 \sum_{u,i,c} n_{uic} \ln \sum_{t\in T} p(t|i)\, p(t|u)\, p(t|c)\, p^{-2}(t) \\
+\,& \tau_4 \sum_{u,i,c,t} n_{uic\tau} \ln \sum_{t\in T} p(t|i)\, p(t|u)\, p(t|c)\, p(t|\tau)\, p^{-3}(t) \;\to\; \max
\end{aligned}
$$

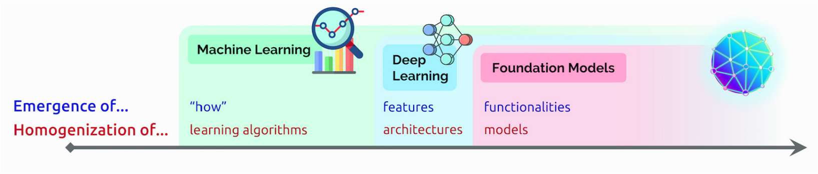
Оптимизация по всем эмбедингам — векторам вида $p(t|\bullet)$

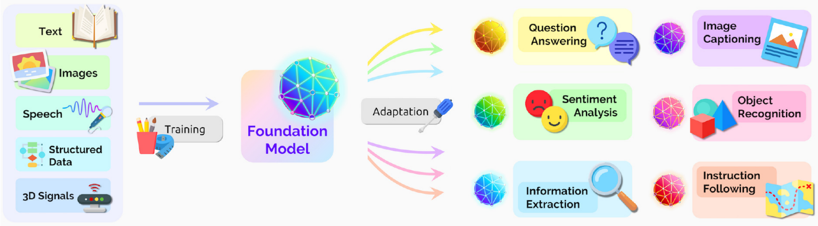
Для вероятностных тематических эмбедингов — EM-алгоритм

Для обычных эмбедингов — стохастический градиент

Обучаемая векторизация данных — глобальный тренд AI/ML

Foundation Models — гомогенизация векторных представлений





R.Bommasani et al. (Center for Research on Foundation Models, Stanford University)
On the opportunities and risks of foundation models // CoRR, 20 August 2021.

Гиперграфовые тематические модели языка

Гипер-рёбрами могут быть *сигментоиды* — подмножества термов, связанные по смыслу и порождаемые общей темой:

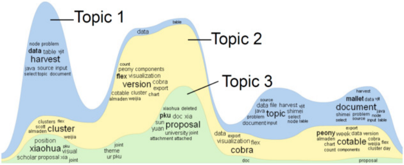
- предложение / фраза / синтагма
- ветка синтаксического дерева / именная группа
- факт «объект, субъект, действие»
- пары синонимов, гипоним–гипероним, мероним–холоним
- лексическая цепочка
- текст комментария, дата–время, автор

Модель даёт интерпретируемые тематические эмбединги:

- $p(t|d)$ — каждого контейнера, в частности, документа
- $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ — каждого терма, в частности, слова
- $p(t|d,x)$ — каждой отдельной транзакции (фразы, факта)

Мотивации. Что хотим:

- вместо «мешка слов» — последовательность $w_1, \dots, w_n$
- вместо документов — локальные контексты слов
- определять тематику любого фрагмента текста
- быстро находить фрагменты, относящиеся к данной теме
- в том числе фразы для суммаризации документа или темы
- разделять документ на тематически однородные сегменты
- визуализировать тематическую структуру документа



Идея тематизации текста за один проход

Дано: s — фрагмент текста d, Φ — тематическая модель
Найти: $p(t|s)$ — тематический вектор фрагмента текста

Проблемы:

- как не переобучить вектор $p(t|s)$, если текст короткий?
- как согласовать $p(t|s)$ с объемлющим контекстом $p(t|d)$?
- как согласовать $p(t|s)$ с $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ термов $w \in s$?

Наводящие соображения:

- первая итерация EM-алгоритма с инициализацией $\theta^0_{td} = \frac{1}{|T|}$:
$$\theta_{td}(\Phi) = \mathop{\mathrm{norm}}_{t\in T}\Bigl(\sum_{w\in W} n_{dw}p_{tdw}\Bigr) = \sum_{w\in d}\frac{n_{dw}}{n_d}\mathop{\mathrm{norm}}_{t\in T}\bigl(\phi_{wt}\theta^0_{td}\bigr)$$
- формула полной вероятности + гипотеза усл. независ.:
$$\theta_{td}(\Phi) = \sum_{w\in d} p(w|d)\,p(t|w,\cancel{d}) = \sum_{w\in d}\frac{n_{dw}}{n_d}\mathop{\mathrm{norm}}_{t\in T}\bigl(\phi_{wt}p_t\bigr)$$

EM-алгоритм для ARTM с явным выражением Θ через Φ

Максимизация логарифма правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$p_{tdw} \equiv p(t|d,w) = \mathop{\mathrm{norm}}_{t\in T}\big(\phi_{wt}\theta_{td}\big); \quad n_{td} = \sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}$$

$$p'_{tdw} = p_{tdw} + \frac{1}{n_{dw}} \sum_{s\in T} \frac{n_{sd}}{\theta_{sd}} \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}}$$

$$\phi_{wt} = \mathop{\mathrm{norm}}_{w\in W}\left(\sum_{d\in D} n_{dw} p'_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$$

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста. КиМ, 2020.

Доказательство (по Лемме о максимизации на симплексах)

Оптимизационная задача М-шага относительно Φ и $\Theta(\Phi)$:

$$Q(\Phi) = \sum_{d\in D}\sum_{u\in W}\sum_{s\in T} n_{du} p_{sdu} \ln\bigl(\phi_{us}\theta_{sd}(\Phi)\bigr) + R\bigl(\Phi, \Theta(\Phi)\bigr) \to \max_{\Phi}$$

Применим Лемму к регуляризованному log-правдоподобию Q:

$$\phi_{wt}\frac{\partial Q}{\partial \phi_{wt}} = \sum_{d\in D} n_{dw} p_{tdw} + \sum_{d,s,u} n_{du} p_{sdu} \frac{\phi_{wt}}{\theta_{sd}} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} + \phi_{wt} \sum_{d,s} \frac{\partial R}{\partial \theta_{sd}} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} =$$

$$= \sum_{d\in D} n_{dw}\left(p_{tdw} + \frac{1}{n_{dw}} \sum_{s\in T} \frac{\phi_{wt}}{\theta_{sd}} \underbrace{\left(\sum_{u\in d} n_{du} p_{sdu} + \theta_{sd}\frac{\partial R}{\partial \theta_{sd}}\right)}_{n_{sd}} \frac{\partial \theta_{sd}}{\partial \phi_{wt}}\right) + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} =$$

$$= \sum_{d\in D} n_{dw} \underbrace{\left(p_{tdw} + \frac{1}{n_{dw}} \sum_{s\in T} \frac{n_{sd}}{\theta_{sd}} \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}}\right)}_{p'_{tdw}} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}. \quad \blacksquare$$

EM-алгоритм для ARTM с линейной тематизацией документов

$$\theta_{td}(\Phi) = \sum_{w\in d} \frac{n_{dw}}{n_d} \operatorname*{norm}_{t\in T}(\phi_{wt} p_t) \Rightarrow \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} = \frac{n_{dw}}{n_d} \phi'_{tw} \big(\delta_{st} - \phi'_{sw}\big)$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} \equiv p(t|w) = \operatorname*{norm}_{t\in T}\big(\phi_{wt} n_t\big); \qquad \theta_{td} = \sum_{w\in d} \frac{n_{dw}}{n_d} \phi'_{tw}$$

$$p_{tdw} \equiv p(t|d,w) = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big); \qquad n_t = \sum_{d\in D}\sum_{w\in d} n_{dw} p_{tdw}$$

$$n_{td} = \sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}$$

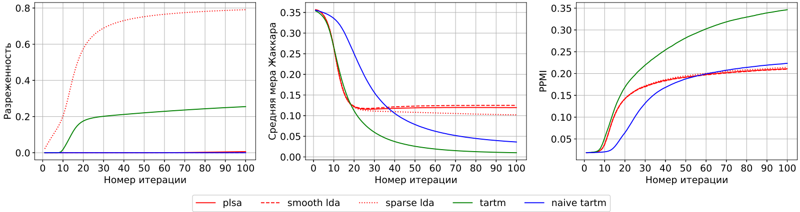
$$p'_{tdw} = p_{tdw} + \frac{\phi'_{tw}}{n_d}\left(\frac{n_{td}}{\theta_{td}} - \sum_{s\in T} \phi'_{sw} \frac{n_{sd}}{\theta_{sd}}\right)$$

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

Эксперимент. Проверка модифицированного EM-алгоритма

Коллекция NIPS, $|T| = 50$, модели:

- TARTM (Θless ARTM) — модифицированный EM-алгоритм
- naive TARTM — одна итерация обычного EM-алгоритма



- TARTM очищает темы от общеупотребительных слов,
- улучшает разреженность, различность и когерентность тем

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста, 2020.

https://github.com/ilirhin/python_artm

Упрощение EM-алгоритма для линейной тематизации

- Нет регуляризации по Θ, следовательно, $\frac{\partial R}{\partial \theta_{td}} = 0$
- Значение отношения $\frac{n_{td}}{\theta_{td}} \approx n_d$ не зависит от t, подстановка в формулу M-шага приводит к упрощению: $p'_{tdw} = p_{tdw}$

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt} n_t\big); \qquad \theta_{td} = \sum_{w\in d} \frac{n_{dw}}{n_d} \phi'_{tw};$$

$$p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big); \qquad n_t = \sum_{d\in D}\sum_{w\in d} n_{dw} p_{tdw};$$

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

Это обычный EM-алгоритм, только с однопроходным E-шагом!
ОГО! И ТАК МОЖНО БЫЛО?!

Линейная тематизация: от документа к локальным контекстам

Тематизация документа $d = (w_1, \dots, w_{n_d})$ за один проход:

$$\theta_{td}(\Phi) \equiv p(t|d) = \frac{1}{n_d} \sum_{i=1}^{n_d} p(t|w_i) = \frac{1}{n_d} \sum_{i=1}^{n_d} \phi'_{tw_i}$$

Тематизация *локального контекста* $C_i = (\dots, w_i, \dots)$ терма w_i:

$$\theta_{ti}(\Phi) \equiv p(t|C_i) = \frac{1}{|C_i|} \sum_{u \in C_i} p(t|u) = \frac{1}{|C_i|} \sum_{u \in C_i} \phi'_{tu}$$

Тематизация локального контекста с распределением весов:

$$\theta_{ti}(\Phi) \equiv p(t|C_i) = \sum_{u \in C_i} \phi'_{tu} \alpha(u|i), \quad \sum_{u \in C_i} \alpha(u|i) = 1, \quad \alpha(u|i) \geqslant 0$$

Локализованная тематическая модель:

$$p(w|C_i) = \sum_{t \in T} p(w|t)\, p(t|C_i) = \sum_{t \in T} \phi_{wt} \sum_{u \in C_i} \phi'_{tu} \alpha(u|i)$$

EM-алгоритм с локализованным E-шагом

$w_1, \dots, w_n$ — сквозная нумерация термов во всей коллекции
C_i — локальный контекст (окружение) терма w_i
$\alpha(u|i)$ — распределение важности термов $u \in C_i$ для терма w_i

- не нужна гипотеза «мешка слов»
- не нужно разбиение коллекции на документы

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} \equiv p(t|w) = \operatorname*{norm}_{t \in T}\bigl(\phi_{wt} p_t\bigr); \qquad \theta_{ti} \equiv p(t|C_i) = \sum_{u \in C_i} \phi'_{tu} \alpha(u|i);$$

$$p_{ti} \equiv p(t|C_i, w_i) = \operatorname*{norm}_{t \in T}\bigl(\phi_{w_i t} \theta_{ti}\bigr); \qquad p_t \equiv p(t) = \frac{1}{n} \sum_{i=1}^{n} p_{ti};$$

$$\phi_{wt} = \operatorname*{norm}_{w \in W}\left(\sum_{i=1}^{n} [w_i = w] p_{ti} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\right).$$

Быстрое вычисление двунаправленных векторов контекста

Два прохода по тексту — «слева направо» и «справа налево» для вычисления *экспоненциальных скользящих средних* (ЭСС):

$$\vec{p}(t|i) = \vec{\gamma}_i\, p(t|w_i) + (1-\vec{\gamma}_i)\, \vec{p}(t|i-1), \;\; i = 1,\dots,n, \;\; \vec{\gamma}_1 = 1$$

$$\overleftarrow{p}(t|i) = \overleftarrow{\gamma}_i\, p(t|w_i) + (1-\overleftarrow{\gamma}_i)\, \overleftarrow{p}(t|i+1), \;\; i = n,\dots,1, \;\; \overleftarrow{\gamma}_n = 1$$

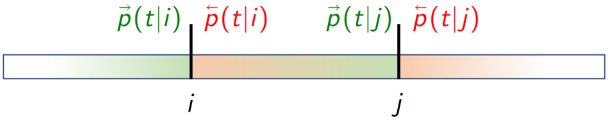
где $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — коэффициенты сглаживания в позиции i

Основное свойство: если $\gamma_i = \gamma$, то $\alpha(w_k|i) = \gamma(1-\gamma)^{|i-k|}$

Несколько соображений, как распоряжаться выбором $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$:

- $\gamma_i \approx \frac{1}{h}$, где h — ширина окна, размер контекста
- $\gamma_i = 1$, если надо забыть контекст, сменить документ
- $\gamma_i = 0$, если надо проигнорировать терм
- γ_i можно умножать на оценку важности терма

Использование двунаправленных векторов контекста



Через *двунаправленные тематические векторы* определяется:

- $\vec{p}(t|i)$ — тематика левого контекста терма w_i
- $\overleftarrow{p}(t|i)$ — тематика правого контекста терма w_i
- $\frac{1}{2}\big(\vec{p}(t|i) + \overleftarrow{p}(t|i)\big)$ — тематика двустороннего контекста w_i
- $p(t|i\dots j) = \frac{1}{2}\big(\overleftarrow{p}(t|i) + \vec{p}(t|j)\big)$ — тематика сегмента $[i\dots j]$
- $\overleftarrow{p}(t|i) \approx \vec{p}(t|j)$ — однородность тематики сегмента $[i\dots j]$
- $\max_i \big\|\vec{p}(t|i) - \overleftarrow{p}(t|i)\big\|$ — граница i между сегментами
- при различных γ_i — короткие и длинные контексты

Гипотеза: есть аналогия с моделью внимания и трансформером

Онлайновый EM-алгоритм с локализованным E-шагом

Вход: коллекция, число тем $|T|$, параметры β, $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, α, δ;
Выход: матрица Φ, векторы термов документов p_{ti};

инициализация: $n_{wt} := 0$; $\tilde{n}_{wt} := 0$; $n_t := 1$; $\phi_{wt} :=$ random;
для всех документов $d \in D$

- $p_{ti} := \operatorname{norm}_t(\phi_{w_i t} n_t), \quad i = 1, \dots, n_d, \quad t \in T$;
- $\vec{\theta}_{ti} := \vec{\gamma}_i\, p_{ti} + (1 - \vec{\gamma}_i)\, \vec{\theta}_{t,i-1}, \quad i = 1, \dots, n_d, \quad \vec{\gamma}_1 = 1, \quad t \in T$;
- $\overleftarrow{\theta}_{ti} := \overleftarrow{\gamma}_i\, p_{ti} + (1 - \overleftarrow{\gamma}_i)\, \overleftarrow{\theta}_{t,i+1}, \quad i = n_d, \dots, 1, \quad \overleftarrow{\gamma}_{n_d} = 1, \quad t \in T$;
- $p_{ti} := \operatorname{norm}_t\bigl(\phi_{w_i t}\bigl(\beta \vec{\theta}_{ti} + (1-\beta)\overleftarrow{\theta}_{ti}\bigr)\bigr), \quad i = 1, \dots, n_d, \quad t \in T$;
- $\tilde{n}_{w_i t} := \tilde{n}_{w_i t} + p_{ti}; \quad n_t := n_t + p_{ti}, \quad i = 1, \dots, n_d, \quad t \in T$;
- **если** пора обновить матрицу Φ **то**
  - $n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}; \quad \tilde{n}_{wt} := 0$;
  - $\phi_{wt} := \operatorname*{norm}_{w \in W}\left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\right)$;

Модель внимания Query–Key–Value

q — *вектор-запрос*, трансформируемый в контекстный вектор z.
Контекст задаётся последовательностью n пар ключ-значение:
$K = (k_1, \dots, k_n)$ — *векторы-ключи*,
$V = (v_1, \dots, v_n)$ — *векторы-значения*.

Модель внимания — это выпуклая комбинация векторов v_i, взвешенных по сходству их ключей k_i с запросом q:

$$z = \mathrm{Attn}(q, K, V) = \sum_{i=1}^{n} v_i \operatorname*{SoftMax}_i \langle k_i, q \rangle$$

Модель само-внимания (self-attention) трансформирует
$X = (x_1, \dots, x_n)$ — входные *бесконтекстные векторы* в
$Z = (z_1, \dots, z_n)$ — выходные *контекстные векторы*:

$$z_i = \mathrm{Attn}(W_q x_i, W_k X, W_v X),$$

где W_q, W_k, W_v — обучаемые матрицы параметров.

Vaswani et al. Attention is all you need. 2017.

Аналогия локализованного E-шага с моделью само-внимания

Контекстный тематический вектор на выходе E-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \underset{t \in T}{\mathrm{norm}}\left(\phi_{w_i t}\theta_{ti}\right) = \underset{t \in T}{\mathrm{norm}}\Biggl(\sum_{u \in C_i} \phi'_{tu}\phi_{w_i t}\alpha(u|i)\Biggr)$$

Контекстный вектор на выходе модели само-внимания:

$$z_i = \sum_{u \in C_i} W_v x_u \, \alpha(u|i) = \sum_{u \in C_i} W_v x_u \underset{u \in C_i}{\mathrm{SoftMax}} \langle W_q x_i, W_k x_u \rangle$$

Сходство:

- вектор терма w_i трансформируется в контекстный вектор
- путём усреднения векторов ϕ'_u из контекста терма w_i,
- наиболее (семантически) схожих с вектором терма w_i.

Отличия:

- адамарово умножение вектора ϕ'_u на вектор-фильтр ϕ_{w_i};
- нет обучаемых матриц W_q, W_k, W_v как у модели внимания;
- проецирование итогового вектора на единичный симплекс.

Аналогия локализованного E-шага с моделью трансформера

Один проход документа аналогичен модели внимания:

— для каждого $d \in D$, для каждой позиции $i = 1, \dots, n_d$ вычисляются 5 тематических векторов, связанных с термом w_i:

$\phi'_{tw_i} = \mathrm{norm}_t(\phi_{w_i t} p_t)$ — бесконтекстный вектор терма $p(t|w_i)$

$\vec{p}(t|i) = \vec{\theta}_{ti}$, $\overleftarrow{p}(t|i) = \overleftarrow{\theta}_{ti}$ — векторы левого и правого контекста

$\theta_{ti} = \beta \vec{\theta}_{ti} + (1 - \beta) \overleftarrow{\theta}_{ti}$ — вектор двустороннего контекста

$p_{ti} = \mathrm{norm}_t\big(\phi_{w_i t} \theta_{ti}\big)$ — контекстный вектор терма $p(t|C_i, w_i)$

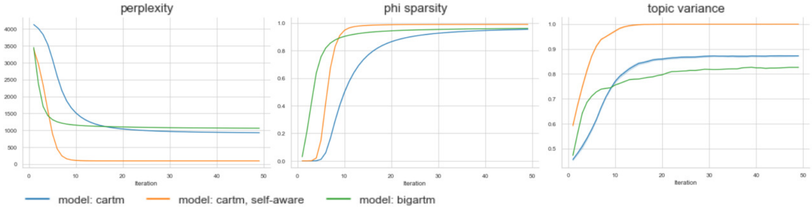
Несколько таких проходов аналогичны трансформеру:

контекстный вектор терма $p_{ti} = p(t|C_i, w_i)$ с предыдущего прохода используется вместо его бесконтекстного вектора $\phi'_{tw_i} = p(t|w_i)$

L таких итераций аналогичны проходу L блоков внимания

Первые эксперименты с реализацией Context-ARTM

Коллекция «20 Newsgroups»: $|D| = 18846$, $|W| = 107672$.



— улучшилась перплексия, разреженность Φ, различность тем

model	time@10 topics	time@30 topics	time@70 topics	time@100 topics
CARTM@CPU	4min 55s ± 1.7s	9min 20s ± 9.52s	20min 37s ± 2.36s	25min 52s ± 4.63s
BigARTM	1min 30s ± 1.6s	3min 5s ± 1.98s	4min 55s ± 653ms	6min 22s ± 5.81s

— время хуже в несколько раз, при этом реализация CARTM на Python/JAX, тогда как ядро BigARTM на C++

В сухом остатке: что сделано, и что дальше

- Теория ARTM — оптимизация без байесовского обучения
- Оптимизация на симплексах уже реализована в pyTorch
- Гиперграфовая модель транзакционных данных
- Предельно упрощённая тематическая модель внимания

- Как параметризовать тематическую модель внимания?
- Будет ли полезен тематический трансформер?
- Не выделить ли вероятностную интерпретируемую часть эмбедингов в нейросетевых языковых моделях?

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

Воронцов К.В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. 2025 http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

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

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, Wray Buntine. Topic modelling meets deep neural networks: A survey. 2021