Теория вероятностного тематического моделирования больших текстовых коллекций

Воронцов Константин Вячеславович
k.vorontsov@iai.msu.ru
д.ф.-м.н., профессор РАН • профессор ВМК МГУ,
руководитель лаборатории машинного обучения
и семантического анализа Института ИИ МГУ,
г.н.с. ФИЦ ИУ РАН, профессор МФТИ

Тольятти • 4–14 августа 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

- $p(w|d) = \sum_{t \in T} p(w|\cancel{d}, t)\, p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$
- $p(w|t) = \phi_{wt}$ — 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 $R_i(\Phi, \Theta)$:

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

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

Напоминание. Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

$$\prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}$$

Максимизация логарифма правдоподобия

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

эквивалентна максимизации функционала

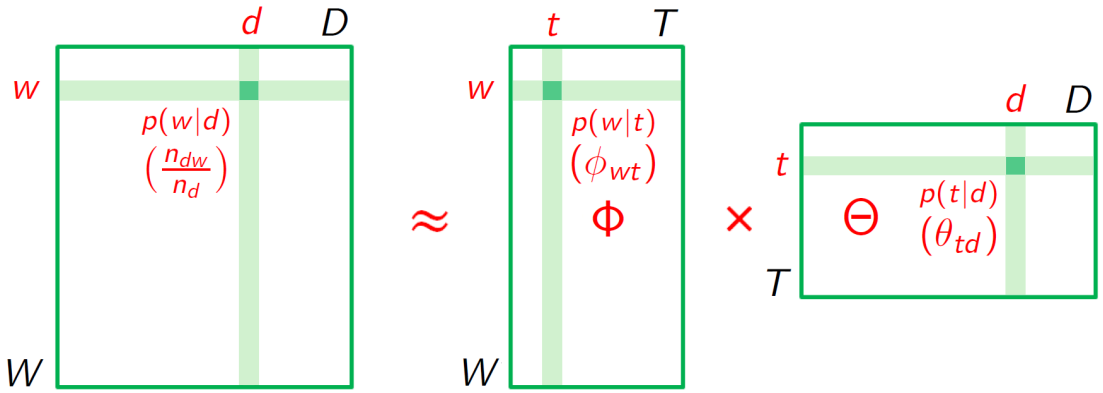
$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

при ограничениях неотрицательности и нормировки

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

Три интерпретации задачи тематического моделирования

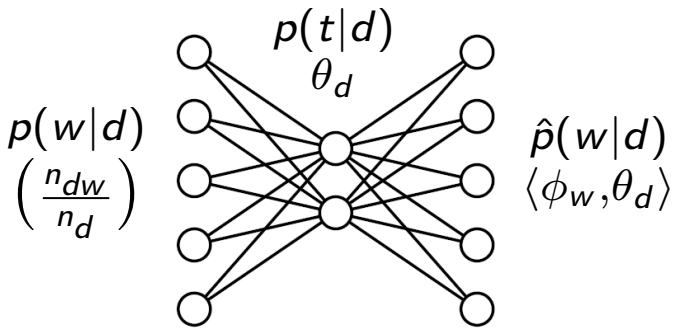
1. **Мягкая би-кластеризация** документов и слов по темам
2. **Матричное разложение** — низкоранговое, стохастическое:



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

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

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



задача реконструкции:

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

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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей, 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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример 2. Биграммная модель научных конференций

Коллекция 1000 статей конференций ММРО, ИОИ на русском

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

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

Цели и не-цели тематического моделирования

Цели:

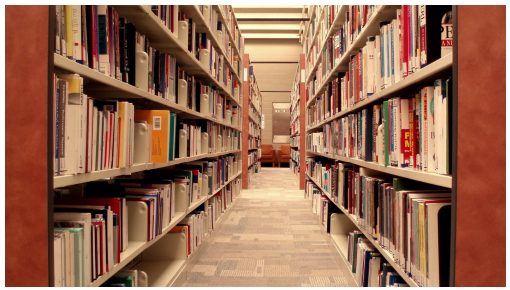
- выявлять тематическую кластерную структуру текстовой коллекции (сколько в ней тем, и о чём они), представляя результат в удобной для человека форме
- получать *интерпретируемые* тематические векторы (эмбединги) слов $p(t|w)$, слов-в-контексте $p(t|d,w)$, документов $p(t|d)$, фрагментов $p(t|s)$, объектов $p(t|x)$
- 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)

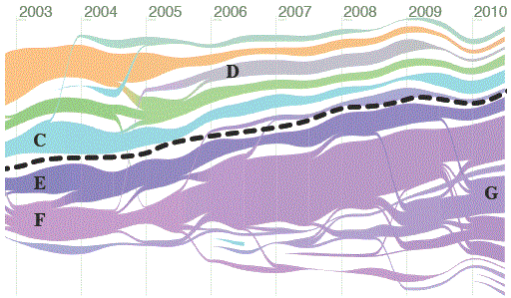
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

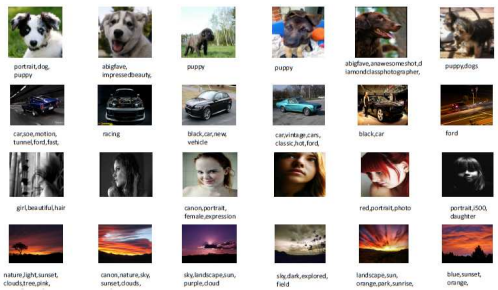


поиск тематических сообществ в соцсетях

выявление и отслеживание цепочек новостей

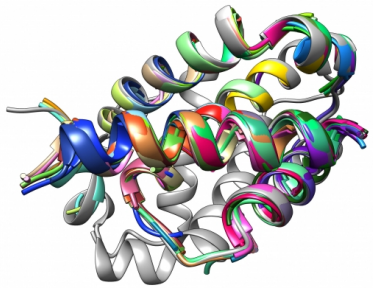


мультимодальный поиск текстов и изображений



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.

Исторические исследования: газетные архивы

[1] Корпус *Pennsylvania Gazette* 1728–1800, 25M слов:
— выделение последовательности событийных тем;
— изучение синхронности событий;
— комбинирование автоматического анализа и ручного.

[2] *Газеты Техаса* от гражданской войны до наших дней:
— выделение всех тем, связанных с хлопком;
— построение серии моделей в скользящих окнах;
— важность качественной предобработки текстов.

[3] Газеты и периодика Финляндии (1854–1917):
— выделение тем о церкви, религии, образовании;
— тренды модернизации и секуляризации финского общества.

---

1. *D.Newman, S.Block.* Probabilistic topic decomposition of an eighteenth-century American newspaper. 2006.

2. *Tze-I Yang, A.J.Torget, R.Mihalcea.* Topic modeling on historical newspapers. 2011.

3. *J.Marjanen et al.* Topic modelling discourse dynamics in historical newspapers. 2021.

Исторические исследования: летописи и дневники

[1] Двуязычный корпус книг на английском и немецком:
— все темы, связанные с эпистемологией

[2] Корпус текстов на китайском языке (1644–1912):
— все темы, связанные с бандитизмом, преступлениями;
— необходим контекст для установления типа преступления;
— важность правильной токенизации для китайского языка.

[3] Дневник Martha Ballard (1735–1812), охватывает 27 лет:
— выделение событийных и перманентных тем;
— выделение персональных и исторических тем;
— специфичный английский XVIII века.

---

1. *M.Erlin.* Topic modeling, epistemology, and the English and German novel. 2017.
2. *Ian Matthew Miller.* Rebellion, crime and violence in Qing China, 2013.
3. *Cameron Blevins.*
http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary.

Исторические исследования: научная и литературная периодика

Статьи коллекции JSTOR доступны в виде «мешков слов».

[1] Научные журналы XX века:
— различия тематики на английском и немецком языках;
— особенно исследовались различия, связанные со 2MB;
— для объединения тем использовались интервики Википедии.

[2] Более 100 лет литературно-художественной периодики:
— как менялись темы;
— как менялись значения слов внутри каждой темы;
— как менялась тема насилия (violence, power, fear, blood, death, murder, act, guilt).

---

1. *D.Mimno.* Computational historiography: Data mining in a century of classics journals. 2012.

2. *A.Goldstone, T.Underwood.* The quiet transformations of literary studies: What thirteen thousand scholars could tell us. 2014.

ТМ в политологии: анализ публичных выступлений

[1] Выступления (210K) в Европарламенте, 1999–2014:
— выявление событийных тем и эволюции перманентных тем;
— как члены и комитеты ЕП влияют на формирование тем

[2] Модель контрастных мнений (Contrastive Opinion Modeling)
— выступления в Сенате США (www.votesmart.org);
— СМИ: New York Times, Xinhua News, The Hindu, 2009–2010

[3] Выступления в Совбезе ООН по Афганистану, 2001–2017:
— динамика отношения разных стран к проблемам Афганистана

---

[1] *D.Greene, J.P.Cross.* Unveiling the political agenda of the European Parliament plenary: a topical analysis. 2015.

[2] *Fang, Y., et al.* Mining contrastive opinions on political texts using cross-perspective topic model. 2012.

[3] *M.Schönfeld.* Discursive landscapes and unsupervised topic modeling in IR: a validation of text-as-data approaches through a new corpus of UN Security Council speeches on Afghanistan. 2018.

ТМ в политологии: анализ СМИ и социальных медиа

[1] Тематика изменения климата в СМИ Пакистана, 2010–2021
— выявление, группирование и динамика тем

[2] Выявление поляризации новостей (AYLIEN COVID-19)
— 1,5M новостей, 440 источников СМИ, 11.2019–07.2020

[3] Выявление политических взглядов пользователей Twitter

[4] Что пишет NYT о ядерных технологиях с 1945 по н/в

---

[1] *W.Ejaz et al.* Politics triumphs: A topic modeling approach for analyzing news media coverage of climate change in Pakistan. 2023

[2] *Zihao He.* Detecting polarized topics using partisanship-aware contextualized topic embeddings. 2021

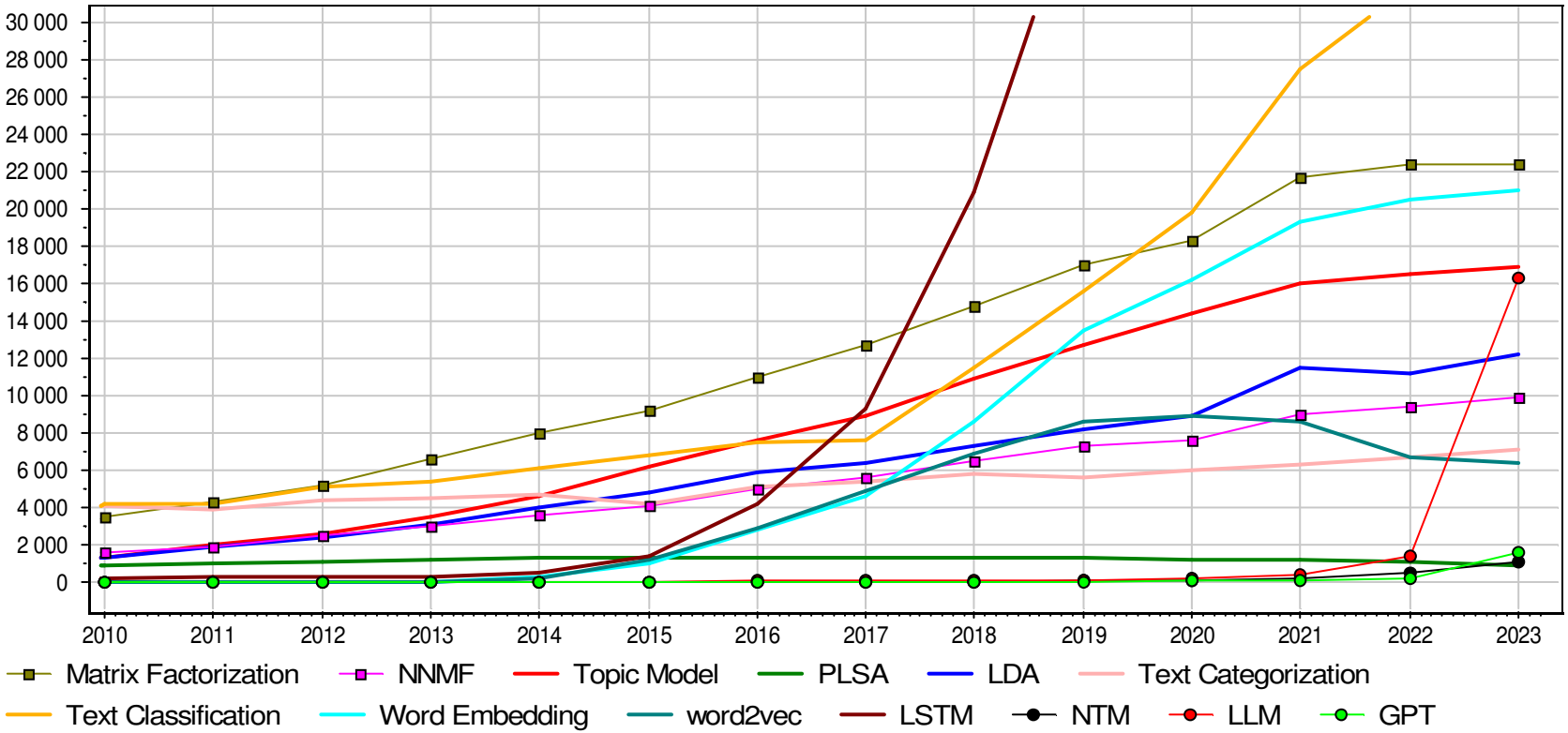
[3] *R.Cohen, D.Ruths.* Classifying Political Orientation on Twitter: It's Not Easy! 2013.

[4] *C.Jacobi.* Quantitative analysis of large amounts of journalistic texts using topic modelling. 2015.

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

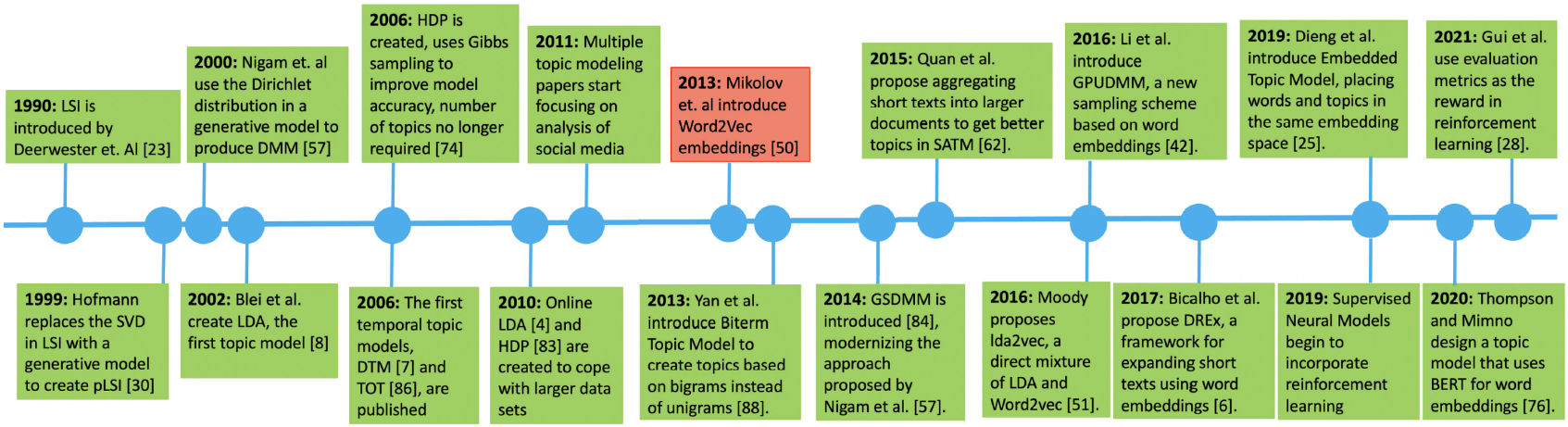
Научные тренды: PTM, LLM и смежные с ними

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
- обобщаются на мультимодальные, мультиязычные 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 $\rightarrow$ алгоритм ближе к SGD
- транзакционные данные $\rightarrow$ ближе к foundation models
- отказ от мешка слов $\rightarrow$ ближе к модели внимания

«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
- PTM решают узкий класс задач лучше и быстрее, чем LLM благодаря интерпретируемости, простоте и полноте

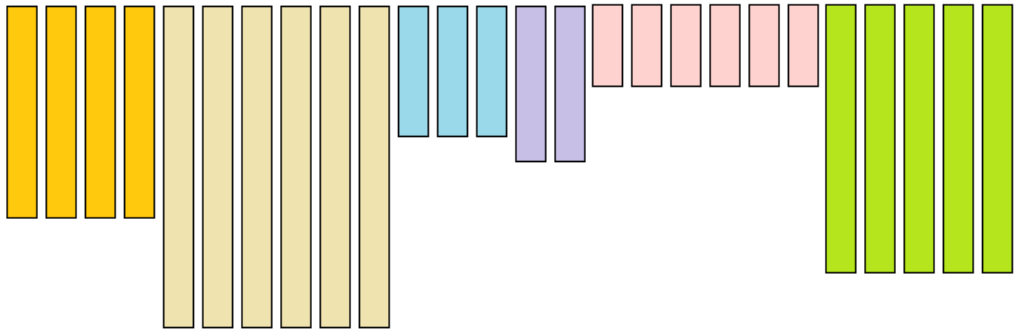
Что необходимо улучшать, какие слабости устранять

- ушли от байесовского вывода к комбинированию моделей
- уходим от «мешка слов» к контекстному вниманию
- уходим от документов к локальным контекстам слов
- устранять причины плохой (иногда) интерпретируемости (*гипотеза:* дублирующие и мусорные темы возникают из-за тематической несбалансированности коллекции)
- генерировать заголовки и аннотации тем с помощью LLM

Задача максимизации функции на единичных симплексах

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

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



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

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

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

Необходимые условия экстремума и метод простых итераций

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

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

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

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

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

но учитывают ограничения и не требуют подбора шага η

Напоминание. Условия Каруша–Куна–Таккера

Задача математического программирования:

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

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

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

Доказательство леммы о максимизации на симплексах

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

Функция Лагранжа:

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

Условия Каруша–Куна–Таккера для вектора ω_j:

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

Умножим обе части первого равенства на ω_{ij}:

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

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

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

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

Теорема о сходимости итерационного процесса

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

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

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

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

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

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

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

А.Н.Тихонов
(1906–1993)

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

Доказательство (по лемме о максимизации на симплексах)

Применим лемму к log-правдоподобию с регуляризатором:

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

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

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

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

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

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

PLSA и LDA — две самые известные тематические модели

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

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

M-шаг — частотные оценки условных вероятностей:

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

LDA: latent Dirichlet allocation (латентное размещение Дирихле):

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

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

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

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

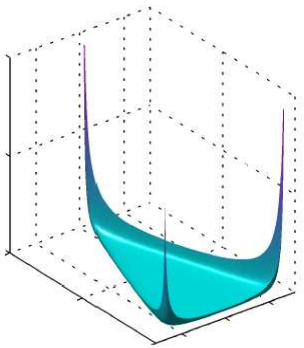
Байесовская модель LDA: априорные распределения Дирихле

Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})$ и $\theta_d = (\theta_{td})$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

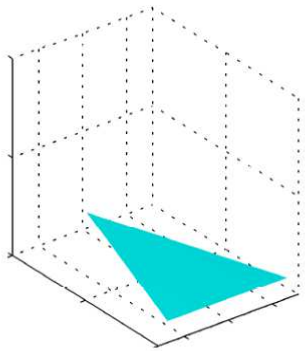
$$\text{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \ \beta_w > 0;$$

$$\text{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t > 0;$$

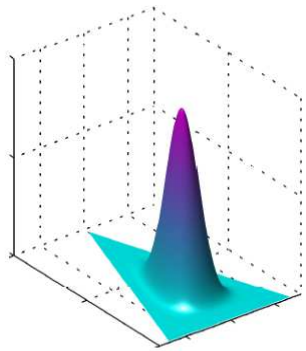
Пример. Распределение $\text{Dir}(\theta|\alpha)$ при $|T| = 3$, $\theta, \alpha \in \mathbb{R}^3$



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$

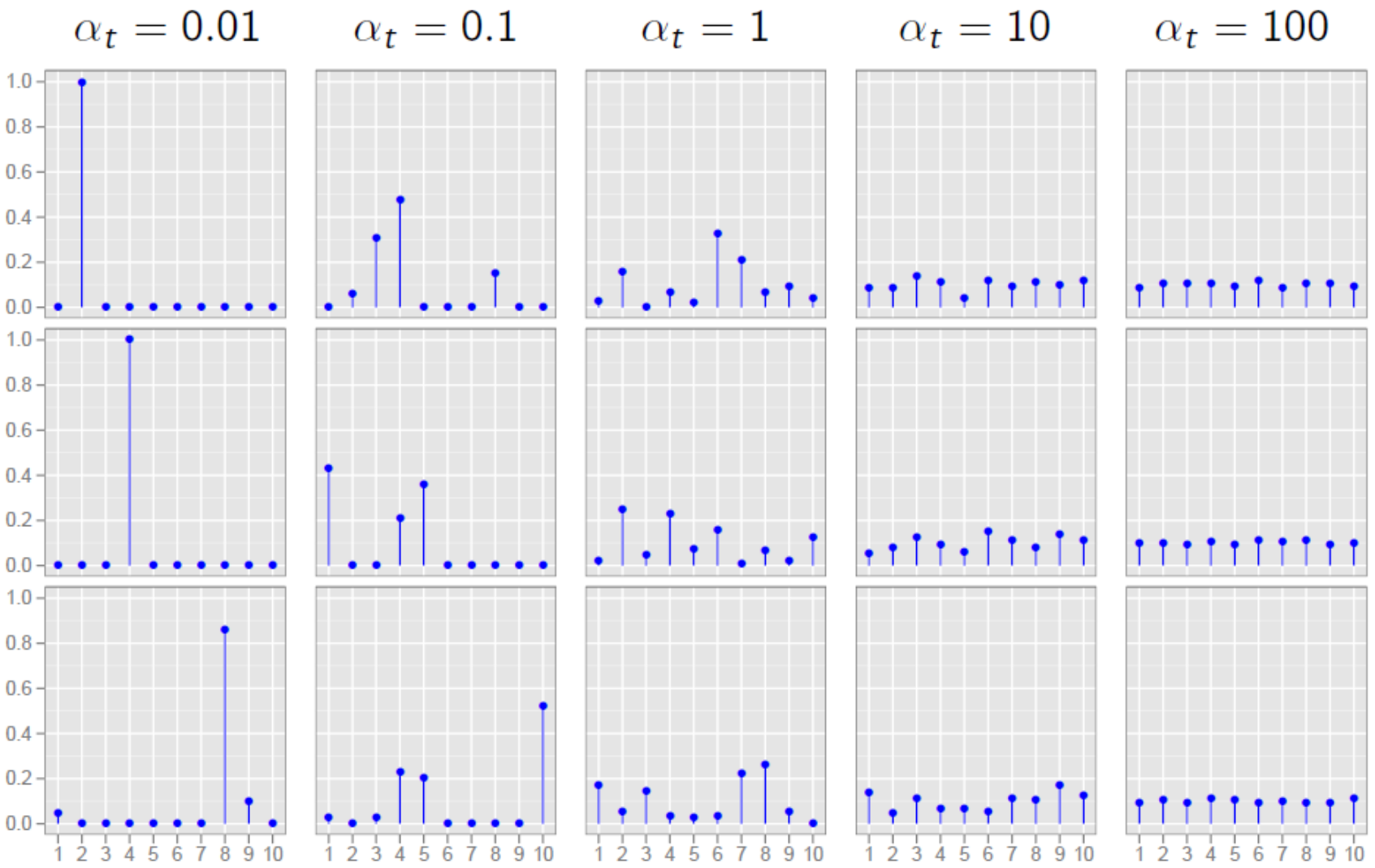


$\alpha_1 = \alpha_2 = \alpha_3 = 1$



$\alpha_1 = \alpha_2 = \alpha_3 = 10$

Пример. Выборки из трёх 10-мерных векторов $\theta \sim \mathrm{Dir}(\theta|\alpha)$



LDA: максимизация апостериорной вероятности

Совместное правдоподобие данных и модели:

$$\ln \prod_{d\in D}\prod_{w\in d} p(w,d|\Phi,\Theta)^{n_{dw}} \prod_{t\in T}\mathrm{Dir}(\phi_t|\beta) \prod_{d\in D}\mathrm{Dir}(\theta_d|\alpha) \to \max_{\Phi,\Theta}$$

Регуляризатор — логарифм априорного распределения:

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

M-шаг — сглаженные или разреженные частотные оценки:

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

при $\beta_w > 1$, $\alpha_t > 1$ — сглаживание,
при $0 < \beta_w < 1$, $0 < \alpha_t < 1$ — слабое разреживание,
при $\beta_w = 1$, $\alpha_t = 1$ априорное распределение равномерно, PLSA.

От байесовского вывода к аддитивной регуляризации

X — исходные данные, Ω — параметры порождающей модели

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

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

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

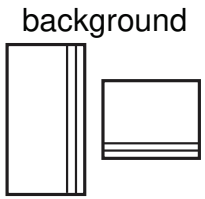
Максимизация апостериорной вероятности (MAP) даёт точечную оценку Ω напрямую, без вывода Posterior:

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

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

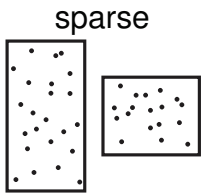
$$\Omega := \arg\max_{\Omega} \Big(\ln p(X|\Omega) + \sum_{i=1} \tau_i R_i(\Omega)\Big)$$

Регуляризаторы для улучшения интерпретируемости тем



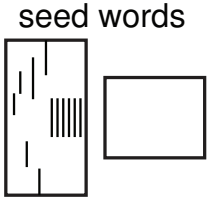
Сглаживание фоновых тем $B \subset T$:

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

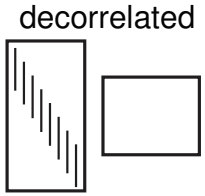


Разреживание предметных тем $S = T \setminus B$:

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

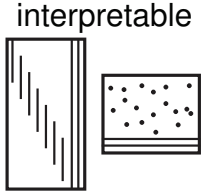


Сглаживание для выделения релевантных тем
с помощью словаря «затравочных» ключевых слов



Декоррелирование для повышения различности тем:

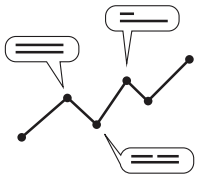
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

Регуляризаторы для учёта дополнительной информации

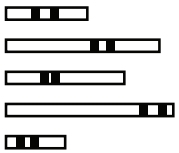
regression



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

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

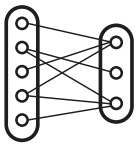
biterm



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

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

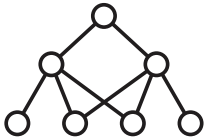
relational



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:

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

Мультимодальная тематическая модель ARTM

W_m — словарь термов m-й модальности, $m \in M$

Максимизация суммы log-правдоподобий с регуляризацией:

$$\sum_{m\in M} \tau_m \sum_{d\in D} \sum_{w\in W_m} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

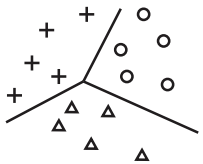
$$\text{E-шаг:}\quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

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

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

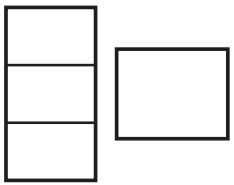
Регуляризаторы для мультимодальных тематических моделей

supervised



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

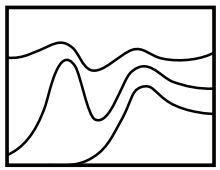
multilanguage



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:

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

geospatial

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

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

Трёхматричная тематическая модель ARTM

Темы порождаются модальностью C (категории, авторы):

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T}\sum_{c\in C} \phi_{wt}\psi_{tc}\theta_{cd} + R(\Phi,\Psi,\Theta) \to \max_{\Phi,\Psi,\Theta};$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:}\quad p_{tcdw} \equiv p(t,c|d,w) = \underset{(t,c)\in T\times C}{\text{norm}}\big(\phi_{wt}\psi_{tc}\theta_{cd}\big);$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{wt} = \underset{w\in W_m}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right); & n_{wt} = \sum\limits_{d,c} n_{dw}p_{tcdw} \\ \psi_{tc} = \underset{t\in T}{\text{norm}}\left(n_{tc} + \psi_{tc}\frac{\partial R}{\partial \psi_{tc}}\right); & n_{tc} = \sum\limits_{d,w} n_{dw}p_{tcdw} \\ \theta_{cd} = \underset{c\in C}{\text{norm}}\left(n_{cd} + \theta_{cd}\frac{\partial R}{\partial \theta_{cd}}\right); & n_{cd} = \sum\limits_{w,t} n_{dw}p_{tcdw} \end{cases}$$

M. Rosen-Zvi et al. The author-topic model for authors and documents. 2004.

Транзакционные данные

Текст — это двудольный граф с рёбрами вида (d, w), $w \in W_m$.
Когда данные содержат 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 тематическая модель транзакционных данных

Дано: E_k — выборка транзакций (рёбер гиперграфа) типа k
n_{kdx} — число вхождений ребра (d,x) в выборку E_k

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

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdx} = p(t|d,x)$:

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

M-шаг:
$$\begin{cases} \phi_{vt} = \operatorname*{norm}_{v\in V^m}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k} \big[v\in x\big] n_{kdx} p_{tdx} + \phi_{vt}\dfrac{\partial R}{\partial \phi_{vt}}\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k} n_{kdx} p_{tdx} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

Vorontsov K.V. Rethinking probabilistic topic modeling from the point of view. . . 2023.

Частный случай: тематическая модель предложений

S_d — множество предложений документа d

n_{sw} — сколько раз терм w встречается в предложении s

Критерий: максимум регуляризованного log-правдоподобия

$$\sum_{d\in D}\sum_{s\in S_d}\ln\sum_{t\in T}\theta_{td}\prod_{w\in s}\phi_{wt}^{n_{sw}} + R(\Phi,\Theta)\to\max_{\Phi,\Theta}$$

Свобода выбора гипер-рёбер *сегментоидов* — подмножеств термов, связанных по смыслу и порождаемых общей темой:

- предложение / фраза / синтагма / именная группа
- факт «объект, субъект, действие»
- лексическая цепочка: синонимы, гипонимы, меронимы
- текст комментария, дата–время, автор

Wayne Xin Zhao et al. Comparing Twitter and traditional media using topic models. 2011.
G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Почему отказ от байесовского обучения даёт больше свободы

В байесовском подходе регуляризаторы являются априорным распределением — частью генеративной модели.

В подходе ARTM это инструмент управления сходимостью итерационного процесса в некорректно поставленной задаче.

ARTM позволяет свободно:

- подбирать комбинацию регуляризаторов под задачу
- менять коэффициенты регуляризации в ходе итераций
- включать и отключать регуляризаторы в ходе итераций
- ставить задачу управления траекторией регуляризации
- отключать все регуляризаторы в конце итераций, чтобы получать несмещённые оценки параметров

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

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

K.Vorontsov, O.Frei, 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

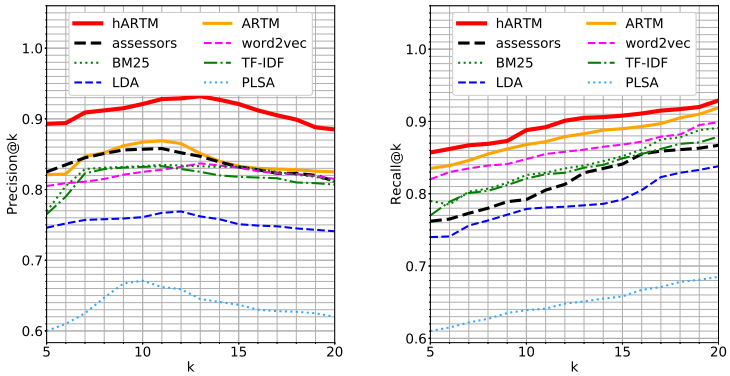
3.7M статей Википедии, 100K слов: время min (перплексия)

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.

Разведочный поиск в технологических блогах

Цель: поиск документов
по длинным текстовым запросам
— Habr.ru (175K документов),
— TechCrunch.com (760K док.).



Регуляризаторы:

$$\mathscr{L}\left(\Phi,\Theta\right) + R\left(\text{hierarchy}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) \to \max$$

Результаты:

- Точность и полнота 93%, превосходит асессоров и другие методы (tf-idf, BM25, word2vec, PLSA, LDA, ARTM).
- Увеличилась оптимальная размерность векторов:
  $200 \to 1400$ (Habr.ru), $475 \to 2800$ (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: мультиязычный поиск и классификация научных публикаций по рубрикаторам 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:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{multilanguage}\right) + R\left(\text{supervised}\right) \to \max$$

Результаты:

- точность мультиязычного поиска 94%
- сокращение модели 128 Гб $\to$ 4.8 Гб при редукции словарей (BPE-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).

(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, сообщать, океан, станция, хатико, 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, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,
(китайцы): китайский, россия, производство, китай, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,
(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, румын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвале,
(осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алахай, российский, население, конфликт,
(цыгане): наркотик, цыган, цыганка, хороший, место, страна, деньга, время, работать, жизнь, жить, рука, дом, цыганский, наркоманка,

Регуляризаторы:

$$\mathscr{L}\left(\Phi,\Theta\right) + R\left(\text{seed words}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{temporal}\right) + R\left(\text{geospatial}\right) + R\left(\text{sentiment}\right) \to \max$$

Результаты: число релевантных тем: 45 (LDA) $\to$ 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.

Аналогичные исследования по выделению узкой тематики

Задачи «поиска и классификации иголок в стоге сена»

- поиск и кластеризация новостей [1]
- поиск в социальных медиа информации, связанной с болезнями, симптомами и методами лечения [2]
- поиск и тематическая классификация чатов, связанных с преступностью и экстремизмом [3, 4]
- поиск выступлений о правах человека в ООН [5]

---

1. *J.Jagarlamudi, H.Daumé III, R.Udupa.* Incorporating lexical priors into topic models. 2012.

2. *M.Paul, M.Dredze.* Discovering health topics in social media using topic models. 2014.

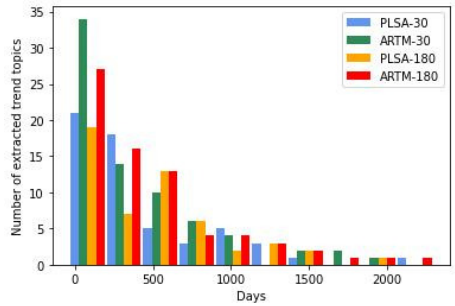
3. *M.A.Basher, A.Rahman, B.C.M.Fung.* Analyzing topics and authors in chat logs for crime investigation. 2014.

4. *A.Sharma, M.Pawar.* Survey paper on topic modeling techniques to gain useful forecasting information on violant extremist activities over cyber space. 2015.

5. *Kohei Watanabe, Yuan Zhou.* Theory-driven analysis of large corpora: semisupervised topic classification of the UN speeches. 2022.

Выявление трендов в коллекции научных публикаций

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



Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{interpretable}\right) + R\left(\text{dynamic}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) \to \max$$

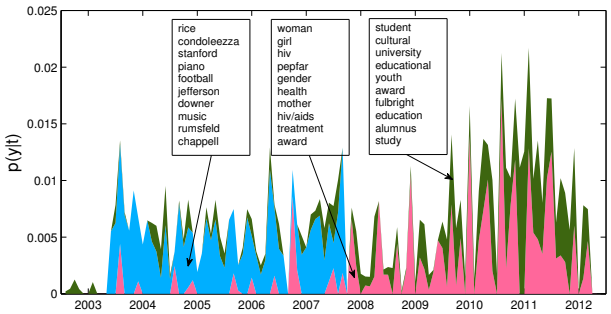
Результаты:

- 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:

$$\mathscr{L}\left(\Phi,\Theta\right) + R\left(\text{interpretable}\right) + R\left(\text{temporal}\right) + R\left(\text{multimodal}\right)$$
$$+ R\left(\text{n-gram}\right) + R\left(\text{multilanguage}\right) \to \max$$

Результаты:

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

Modalities	*Pr*	*Rec*	*F1*
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**

Регуляризаторы:

$$\mathscr{L}\left(\underset{\Phi\ \Theta}{\text{PLSA}}\right)+R\left(\underset{}{\text{interpretable}}\right)+R\left(\underset{}{\text{multimodal}}\right)+R\left(\underset{}{\text{n-gram}}\right)+R\left(\underset{}{\text{syntax}}\right)\to\max$$

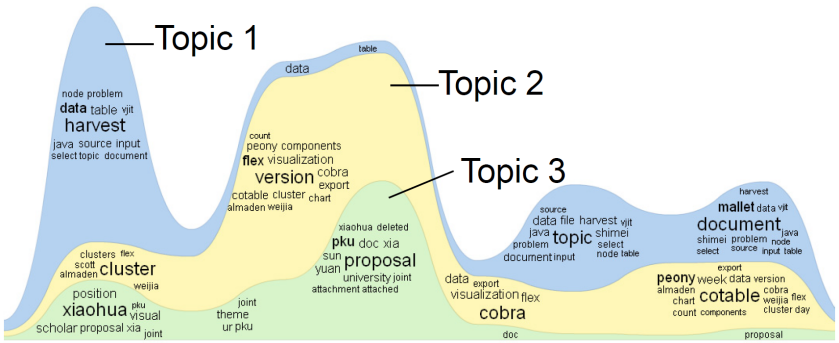
Результаты:

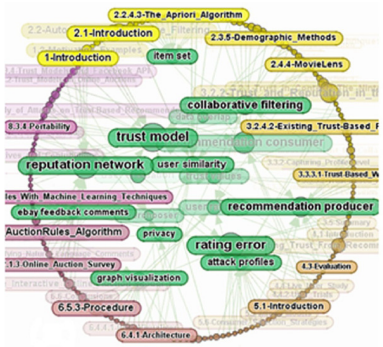
- выделение мнений внутри тем: F1-мера = 0.86%
- совместное использование трёх модальностей:
  - SPO — факты как триплеты «субъект–предикат–объект»
  - FR — семантические роли слов по Филлмору
  - Sent — тональности именованных сущностей

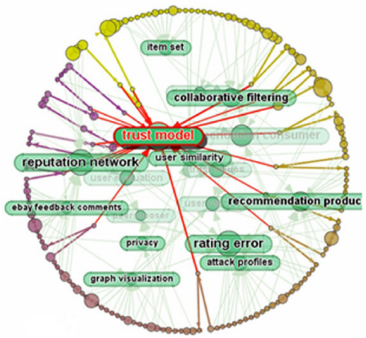
D.Feldman, T.Sadekova, K.Vorontsov. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. Dialogue 2020.

Мотивации. Что хотим от PTMs глядя на LLM (BERT, GPT)

- вместо «мешка слов» — последовательность $w_1, \ldots, w_n$
- 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







Сегментная структура текста и пост-обработка E-шага

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

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

Регуляризация E-шага

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

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

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

EM-алгоритм: метод простой итерации для системы уравнений

E-шаг:
$$\begin{cases} p_{tdw} = \underset{t\in T}{\operatorname{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}\right)\right) \qquad (*) \end{cases}$$

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

Набросок доказательства: три шага

1. Для функции $p_{tdw}(\Phi,\Theta) = \frac{\phi_{wt}\theta_{td}}{\sum_z \phi_{wz}\theta_{zd}}$ и любого $z \in T$

$$\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}} = \theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

2. Введём вспомогательную функцию от переменных Π, Φ, Θ:

$$Q_{tdw}(\Pi,\Phi,\Theta) = \frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{zdw}}.$$

Если $R(\Pi,\Phi,\Theta)$ не зависит от p_{tdw} при $w \notin d$, то

$$\phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}} = \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \sum_{d\in D} p_{tdw}Q_{tdw}; \quad \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_{w\in d} p_{tdw}Q_{tdw}.$$

3. Подставляем это в формулы M-шага:

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

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

■

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

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

Верно и обратное:

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

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

$p(t|d, w)$ 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. ■

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 $\theta_{td}^0 = \frac{1}{|T|}$:

$$\theta_{td}(\Phi) = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in W} n_{dw} p_{tdw}\Big) = \sum_{w\in d}\frac{n_{dw}}{n_d}\operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}^0\big)$$

- формула полной вероятности + гипотеза усл. независ.:

$$\theta_{td}(\Phi) = \sum_{w\in d} p(w|d)\, p(t|w,\cancel{d}) = \sum_{w\in d}\frac{n_{dw}}{n_d}\operatorname*{norm}_{t\in T}\big(\phi_{wt} p_t\big)$$

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

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

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

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

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 M-шага относительно Φ и $\Theta(\Phi)$:

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

Применим Лемму к регуляризованному log-правдоподобию Q:

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

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

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

■

EM-алгоритм для ARTM с линейной тематизацией документов

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

EM-алгоритм: метод простой итерации для системы уравнений

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

$$p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right); \qquad n_t = \sum_{d \in D} \sum_{w \in d} n_{dw} p_{tdw}$$

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

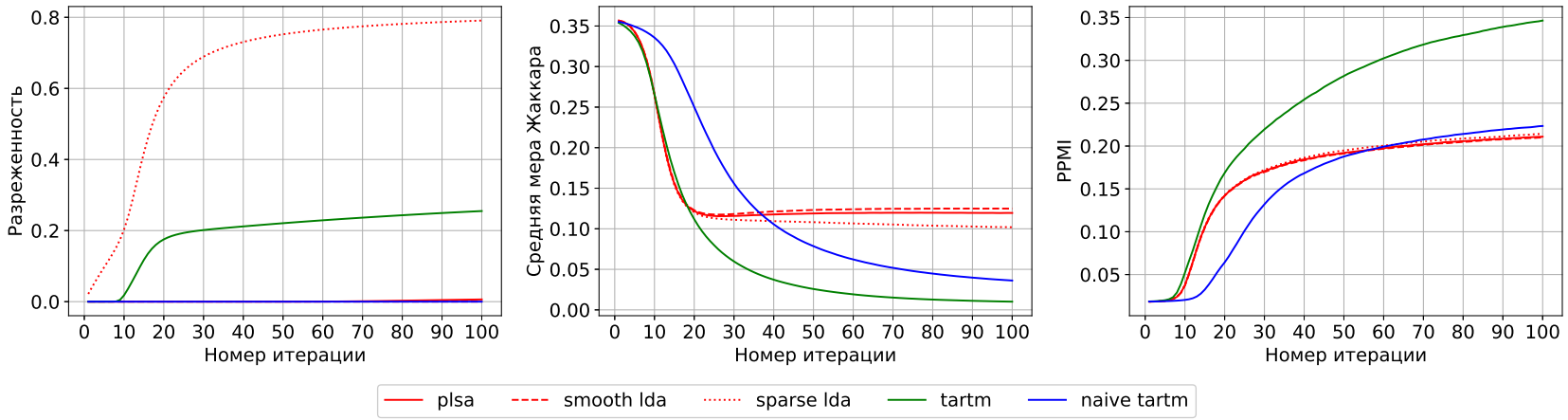
$$p'_{tdw} = p_{tdw} + \frac{\phi'_{tw}}{n_d} \left(\frac{n_{td}}{\theta_{td}} - \sum_{s \in T} \phi'_{sw} \frac{n_{sd}}{\theta_{sd}} \right)$$

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

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

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

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

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

Это обычный EM-алгоритм, только с однопроходным E-шагом!

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

Линейная тематизация: от документа к локальным контекстам

Тематизация документа $d = (w_1, \ldots, w_{n_d})$ 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* $C_i = (\ldots, w_i, \ldots)$ 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 α_{ui}:

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

Локализованная тематическая модель:

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

EM-алгоритм с локализованным E-шагом

$w_1, \ldots, w_n$ — сквозная нумерация термов во всей коллекции

C_i — локальный контекст (окружение) терма w_i

α_{ui} — распределение весов термов $u \in C_i$ около терма w_i

- отказались от гипотезы «мешка слов»
- отказались от разбиения коллекции на документы

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} \equiv p(t|w) = \underset{t\in T}{\operatorname{norm}}\big(\phi_{wt} p_t\big); \qquad \theta_{ti} \equiv p(t|C_i) = \sum_{u\in C_i} \alpha_{ui}\phi'_{tu};$$

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

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

Быстрое вычисление двунаправленных векторов контекста

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

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

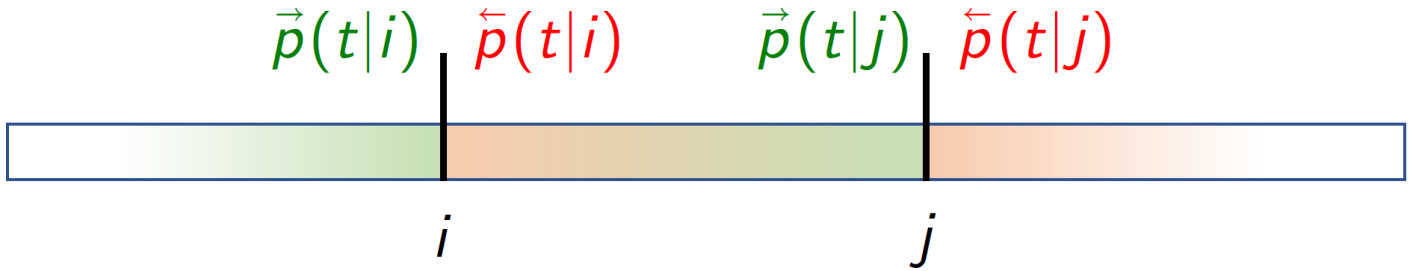
где $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — коэффициенты сглаживания в позиции i

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

Аналогия с моделями языка GCNN, Attention, Transformer

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};

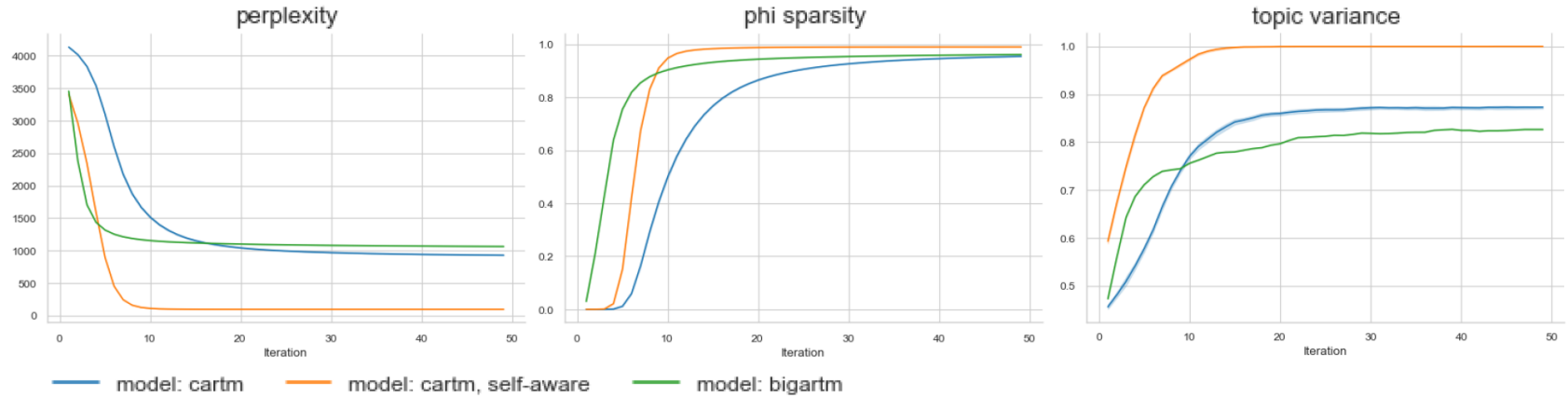
инициализация: $n_{wt} := 0$; $\tilde{n}_{wt} := 0$; $n_t := 1$; $\phi_{wt} := \text{random}$;
для всех документов $d \in D$

$p_{ti} := \text{norm}_t(\phi_{w_i t} n_t)$, $i = 1, \ldots, n_d$, $t \in T$;
$\vec{\theta}_{ti} := \vec{\gamma}_i\, p_{ti} + (1 - \vec{\gamma}_i)\, \vec{\theta}_{t,i-1}$, $i = 1, \ldots, n_d$, $\vec{\gamma}_1 = 1$, $t \in T$;
$\overleftarrow{\theta}_{ti} := \overleftarrow{\gamma}_i\, p_{ti} + (1 - \overleftarrow{\gamma}_i)\, \overleftarrow{\theta}_{t,i+1}$, $i = n_d, \ldots, 1$, $\overleftarrow{\gamma}_{n_d} = 1$, $t \in T$;
$p_{ti} := \text{norm}_t\left(\phi_{w_i t}\left(\beta \vec{\theta}_{ti} + (1-\beta)\overleftarrow{\theta}_{ti}\right)\right)$, $i = 1, \ldots, n_d$, $t \in T$;
$\tilde{n}_{w_i t} := \tilde{n}_{w_i t} + p_{ti}$; $n_t := n_t + p_{ti}$, $i = 1, \ldots, n_d$, $t \in T$;
если пора обновить матрицу Φ **то**

$n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}$; $\tilde{n}_{wt} := 0$;
$\phi_{wt} := \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\right)$;

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++

Дьяков И.А. Тематические модели внимания для анализа связного текста. ВКР бакалавра, ВМК МГУ, 2025.

Свёрточная нейросеть GCNN (Gated Convolutional Network)

Входные векторы слов (эмбединги)

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

трансформируются в векторы слов, зависящие от контекстов C_i:

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

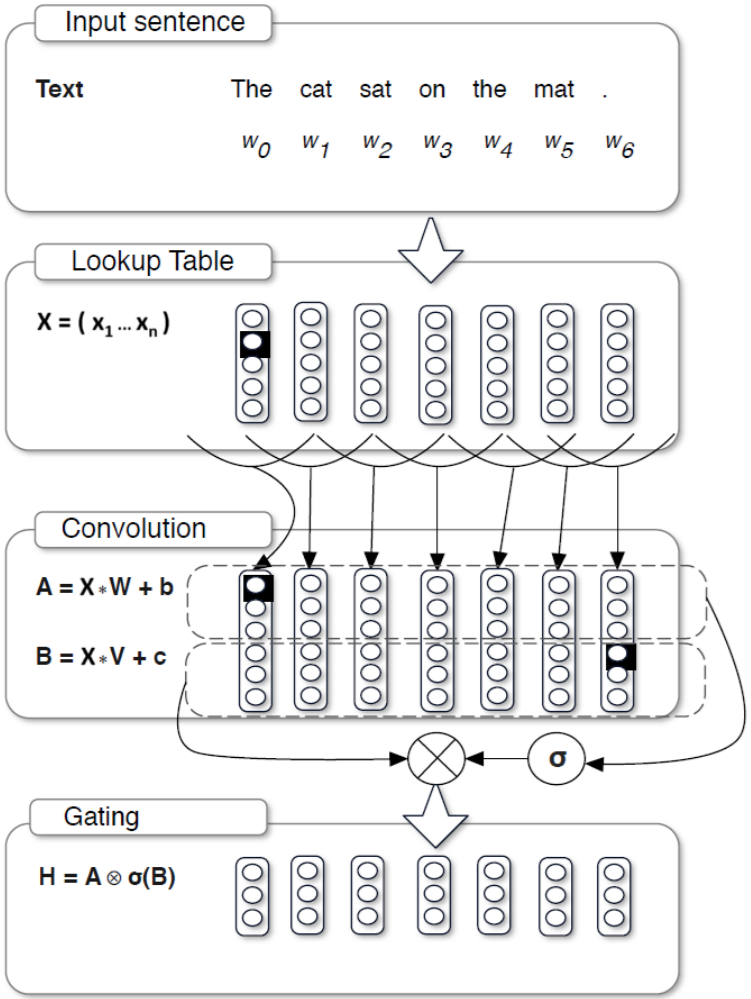
через адамарово произведение:

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

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

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

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



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

Аналогия локализованного E-шага с моделью GCNN

Контекстный тематический вектор на выходе E-шага:

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

Контекстный вектор на выходе модели GCNN:

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

Сходство:

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

Отличия локализованного E-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)

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

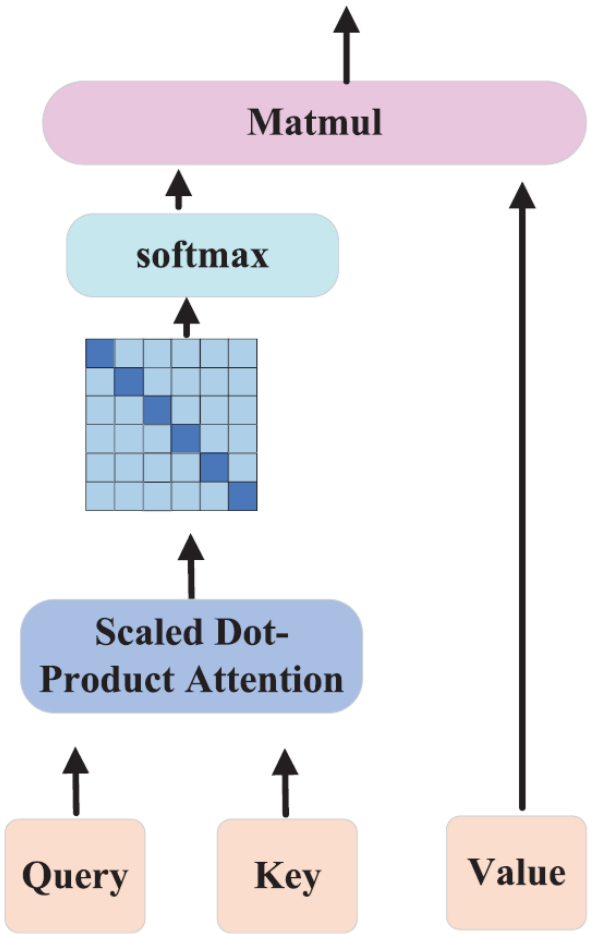
трансформируются в векторы слов, зависящие от контекстов C_i:

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

Модель внимания (self-attention):

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

$W_v x_u$ — вектор-значение (value)
$W_k x_u$ — вектор-ключ (key)
$W_q x_i$ — вектор-запрос (query)
W_q, W_k, W_v — матрицы параметров



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

Аналогия локализованного E-шага с моделью само-внимания

Контекстный тематический вектор на выходе E-шага:

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

Контекстный вектор на выходе модели само-внимания:

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

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 локализованного E-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

Аналогия локализованного E-шага с моделью трансформера

Один проход документа аналогичен модели внимания:

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

$\phi'_{tw_i} = \text{norm}_t(\phi_{w_i t} p_t)$ — 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

$p_{ti} = \text{norm}_t\left(\phi_{w_i t}\theta_{ti}\right)$ — 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 итераций аналогичны L необучаемым блокам внимания

Онлайновый EM с многопроходным локализованным E-шагом

Вход: коллекция, число тем $|T|$, параметры L, 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 $d \in D$

$p_{ti} := \text{norm}_t(\phi_{wt} n_t)$;
для всех $l = 1, \ldots, L$ (аналог L блоков внимания)

$\vec{\theta}_{ti} := \vec{\gamma}_i\, p_{ti} + (1 - \vec{\gamma}_i)\, \vec{\theta}_{t,i-1}, \quad i = 1, \ldots, n_d, \quad \vec{\gamma}_1 = 1;$

$\overleftarrow{\theta}_{ti} := \overleftarrow{\gamma}_i\, p_{ti} + (1 - \overleftarrow{\gamma}_i)\, \overleftarrow{\theta}_{t,i+1}, \quad i = n_d, \ldots, 1, \quad \overleftarrow{\gamma}_{n_d} = 1;$

$p_{ti} := \text{norm}_t\big(\big(\beta \vec{\theta}_{ti} + (1-\beta)\overleftarrow{\theta}_{ti}\big) p_{ti}/n_t\big)$;

$\tilde{n}_{w_i t} := \tilde{n}_{w_i t} + p_{ti}; \quad n_t := n_t + p_{ti};$
если пора обновить матрицу Φ **то**

$n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}; \quad \tilde{n}_{wt} := 0;$

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

Нейросетевая тематическая модель Contextual-Top2Vec

Вместо PTM — 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. иерархичекое 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. именование 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 локализованного E-шага:

- 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.

Поучительные выводы

- Лемма о максимизации на единичных симплексах применима не только в тематическом моделировании
- Польза обобщения моделей путём параметризации:
  1) «что общего между PLSA, LDA, SWB?» (2012)
  2) ансамблирование регуляризаторов, ARTM (2014)
  3) модальности (2015), иерархии (2016), транзакции (2018)
  4) модели внимания и дальнейшее сближение с LLM (2020)
- 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.
Воронцов К.В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. URSS, 2025.
http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf