

Вероятностное тематическое моделирование в эпоху больших языковых моделей

Воронцов Константин Вячеславович

k.vorontsov@iai.msu.ru

д.ф.-м.н., профессор РАН • зав. каф. ММП ВМК МГУ,
руководитель лаборатории МОиСА Института ИИ МГУ,
зав. каф. МОЦГ МФТИ • г.н.с. ФИЦ ИУ РАН

Научная конференция • 100-летие академика О. М. Белоцерковского
«Численное моделирование в механике сплошных сред»
Секция «Машинное обучение и искусственный интеллект»
МФТИ • 29 ноября 2025

1 Вероятностное тематическое моделирование

- постановка задачи
- примеры приложений
- сравнение с большими языковыми моделями

2 Теория аддитивной регуляризации

- основная лемма
- отказ от байесовского обучения
- небайесовская формализация постановок задач

3 От «мешка слов» к тематическому вниманию

- окончательный отказ от гипотезы «мешка слов»
- аналогии с нейросетевыми моделями языка
- нейросетевые тематические модели

Задача вероятностного тематического моделирования

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

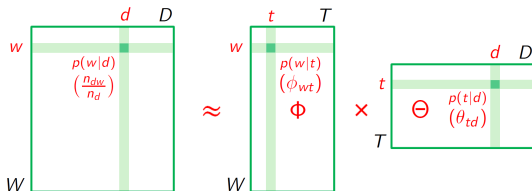
- $p(w|d) = \sum_{t \in T} p(w|\cancel{d}, t) p(t|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$
- $p(w|t) = \phi_{wt}$ — из каких слов w состоит каждая тема $t \in T$
- $p(t|d) = \theta_{td}$ — из каких тем t состоит каждый документ d

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

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

Три интерпретации задачи тематического моделирования

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



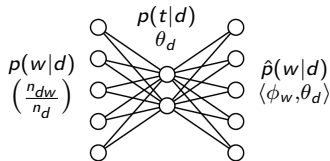
3. **Автокодировщик** документов в тематические эмбединги:

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

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

задача реконструкции:

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



Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей, число тем $|T| = 400$
Первые 10 слов и их частоты $p(w|t)$ в %:

Тема №68				Тема №79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей, число тем $|T| = 400$
Первые 10 слов и их частоты $p(w|t)$ в %:

Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример 2. Биграммная модель научных конференций

Коллекция 1000 статей конференций ММРО, ИОИ на русском

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

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

Цели и не-цели тематического моделирования

Цели:

- выявлять тематическую кластерную структуру текстовой коллекции (сколько в ней тем, и о чём они), представляя 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, классификации, фильтрации, сегментации, суммаризации текстов

Не-цели:

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

Некоторые приложения Тематического Моделирования (ТМ)

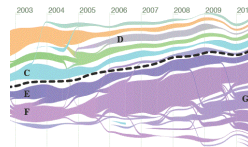
разведочный поиск в
электронных библиотеках



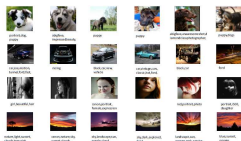
поиск тематических
сообществ в соцсетях



выявление и отслеживание
цепочек новостей



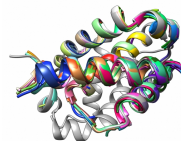
мультимодальный поиск
текстов и изображений



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.

Сходства и отличия от LLM

PTM и LLM — что общего

- языковая модель, которая предсказывает слова в тексте
- автокодировщик, который переводит текст в эмбединг
- обобщаются на мультимодальные, мультязычные данные
- возможно многозадачное, многокритериальное обучение

PTM — принципиальные отличия от LLM

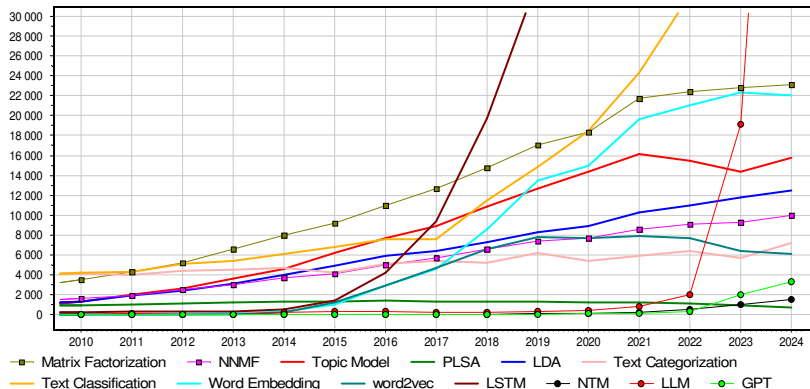
- намного более слабая языковая модель
- эмбединги вероятностные, разреженные, интерпретируемые
- простота и скорость матричного разложения

PTM — дальнейшее развитие навстречу LLM

- отказ от байесовского обучения → алгоритм ближе к SGD
- транзакционные данные → ближе к foundation models
- отказ от мешка слов → ближе к модели внимания

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

Максимизация 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 $|I_j|$:

[illegible]

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

$$\begin{cases} f(\Omega) \rightarrow \max_{\Omega}; \\ \sum_{i \in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geq 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

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: $p_i = \text{norm}_{i \in I}(x_i) = \frac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$

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

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

- 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: $f(\Omega) \rightarrow \max_{\Omega}; \quad \sum_{i \in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geq 0, \quad i \in I_j, \quad j \in J.$

Функция Лагранжа:

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

Условия Каруша–Куна–Таккера для вектора ω_j :

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

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 $\omega_{ij} \lambda_j = (A_{ij})_+; \quad \lambda_j = \sum_i (A_{ij})_+ \Rightarrow \omega_{ij} = \text{norm}_i(A_{ij}).$

■

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

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

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)

Аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

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

Е-шаг совпадает с формулой Байеса: $p(t|d, w) = \frac{p(w|t)p(t|d)}{p(w|d)}$

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:

$$\begin{aligned}
 f(\Phi, \Theta) &= \sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta} \\
 \phi_{wt} &= \text{norm}_{w \in W} \left(\phi_{wt} \frac{\partial f}{\partial \phi_{wt}} \right) = \text{norm}_{w \in W} \left(\phi_{wt} \sum_{d \in D} n_{dw} \frac{\theta_{td}}{p(w|d)} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) = \\
 &= \text{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right); \\
 \theta_{td} &= \text{norm}_{t \in T} \left(\theta_{td} \frac{\partial f}{\partial \theta_{td}} \right) = \text{norm}_{t \in T} \left(\theta_{td} \sum_{w \in W} n_{dw} \frac{\phi_{wt}}{p(w|d)} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) = \\
 &= \text{norm}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right);
 \end{aligned}$$

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

PLSA, LDA: первые и 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:

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



Thomas
Hofmann

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 со смещением β_w, α_t :

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



David Blei

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. NIPS-2001. JMLR 2003.

От байесовского вывода к аддитивной регуляризации

X — исходные данные, $\Omega = (\Phi, \Theta)$ — параметры модели

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

$$\text{Posterior}(\Omega|X, \gamma) = \frac{p(X|\Omega) \text{Prior}(\Omega|\gamma)}{\int p(X|\Omega) \text{Prior}(\Omega|\gamma) d\Omega}$$

$$\Omega := \arg \max_{\Omega} \text{Posterior}(\Omega|X, \gamma)$$

Максимизация апостериорной вероятности (MAP)

даёт точечную оценку Ω напрямую, без вывода Posterior:

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

Многокритериальная аддитивная регуляризация (ARTM)

обобщает MAP на любые регуляризаторы и их комбинации:

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

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	

-- нестандартизируемые этапы, уникальная разработка для каждой задачи

-- стандартизуемые этапы

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

Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

3.7М статей Википедии, 100К слов: время min (перплексия)

проц.	$ T $	Gensim	Vowpal Wabbit	BigARTM	BigARTM асинхрон
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

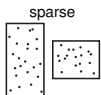
D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov. Fast and modular regularized topic modelling. FRUCT ISMW, 2017.

Регуляризаторы для улучшения интерпретируемости тем



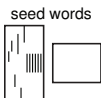
Сглаживание фоновых тем $B \subset T$:

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

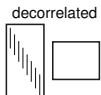


Разреживание предметных тем $S = T \setminus B$:

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

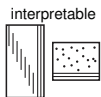


Сглаживание для выделения релевантных тем с помощью словаря «затравочных» ключевых слов



Декоррелирование для повышения различности тем:

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



Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем

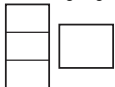
Регуляризаторы для мультимодальных тематических моделей

supervised



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

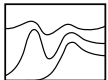
multilanguage



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 :

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

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 взаимосвязей и зависимостей

regression



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

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

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 :

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

Регуляризаторы для моделирования последовательного текста

sentence



Тематические модели, учитывающие границы предложений, абзацев и секций документов

n-gram



Модели с модальностями n -грамм, коллокаций, именованных сущностей (используем TopMine)

syntax



Модели, учитывающие результаты автоматического синтаксического разбора (используем UDPipe)

sentiment



Модели выделения мнений на основе тональностей, фактов, семантических ролей именованных сущностей

segmentation



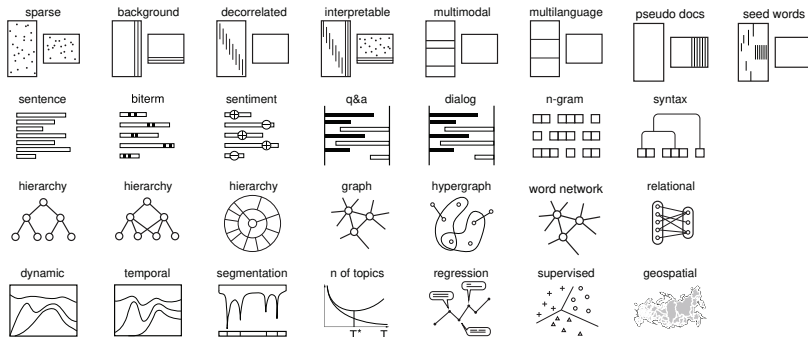
Тематические модели сегментации с автоматическим определением границ сегментов

Язык пиктограмм для описания комбинаций моделей

Структуры матричных разложений в вероятностных моделях:



Регуляризаторы — дополнительные критерии и ограничения:



Разведочный поиск в технологических блогах

Цель: поиск документов

по длинным текстовым запросам

— Habr.ru (175K документов),

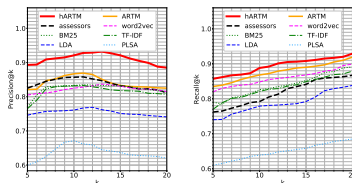
— TechCrunch.com (760K док.).

Регуляризаторы:

$$\mathcal{L}\left(\begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array}\right) + R\left(\begin{array}{c} \text{hierarchy} \\ \text{graph} \end{array}\right) + R\left(\begin{array}{c} \text{interpretable} \\ \text{matrix} \end{array}\right) + R\left(\begin{array}{c} \text{multimodal} \\ \text{image} \end{array}\right) + R\left(\begin{array}{c} \text{n-gram} \\ \text{tokens} \end{array}\right) \rightarrow \max$$

Результаты:

- Точность и полнота **93%**, превосходит ассессоров и другие методы (tf-idf, BM25, word2vec, PLSA, LDA, ARTM).
- Увеличилась оптимальная размерность векторов:
200 → 1400 (Habr.ru), 475 → 2800 (TechCrunch.com).



А.Янина. Тематические и нейросетевые модели языка для разведочного информационного поиска // диссертация к.ф.-м.н. МФТИ, 2022.

Поиск и рубрикация научных публикаций на 100 языках

Цель: мультязычный поиск
и классификация научных
публикаций по рубрикам
УДК, ГРНТИ, ОЭСР, ВАК

модель	ср.ч. УДК	ср.% УДК	ср.ч. ГРНТИ	ср.% ГРНТИ
Базовая TM	0.558	0.165	0.536	0.220
XLM-RoBERTa	0.835	0.179	0.832	0.288
ARTM	0.995	0.225	0.852	0.366

Регуляризаторы:

$$\mathcal{L}\left(\begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array}\right) + R\left(\begin{array}{c} \text{interpretable} \\ \text{[stack of bar charts]} \end{array}\right) + R\left(\begin{array}{c} \text{multimodal} \\ \text{[stack of images]} \end{array}\right) + R\left(\begin{array}{c} \text{multilanguage} \\ \text{[stack of boxes]} \end{array}\right) + R\left(\begin{array}{c} \text{supervised} \\ \text{[scatter plot]} \end{array}\right) \rightarrow \max$$

Результаты:

- точность мультязычного поиска 94%
- сокращение модели 128 Гб → 4.8 Гб при редукции словарей (ВРЕ-токенизация) до 11K токенов на каждый язык.

Н.А.Герасименко, П.С.Потапова, А.О.Янина, К.В.Воронцов. Применение вероятностного тематического моделирования в четырёх задачах разведочного информационного поиска // Информационный бюллетень РБА, 2022, №98, С.43–48.

Поиск и классификация этно-релевантных тем в соцсетях

Цель: выявление как можно большего числа тем о национальностях и межнациональных отношениях (затравка — словарь 300 этнонимов).

Регуляризаторы:

(японцы): японский, япония, корей, китайский, жилища, авария, фукусима, цунами, сообщать, океан, станция, катюка, район, правительство, атомный, (норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,
(венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадуро, боливия, фидель, глава, латиноский, венесуэльский, лидер, боливарианской, президентский, алиенде, гевару,
(китайцы): китайский, россия, производство, китай, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кир, (азербайджанцы): русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, москва, страна, армянин, слово, рынок,
(грузины): грузинский, спецназ, военный, август, баташева, российский, спецназовец, миротворец, операция, дурын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвалы,
(осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алатай, российский, население, конфликт, (цыгане): народити, цыган, цыганка, хороший, место, страна, денюга, время, работата, жилье, жить, рука, дом, цыганский, наркоманка,

$$\mathcal{L}\left(\begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array}\right) + R\left(\begin{array}{c} \text{seed words} \\ \text{[bar chart]} \quad \square \end{array}\right) + R\left(\begin{array}{c} \text{interpretable} \\ \text{[bar chart]} \quad \text{[scatter plot]} \end{array}\right) + R\left(\begin{array}{c} \text{multimodal} \\ \text{[stack of images]} \quad \square \end{array}\right) \\ + R\left(\begin{array}{c} \text{temporal} \\ \text{[waveform]} \end{array}\right) + R\left(\begin{array}{c} \text{geospatial} \\ \text{[map of Russia]} \end{array}\right) + R\left(\begin{array}{c} \text{sentiment} \\ \text{[sentiment scale]} \end{array}\right) \rightarrow \max$$

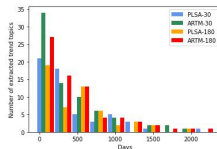
Результаты: число релевантных тем: 45 (LDA) $\rightarrow$ 83 (ARTM).

M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko, K. Vorontsov. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI, 2016.

–, –, –, –. Mining ethnic content online with additively regularized topic models. 2016.

Выявление трендов в коллекции научных публикаций

Цель: раннее обнаружение трендовых тем с начальным экспоненциальным ростом в области AI/ML 2009–2021 гг.



Регуляризаторы:

$$\mathcal{L}\left(\begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array}\right) + R\left(\begin{array}{c} \text{interpretable} \\ \text{[stack of bars]} \quad \text{[matrix]} \end{array}\right) + R\left(\begin{array}{c} \text{dynamic} \\ \text{[line graph]} \end{array}\right) + R\left(\begin{array}{c} \text{multimodal} \\ \text{[stack of boxes]} \quad \text{[box]} \end{array}\right) + R\left(\begin{array}{c} \text{n-gram} \\ \text{[matrix]} \end{array}\right) \rightarrow \max$$

Результаты:

- выделение 90 из 91 тренда в области машинного обучения
- 63% тем выделяется за год, 79% за два года

Н.Герасименко, А.Чернявский, М.Никифорова, М.Никитин, К.Воронцов.

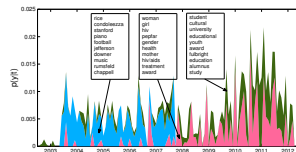
Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях. Доклады РАН, 2022.

Выявление динамики тем в новостных потоках

Цель: выделение тем в коллекции пресс-релизов МИДов 4х стран, с привязкой ко времени.

Регуляризаторы:

$$\mathcal{L}\left(\begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array}\right) + R\left(\begin{array}{c} \text{interpretable} \\ \begin{array}{|c|} \hline \text{bar chart} \\ \hline \end{array} \quad \begin{array}{|c|} \hline \text{scatter plot} \\ \hline \end{array} \right) + R\left(\begin{array}{c} \text{temporal} \\ \begin{array}{|c|} \hline \text{line graph} \\ \hline \end{array} \right) + R\left(\begin{array}{c} \text{multimodal} \\ \begin{array}{|c|} \hline \text{stacked bars} \\ \hline \end{array} \quad \begin{array}{|c|} \hline \text{empty box} \\ \hline \end{array} \right) \\ + R\left(\begin{array}{c} \text{n-gram} \\ \begin{array}{|c|c|c|} \hline \square \square \square \square \\ \hline \square \square \square \square \\ \hline \square \square \square \square \end{array} \right) + R\left(\begin{array}{c} \text{multilanguage} \\ \begin{array}{|c|} \hline \text{stacked bars} \\ \hline \end{array} \quad \begin{array}{|c|} \hline \text{empty box} \\ \hline \end{array} \right) \rightarrow \max$$



Результаты:

- разделение тем на событийные и перманентные
- когерентность тем: $5.5 \rightarrow 6.5$

Н.Дойков. Адаптивная регуляризация вероятностных тематических моделей.
ВКР бакалавра, ВМК МГУ, 2015.

Выделение поляризованных мнений в политических новостях

Цель: найти признаки, по которым
событийная тема разделяется
на кластеры-мнения

Регуляризаторы:

Modalities	<i>Pr</i>	<i>Rec</i>	<i>F1</i>
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	0.77	0.97	0.86

$$\mathcal{L}\left(\begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array}\right) + R\left(\begin{array}{c} \text{interpretable} \\ \text{matrix} \end{array}\right) + R\left(\begin{array}{c} \text{multimodal} \\ \text{stack} \end{array}\right) + R\left(\begin{array}{c} \text{n-gram} \\ \text{matrix} \end{array}\right) + R\left(\begin{array}{c} \text{syntax} \\ \text{tree} \end{array}\right) \rightarrow \max$$

Результаты:

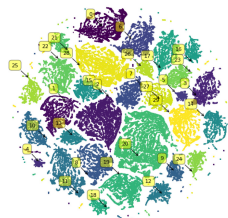
- выделение мнений внутри тем: F1-мера = 0.86%
- совместное использование трёх модальностей:
 - SPO — факты как триплеты «субъект–предикат–объект»
 - FR — семантические роли слов по Филлмору
 - Sent — тональности именованных сущностей

D.Feldman, T.Sadekova, K.Vorontsov. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. Dialogue 2020.

Тематическая модель банковских транзакционных данных

Цель: Выявление паттернов потребительского поведения клиентов банка, причём

- документы $\rightarrow$ клиенты,
- слова $\rightarrow$ MCC-коды продавцов.



Регуляризаторы:

$$\mathcal{L}\left(\begin{array}{|c|} \hline \text{PLSA} \\ \hline \Phi \quad \Theta \\ \hline \end{array}\right) + R\left(\begin{array}{|c|} \hline \text{interpretable} \\ \hline \text{[Bar Chart Icon]} \quad \text{[Scatter Plot Icon]} \\ \hline \end{array}\right) + R\left(\begin{array}{|c|} \hline \text{multimodal} \\ \hline \text{[Stacked Bar Chart Icon]} \quad \text{[Empty Box Icon]} \\ \hline \end{array}\right) + R\left(\begin{array}{|c|} \hline \text{supervised} \\ \hline \text{[Decision Tree Icon]} \\ \hline \end{array}\right) \rightarrow \max$$

Результаты:

- темы — паттерны потребительского поведения
- предсказание пола, возраста, достатка клиентов

E.Egorov, F.Nikitin, A.Goncharov, V.Alekseev, K.Vorontsov. Topic modelling for extracting behavioral patterns from transactions data. 2019.

Поиск обсуждений болезней в социальных сетях

Цель: поиск сообщений в Twitter о болезнях, симптомах, способах лечения, побочных эффектах.

Регуляризаторы:



$$\mathcal{L}\left(\begin{array}{|c|} \hline \text{PLSA} \\ \hline \Phi \quad \Theta \\ \hline \end{array}\right) + R\left(\begin{array}{|c|} \hline \text{seed words} \\ \hline \text{[Bar Chart]} \quad \text{[Box]} \\ \hline \end{array}\right) + R\left(\begin{array}{|c|} \hline \text{interpretable} \\ \hline \text{[Bar Chart]} \quad \text{[Scatter Plot]} \\ \hline \end{array}\right) + R\left(\begin{array}{|c|} \hline \text{multimodal} \\ \hline \text{[Table]} \quad \text{[Box]} \\ \hline \end{array}\right) +$$

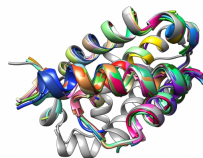
$$+ R\left(\begin{array}{|c|} \hline \text{hierarchy} \\ \hline \text{[Tree Diagram]} \\ \hline \end{array}\right) + R\left(\begin{array}{|c|} \hline \text{temporal} \\ \hline \text{[Line Graph]} \\ \hline \end{array}\right) + R\left(\begin{array}{|c|} \hline \text{geospatial} \\ \hline \text{[Map]} \\ \hline \end{array}\right) \rightarrow \max$$

Модель ATAM (Ailment Topic Aspect Model) похожа на поиск этно-релевантных тем и легко реализуема в BigARTM

Обработка последовательностей нуклеотидов или аминокислот

Цель: поиск мотивов и предсказание функций по нуклеотидным или аминокислотным последовательностям.

Регуляризаторы (гипотеза):



$$\begin{aligned} \mathcal{L} \left(\begin{array}{c} \text{PLSA} \\ \left(\begin{array}{|c|} \hline \Phi \\ \hline \end{array} \begin{array}{|c|} \hline \Theta \\ \hline \end{array} \right) \end{array} \right) + R \left(\begin{array}{c} \text{seed words} \\ \left(\begin{array}{|c|} \hline \text{seed words} \\ \hline \end{array} \begin{array}{|c|} \hline \\ \hline \end{array} \right) \end{array} \right) + R \left(\begin{array}{c} \text{interpretable} \\ \left(\begin{array}{|c|} \hline \text{interpretable} \\ \hline \end{array} \begin{array}{|c|} \hline \text{interpretable} \\ \hline \end{array} \right) \end{array} \right) + \\ + R \left(\begin{array}{c} \text{multimodal} \\ \left(\begin{array}{|c|} \hline \text{multimodal} \\ \hline \end{array} \begin{array}{|c|} \hline \\ \hline \end{array} \right) \end{array} \right) + R \left(\begin{array}{c} \text{hierarchy} \\ \left(\begin{array}{c} \text{hierarchy} \end{array} \right) \end{array} \right) + R \left(\begin{array}{c} \text{n-gram} \\ \left(\begin{array}{|c|} \hline \text{n-gram} \\ \hline \end{array} \begin{array}{|c|} \hline \text{n-gram} \\ \hline \end{array} \right) \end{array} \right) + R \left(\begin{array}{c} \text{segmentation} \\ \left(\begin{array}{|c|} \hline \text{segmentation} \\ \hline \end{array} \begin{array}{|c|} \hline \text{segmentation} \\ \hline \end{array} \right) \end{array} \right) \rightarrow \max \end{aligned}$$

Такая модель легко реализуема в BigARTM.

J.B.Gutierrez, K.Nakai. A study on the application of topic models to motif finding algorithms. 2016.

Lin Liu, Lin Tang, Libo He, Shaowen Yao, Wei Zhou. Predicting protein function via multi-label supervised topic model on gene ontology. 2017.

Lin Liu, Lin Tang, Xin Jin, Wei Zhou. A multi-label supervised topic model conditioned on arbitrary features for gene function prediction. 2019

Функциональная аннотация генома человека

Цель: прогноз тканеспецифических функций некодирующих генетических вариантов для каждой позиции в геноме человека в 127 различных тканях и типах клеток.

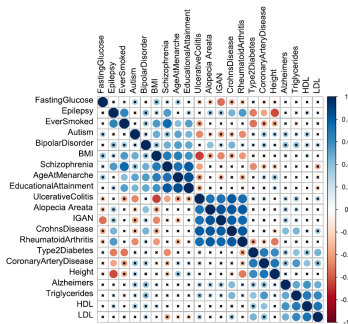
Регуляризаторы (гипотеза):

$$\mathcal{L}\left(\begin{array}{c} \text{PLSA} \\ \left[\begin{array}{c} \Phi \\ \Theta \end{array} \right] \end{array}\right) + R\left(\begin{array}{c} \text{supervised} \\ \text{diagram with points and lines} \end{array}\right) +$$

$$+ R\left(\begin{array}{c} \text{interpretable} \\ \begin{array}{|c|c|} \hline \text{[Bar Chart]} & \text{[Scatter Plot]} \\ \hline \end{array} \end{array}\right) + R\left(\begin{array}{c} \text{multimodal} \\ \begin{array}{|c|c|} \hline \text{[Text Box]} & \text{[Image Box]} \\ \hline \end{array} \end{array}\right) + R\left(\begin{array}{c} \text{hierarchy} \\ \begin{array}{c} \text{[Tree Diagram]} \end{array} \end{array}\right) + R\left(\begin{array}{c} \text{n-gram} \\ \begin{array}{|c|c|c|} \hline \text{[Box]} & \text{[Box]} & \text{[Box]} \\ \hline \text{[Box]} & \text{[Box]} & \text{[Box]} \\ \hline \text{[Box]} & \text{[Box]} & \text{[Box]} \end{array} \end{array}\right) \rightarrow \max$$

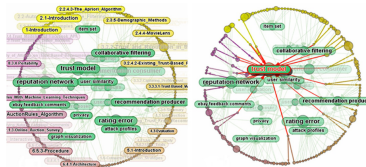
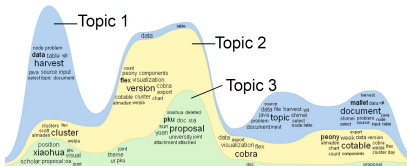
Такая модель легко реализуема в BigARTM.

D.Backenroth et al. FUN-LDA: a latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. 2018



Мотивации: что хотим от PTMs глядя на LLM (BERT, GPT)

- 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



Сегментация текста и пост-обработка Е-шага

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

Тематика термов в документе $p(t|d, w_i)$ — матрица $T \times n_d$:



Регуляризация Е-шага

Трёхмерная матрица $\Pi = (p_{tdw} = p(t|d, w))_{T \times D \times W}$

Регуляризатор Е-шага: $\tilde{R}(\Phi, \Theta) = R(\Pi(\Phi, \Theta), \Phi, \Theta)$:

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\begin{aligned} \text{Е-шаг:} & \begin{cases} p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \tilde{p}_{tdw} = p_{tdw} \left(1 + \frac{1}{n_{dw}} \left(\frac{\partial R}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw} \frac{\partial R}{\partial p_{zdw}} \right) \right) \end{cases} \quad (*) \\ \text{М-шаг:} & \begin{cases} \phi_{wt} = \text{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \text{norm}_{t \in T} \left(\sum_{w \in d} n_{dw} \tilde{p}_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases} \end{aligned}$$

Любая пост-обработка Е-шага — это регуляризатор $R(\Pi)$

Итак, произвольному гладкому регуляризатору $R(\Pi, \Phi, \Theta)$ однозначно соответствует пост-обработка $p_{tdw} \rightarrow \tilde{p}_{tdw}$.

Верно и обратное:

Теорема. Если на k -й итерации ЕМ-алгоритма для каждого (d, w) : $n_{dw} > 0$ в формулах М-шага вместо вектора $(p_{tdw}^k)_{t \in T}$ подставить вектор $(\tilde{p}_{tdw}^k)_{t \in T}$, удовлетворяющий условию нормировки $\sum_t \tilde{p}_{tdw}^k = 1$, то это эквивалентно добавлению регуляризатора сглаживания–разреживания

$$R(\Pi) = \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{t \in T} (\tilde{p}_{tdw}^k - p_{tdw}^k) \ln p_{tdw}.$$

$p(t|d, w)$ можно подвергать любой разумной пост-обработке!

Vorontsov K. V. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization. 2023.

Доказательство

В системе (*) дифф. уравнений относительно R введём переменные x_{tdw} :

$$\underbrace{p_{tdw}^k \frac{\partial R}{\partial p_{tdw}}}_{x_{tdw}} = n_{dw}(\tilde{p}_{tdw}^k - p_{tdw}^k) + p_{tdw}^k \sum_{z \in T} \underbrace{p_{zdw}^k \frac{\partial R}{\partial p_{zdw}}}_{x_{zdw}}, \quad t \in T.$$

Для любой пары (d, w) такой, что $n_{dw} > 0$, это система $|T|$ линейных уравнений относительно $|T|$ переменных x_{tdw} , $t \in T$.

Подстановкой убеждаемся, что $x_{tdw} = n_{dw}(\tilde{p}_{tdw}^k - p_{tdw}^k)$ — решение системы.

Взяв это решение, получим систему дифф. уравнений относительно R :

$$\frac{\partial R}{\partial p_{tdw}} = \frac{x_{tdw}}{p_{tdw}}, \quad d \in D, w \in d, t \in T.$$

Система декомпозируется по переменным p_{tdw} : каждой тройке (d, w, t) соответствует частное решение $R(\Pi) = x_{tdw} \ln p_{tdw} + C$. Общее решение:

$$R(\Pi) = \sum_{d \in D} \sum_{w \in d} \sum_{t \in T} x_{tdw} \ln p_{tdw} + C.$$

Подставляя сюда найденное решение x_{tdw} , получаем требуемое. ■

Контекстная тематическая модель Attentive ARTM

Дано: коллекция текстовых документов, $w_1, \dots, w_n$

$C_i = \{w_{b_i}, \dots, w_{e_i}\}$ — локальный контекст (окружение) термина w_i

α_{ui} — распределение весов термов u в контексте C_i

Найти: параметры $\phi_{wt} = p(w|t)$, $\pi_t = p(t)$ тематической модели

$$p(w|C_i) = \sum_{t \in T} \phi_{wt} \theta_{ti}, \quad \theta_{ti} = \sum_{u \in C_i} \alpha_{ui} p(t|w), \quad p(t|w) = \text{norm}_{t \in T}(\phi_{wt} \pi_t)$$

Критерий: максимум $\log$ правдоподобия с регуляризатором R :

$$\sum_{i=1}^n \ln p(w|C_i) + R(\Phi) \rightarrow \max_{\Phi, \pi}$$

$$\text{EM: } p_{ti} = \text{norm}_{t \in T}(\phi_{w_i t} \theta_{ti}); \quad \theta_{ti} = \sum_{u \in C_i} \alpha_{ui} \text{norm}_{t \in T}(\phi_{wt} \pi_t);$$

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

Как быстро вычислять веса α_{ui} в контекстах

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

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

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

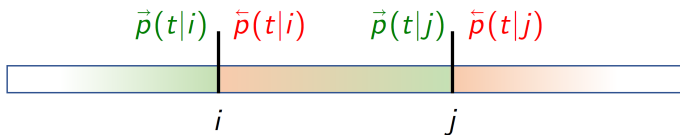
где $\vec{\gamma}_i$, $\bar{\gamma}_i$ — коэффициенты сглаживания в позиции i

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

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

- $\gamma_i \approx \frac{1}{h}$, 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 термина

Использование двунаправленных векторов контекста



Через двунаправленные тематические векторы определяется:

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

Аналогия с моделями языка GCNN, Attention, Transformer

Свёрточная нейросеть GCNN (Gated Convolutional Network)

Входные векторы слов (эмбединги)

$$X = (x_1, \dots, x_n) \in \mathbb{R}^T$$

трансформируются в векторы слов, зависящие от контекстов C_i :

$$H = (h_1, \dots, h_n) \in \mathbb{R}^d$$

через адамарово произведение:

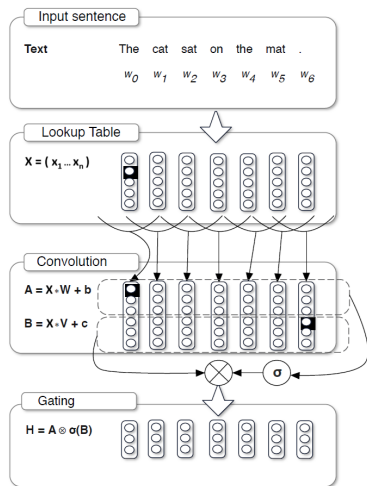
$$h_i = a_i \otimes \sigma(b_i), \text{ где}$$

$$a_i = \sum_{u \in C_i} W_u x_u \text{ — свёртка-контекст,}$$

$$b_i = \sum_{u \in C_i} V_u x_u \text{ — свёртка-фильтр,}$$

W_u, V_u — матрицы размера $d \times T$,
обучаемые параметры модели,

$$\sigma(x) = \frac{1}{1+e^{-x}} \text{ — функция сигмоида}$$



Yann N. Dauphin et al. Language modeling with gated convolutional networks, 2017.

Аналогия локализованного Е-шага с моделью GCNN

Контекстный тематический вектор на выходе Е-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \text{norm}_{t \in T}(\phi_{w_i t} \theta_{ti}) = \text{norm}_{t \in T} \left(\sum_{u \in C_i} \alpha_{ui} \phi'_{tu} \phi_{w_i t} \right)$$

Контекстный вектор на выходе модели GCNN:

$$h_i = \left(\sum_{u \in C_i} W_u x_u \right) \otimes \sigma \left(\sum_{u \in C_i} V_u x_u \right)$$

Сходство:

- вектор термина w_i трансформируется в контекстный вектор
- путём усреднения векторов ϕ'_u его контекста,
- семантически схожих с вектором термина w_i , фильтруемых адамаровым умножением на неотрицательный вектор

Отличия локализованного Е-шага:

- нет обучаемых матриц W_u, V_u как у модели GCNN
- вектор-фильтр ϕ_{w_i} без усреднения по контексту C_i
- проецирование итогового вектора на единичный симплекс

Модель внимания (self-attention) Query–Key–Value

Входные векторы слов (эмбединги)

$$X = (x_1, \dots, x_n) \in \mathbb{R}^T$$

трансформируются в векторы слов,
зависящие от контекстов C_i :

$$H = (h_1, \dots, h_n) \in \mathbb{R}^d$$

Модель внимания (self-attention):

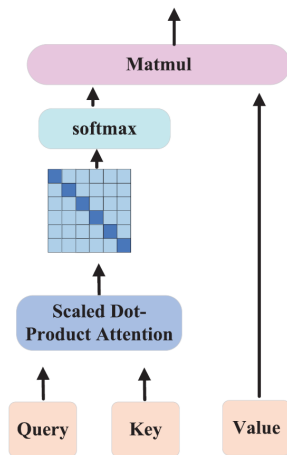
$$h_i = \sum_{u \in C_i} W_v x_u \text{ SoftMax} \langle W_k x_u, W_q x_i \rangle$$

$W_v x_u$ — вектор-значение (value)

$W_k x_u$ — вектор-ключ (key)

$W_q x_i$ — вектор-запрос (query)

W_q, W_k, W_v — обучаемые параметры



Аналогия локализованного Е-шага с моделью само-внимания

Контекстный тематический вектор на выходе Е-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \text{norm}_{t \in T}(\phi_{w_i t} \theta_{ti}) = \text{norm}_{t \in T} \left(\sum_{u \in C_i} \phi'_{tu} \alpha_{ui} \phi_{w_i t} \right)$$

Контекстный вектор на выходе модели само-внимания:

$$h_i = \sum_{u \in C_i} W_v x_u \alpha_{ui} = \sum_{u \in C_i} W_v x_u \text{SoftMax}_{u \in C_i} \langle W_k x_u, W_q x_i \rangle$$

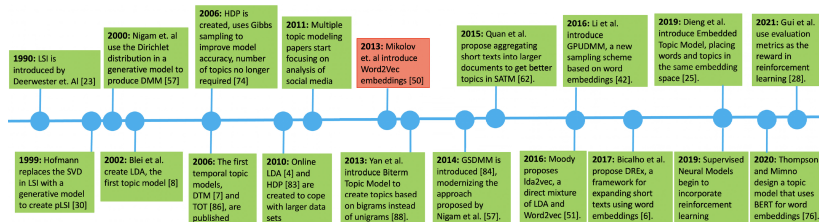
Сходство:

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

Отличия локализованного Е-шага:

- 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

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.

Нейросетевая тематическая модель Contextual-Top2Vec

Вместо РТМ — сумма технологий:

- 1 векторизация токенов (Sentence-BERT)
- 2 векторизация предложений скользящим окном в 50 токенов (mean pooling)
- 3 понижение размерности векторов (UMAP)
- 4 иерархическая кластеризация (hDbSCAN)
с автоматическим определением числа тем
- 5 иерархическое укрупнение тем слиянием мелких кластеров
с ближайшими соседями (Top2Vec)
- 6 разбиение документа на монотематические сегменты
- 7 $p(t|d)$ = доля векторов данной темы в документе
- 8 именованение тем: поиск фраз, ближайших к центроиду темы



Dimo Angelov. Top2vec: Distributed representations of topics. 2020.

D.Angelov, D.Inkpen. Topic modeling: contextual token embeddings are all you need. 2024.

Нейросетевая тематическая модель Contextual-Top2Vec

Достоинства:

- модель BERT предобучена по большим внешним данным, поэтому качество тем не зависит от размера коллекции
- документ разбивается на монотематические сегменты
- автоматически определяется число тем (hDbSCAN, Top2Vec)
- тема описывается фразами, а не отдельными словами

Недостатки:

- это работает долго, особенно на больших коллекциях
- инкрементное добавление документов не предполагается

Сходство с локализованным E-шагом:

- обработка локальных контекстов скользящим окном
- возможно разреживать $p(t|C_i)$ до монотематичности

Dimo Angelov. Top2vec: Distributed representations of topics. 2020.

D. Angelov, D. Inkpen. Topic modeling: contextual token embeddings are all you need. 2024.

«Make PTM Great Again» — почему это разумная цель

Почему рано отказываться от PTM, когда вокруг LLM

- разведочный тематический анализ и фильтрация тем по-прежнему нужны в социогуманитарных исследованиях, научном поиске, анализе трендов, социальных сетей
- PTM решают узкий класс задач лучше и быстрее, чем LLM благодаря **интерпретируемости**, **простоте** и **полноте**

Что необходимо улучшать, какие слабости устранять

- ушли от байесовского вывода к комбинированию моделей
- уходим от «мешка слов» к контекстному вниманию
- уходим от документов к локальным контекстам слов
- будем устранять недостатки интерпретируемости тем в тематически несбалансированных коллекциях
- будем генерировать заголовки и аннотации тем через LLM

Поучительные выводы

- Удалось сделать PTM «теорией одной леммы»
- Основная лемма применима не только для PTM, но и для
 - релаксационных методов дискретной оптимизации,
 - обучения монотонных нейронных сетей, и др.
- Польза обобщения моделей путём параметризации:
 - что общего у PLSA, LDA, SWB? EM-алгоритм [2012]
 - что общего у разных моделей? регуляризация [2014]
 - что общего у E-шага и внимания? локальность [2023]
- Большое научное сообщество может годами не замечать
 - более простых и рациональных методов решения
 - очевидных (задним умом) обобщений

Воронцов К.В., Потапенко А.А. Регуляризация, робастность и разреженность вероятностных тематических моделей. 2012.

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

Vorontsov K. V. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization. 2023.

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.*