

Вероятностные тематические модели локального контекста

Воронцов Константин Вячеславович^{1,2,3,4},

Дьяков Илья Андреевич^{1,2},

Карпов Матвей Леонидович³

(¹ВМК МГУ, ²Институт ИИ МГУ, ³МФТИ, ⁴ФИЦ ИУ РАН)



Интеллектуализация Обработки Информации
Беларусь, Витебск • 28 сентября – 1 октября 2026

1 Эволюция тематического моделирования

- постановка задачи и примеры приложений
- три основных подхода BTM
- мотивация исследования: Make PTMs Great Again

2 Тематическая модель локальных контекстов

- быстрая однократная тематизация текста
- тематическая модель локального контекста
- сравнение с нейросетевыми контекстными моделями

3 Коэффициенты внимания и эксперименты

- определение коэффициентов внимания
- оптимизация коэффициентов внимания
- эксперименты

Задача вероятностного тематического моделирования (BTM)

— это задача «мягкой» кластеризации текстовых документов

Дано: словарь W , коллекция документов D , число тем $|T|$

Найти: оценки дискретных распределений

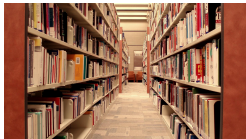
- $p(t|d)$ — вероятность темы $t \in T$ в документе $d \in D$
- $p(w|t)$ — вероятность слова (терма) $w \in W$ в теме t
- $p(t)$ — вероятность темы t во всей коллекции

Три основных подхода:

- *оптимизационный* (PLSA [1999], ARTM [2014]):
обучает языковую модель, используя max правдоподобия
- *байесовский* (Bayesian Learning: LDA [2001] и др.):
обучает языковую модель, используя байесовский вывод
- *нейросетевой* (NTM, Neural Topic Modeling):
использует предобученную большую языковую модель

Некоторые приложения тематического моделирования

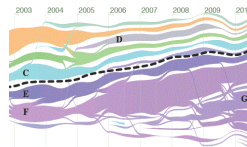
разведочный поиск в
электронных библиотеках



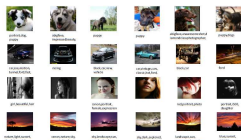
поиск тематических
сообществ в соцсетях



выявление и отслеживание
цепочек новостей



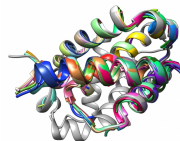
мультимодальный поиск
текстов и изображений



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.

BTM в исторических исследованиях: газетные архивы

[1] Корпус *Pennsylvania Gazette* 1728–1800, 25М 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.

BTM в исторических исследованиях: летописи и дневники

- [1] Двужычный корпус книг на английском и немецком:
 - все темы, связанные с эпистемологией
- [2] Корпус текстов на китайском языке (Цин, 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>.

ВТМ в исторических исследованиях: анализ периодики

Статьи коллекции JSTOR доступны в виде «мешков слов».

[1] Научные периодические журналы XX века:

- различия тематики на английском и немецком языках;
- особенно исследовались различия, связанные со 2МВ;
- для объединения тем использовались интервики Википедии.

[2] Литературно-художественная периодика (более 100 лет):

- 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.

ВТМ 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.

BTM 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)
— 1,5М 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.

ВТМ в некоторых наших прикладных исследованиях

- [1] «Поиск и классификация иголок в стоге сена»
 - тем о национальностях и межнациональных отношениях
 - в постах и комментариях социальных медиа (ЖЖ, ВК)
- [2] Определение политической поляризации мнений в СМИ
 - на основе различий в применении приёмов пропаганды
- [3] Детекция новых трендов в научных публикациях
- [4] Выявление динамики тем о российско-украинских отношениях в дискуссиях Твиттера до начала СВО

-
- 1. *M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko, K. Vorontsov*. Mining ethnic content online with additively regularized topic models. 2016.
 - 2. *D. Feldman, T. Sadekova, K. Vorontsov*. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. Dialogue 2020.
 - 3. *N. Gerasimenko, A. Chernyavskiy, M. Nikiforova, A. Ianina, K. Vorontsov*. Incremental topic modeling for scientific trend topics extraction. Dialogue 2023.
 - 4. *A. Sytnik, P. Chernikova, K. Vorontsov, M. Bazlutckaia*. Digital publics and the Ukraine dilemma: topic modelling of the cumulative twitter discussion. SCSM 2024.

Оптимизационный подход. Постановка задачи BTM

Дано: коллекция текстовых документов как «мешков-слов»

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

Найти: тематическую вероятностную языковую модель

$$p(w|d) = \sum_{t \in T} p(w|\underset{\text{г.у.н.}}{d}, t) p(t|d) = \sum_{t \in T} \phi_{wt} \theta_{td} \text{ с параметрами}$$

- $\phi_{wt} = p(w|t)$ — вероятность термина w в теме t
- $\theta_{td} = p(t|d)$ — вероятность темы t в документе d

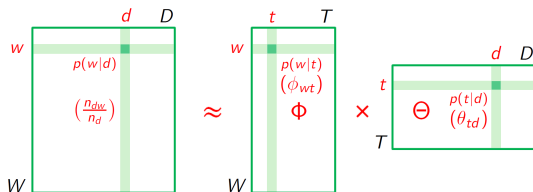
Критерий: максимум регуляризованного log-правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta} \quad (1)$$

$$\phi_{wt} \geq 0; \quad \sum_w \phi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_t \theta_{td} = 1 \quad (2)$$

Три интерпретации постановки задачи ВТМ

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

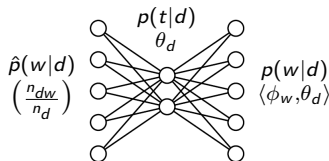


3. **Автокодировщик** документов в тематические эмбединги:

- кодировщик $f_\Phi: \frac{n_{dw}}{n_d} \rightarrow \theta_d$
- декодировщик $g_\Phi: \theta_d \rightarrow \Phi \theta_d$

задача реконструкции:

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



Зачем нужна регуляризация

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 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 (ARTM)

Теорема ARTM. Решение (Φ, Θ) задачи математического программирования (1)–(2) удовлетворяет системе уравнений

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

где $\text{norm}_{t \in T}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$ — операция нормировки вектора

ЕМ-алгоритм — решение данной системы уравнений методом простых итераций, т. е. чередованием Е-шага и М-шага

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

Основная лемма (о максимизации на единичных симплексах)

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

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω . Если ω_j — вектор локального экстремума задачи $f(\Omega) \rightarrow \max$ при ограничениях $\sum_{i \in I_j} \omega_{ij} = 1$, $\omega_{ij} \geq 0$, $i \in I_j$, $j \in J$, и если $\exists i: \omega_{ij} \frac{\partial f}{\partial \omega_{ij}} > 0$, то ω_j удовлетворяет системе уравнений

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

где $\operatorname{norm}_{i \in I} (x_i) = \frac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$ — операция нормировки вектора

Численное решение системы — методом простых итераций

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

Доказательство основной леммы

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

■

Доказательство Теоремы ARTM (по основной лемме)

Применим лемму к log-правдоподобию с регуляризатором:

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

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

Рациональная реализация ЕМ-алгоритма для ARTM

Вход: коллекция D , число тем $|T|$, число итераций $i_{\max}$;

Выход: матрицы термов тем Φ и тем документов Θ ;

$\phi_{wt} := \text{norm}_w(\text{rand})$; $\theta_{td} := 1/|T|$ для всех $w \in W$, $d \in D$, $t \in T$;

для всех итераций $i = 1 \dots i_{\max}$

$n_{wt} := 0$ для всех $w \in W$, $t \in T$;

для всех документов $d \in D$

$n_{td} := 0$ для всех $t \in T$;

для всех термов $w \in d$

$n_{tdw} := n_{dw} \text{norm}_{t \in T}(\phi_{wt} \theta_{td})$ для всех $t \in T$;

$n_{wt} += n_{tdw}$; $n_{td} += n_{tdw}$ для всех $t \in T$;

$\theta_{td} := \text{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$ для всех $t \in T$;

$\phi_{wt} := \text{norm}_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ для всех $w \in W$, $t \in T$;

Модификация М-шага, улучшающая сходимость

В формулах М-шага вместо ϕ_{wt} и θ_{td} от предыдущей итерации лучше подставлять несмещённые частотные оценки (PLSA)

$\phi_{wt}^* = \frac{n_{wt}}{n_t}$ и $\theta_{td}^* = \frac{n_{td}}{n_d}$, полученные в текущей итерации:

$$\phi_{wt} = \text{norm}_{w \in W} \left(n_{wt} + \phi_{wt}^* \frac{\partial R(\Phi^*, \Theta^*)}{\partial \phi_{wt}} \right)$$
$$\theta_{td} = \text{norm}_{t \in T} \left(n_{td} + \theta_{td}^* \frac{\partial R(\Phi^*, \Theta^*)}{\partial \theta_{td}} \right)$$

Доказано и/или экспериментально подтверждено, что

- значение регуляризованного правдоподобия увеличивается
- его монотонный рост начинается уже со второй итерации
- чем больше τ , тем заметнее улучшение сходимости
- не требуется дополнительных затрат времени или памяти

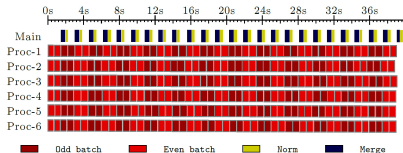
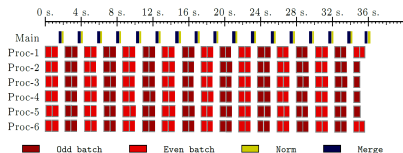
И.А.Ирхин, К.В.Воронцов. Сходимость алгоритма аддитивной регуляризации тематических моделей. 2020.

Онлайновый EM: батчи, параллельность, асинхронность

Батчи: пакеты документов, можно обрабатывать параллельно

Online: обновляем Φ после каждого батча (или нескольких)

DetAsync: используем старую Φ , пока новая ещё не готова

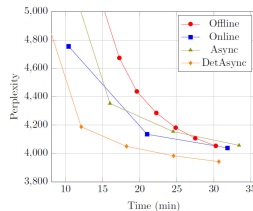
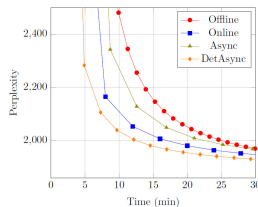


Скорость сходимости

на данных

Wikipedia-3.7M (слева)

Pubmed-8.2M (справа)



O.Frei, M.Apishev. Parallel non-blocking deterministic algorithm for online topic modeling. 2016.

Библиотека BigARTM

Ключевые возможности:

- онлайн-параллельный пакетный EM-алгоритм
- комбинирование регуляризаторов для построения тематических моделей с требуемыми свойствами
- встроенная библиотека регуляризаторов и метрик качества

Открытый код:

- <https://github.com/bigartm>
- 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++, Python

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

Пример 1. Интерпретируемость тем — ключевое свойство ВТМ

216 175 русско-английских пар статей Википедии.

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.

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

Пример 1. Интерпретируемость тем — ключевое свойство ВТМ

216 175 русско-английских пар статей Википедии.

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.

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

Пример 2. Интерпретируемость тем в биграммной модели

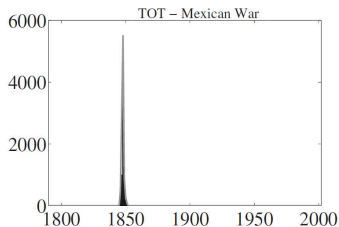
Коллекция 1000 статей конференций ММРО, ИОИ на русском

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

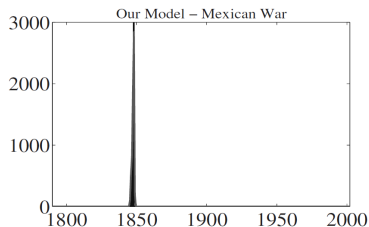
Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели // Магистерская диссертация, МФТИ, 2015.

Пример 3. Модальности униграмм, n -грамм и времени

Коллекция «State of the Union addresses» — выступления президентов США 1790–2008, 10 Мб, $|D| = 20\,431$, $|T| = 50$



1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

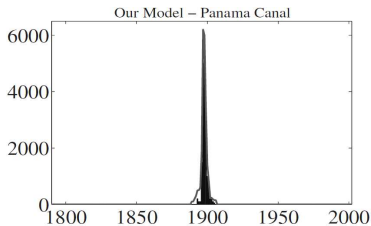
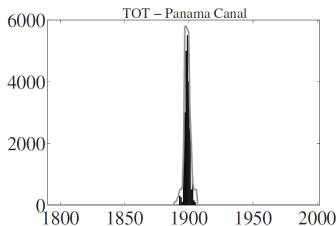


1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An N -gram topic model for time-stamped documents. ECIR-2013.

Пример 3. Модальности униграмм, n -грамм и времени

Коллекция «State of the Union addresses» — выступления президентов США 1790–2008, 10 Мб, $|D| = 20\,431$, $|T| = 50$



1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An N -gram topic model for time-stamped documents. ECIR-2013.

Классические модели PLSA и LDA (Latent Dirichlet Allocation)

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

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

M-шаг — частотные оценки условных вероятностей:

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



Thomas
Hofmann

LDA (латентное размещение Дирихле) [2001]:

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

M-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.

Байесовский подход. Основное отличие от оптимизационного

X — исходные данные, Ω — параметры порождающей 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:

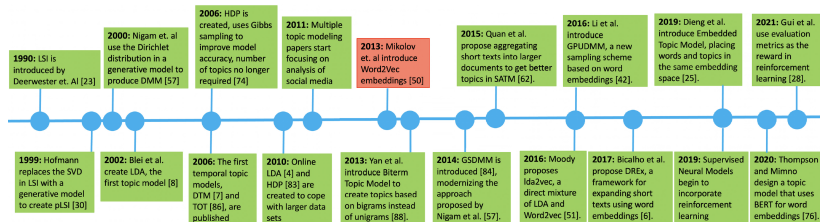
$$\Omega := \arg \max_{\Omega} (\ln p(X|\Omega) + \ln \text{Prior}(\Omega|\gamma))$$

Многокритериальная аддитивная регуляризация (ARTM)

обобщает MAP на любые регуляризаторы и их комбинации:

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

Эволюция тематического моделирования



1999 PLSA — Probabilistic Latent Semantic Analysis

2001 LDA — Latent Dirichlet Allocation

200x мультимодальные, темпоральные, иерархические и др. PTMs

2013 модели битермов и WNTM — аналоги word2vec

2016 тематические модели на основе предобученных word2vec

2020 BERTopic — TM на основе предобученного BERT

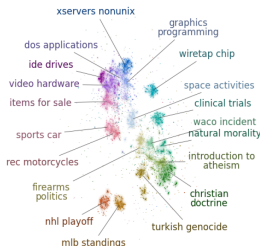
202x огромное разнообразие NTMs — Neural Topic Models...

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

Нейросетевая тематическая модель Contextual-Top2Vec

Вместо PTM — конвейер 8 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), автоматическое определение 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

Недостатки:

- это не единая модель, а конвейер эвристических моделей
- долго-дорого, особенно на больших коллекциях
- инкрементное добавление документов не предполагается

Достоинства — что хотелось бы перенять в ARTM:

- вместо «мешка слов» — последовательность $w_1, \dots, w_n$
- вместо документов — локальные контексты термов
- модель внимания (начнём с упрощённого варианта)
- эмбединг текста = среднее эмбедингов его термов
- быстрый отбор тематичных фраз для суммаризации тем
- быстрое разбиение документа на тематические сегменты

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

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

Нейросетевые и вероятностные языковые модели

Преимущества нейросетевых языковых моделей

- *генеративность*: способны порождать связный текст
- *универсальность*: решают широкий класс задач NLP/NLU
- *предобученность*: «знают всё о языке» (и о мире)

Преимущества вероятностных тематических моделей:

- *интерпретируемость* тематических эмбедингов
- *эффективность* в задачах «мягкой» кластеризации текстов
- *полнота* тематической кластерной структуры коллекции

Как «объединить лучшее от двух миров»?

Что объединяет PTM и LLM, и что их разобщает:

- ⊕ обе — вероятностные языковые модели,
- ⊕ обе — автокодировщики, векторные представления текста
- ⊖ PTM: мешок-слов, архитектура матричного разложения, байесовское обучение, трудности предобучения и др.

Идея тематизации текста за один линейный проход

Наводящие соображения:

- первая итерация ЕМ-алгоритма с инициализацией $\theta_{td}^0 = \frac{1}{|T|}$:

$$\theta_{td}(\Phi) = \text{norm} \left(\sum_{t \in T} \sum_{w \in W} n_{dw} p_{tdw} \right) = \sum_{w \in d} \frac{n_{dw}}{n_d} \text{norm} \left(\phi_{wt} \theta_{td}^0 \right)$$

- формула полной вероятности:

$$\theta_{td}(\Phi) = \sum_{w \in d} p(w|d) p(t|w, \text{г.у.н.} d) = \sum_{w \in d} \frac{n_{dw}}{n_d} \text{norm} \left(\phi_{wt} p(t) \right)$$

Преимущества:

- быстрая обработка документа за одну итерацию
- вычисление $p(t|s)$ фрагмента s усреднением $p(t|w)$, $w \in s$
- ограничение-равенство $\Theta = f(\Phi)$ — по сути регуляризатор
- гипотеза: такая модель более устойчива
- гипотеза: меньше переобучение на коротких текстах

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

ЕМ-алгоритм для ARTM с явным выражением Θ через Φ

Максимизация логарифма правдоподобия:

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

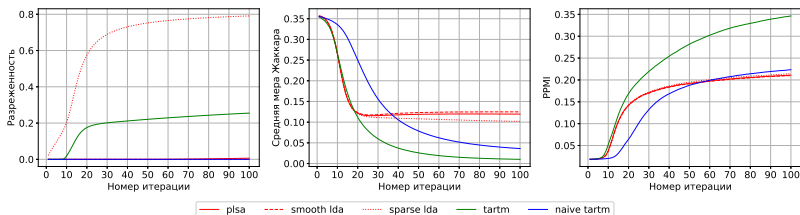
$$\begin{array}{l} \text{Е-шаг:} \\ \text{пост-} \\ \text{обработка} \\ \text{Е-шага} \\ \text{М-шаг:} \end{array} \left\{ \begin{array}{l} p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}); \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \\ p'_{tdw} = p_{tdw} + \frac{1}{n_{dw}} \sum_{s \in T} \frac{n_{sd}}{\theta_{sd}} \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} \\ \phi_{wt} = \text{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p'_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \end{array} \right.$$

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

Эксперимент. Проверка модифицированного EM-алгоритма

Коллекция NIPS, $|T| = 50$, модели:

- TARTM (Θ_{less} ARTM) — новый EM-алгоритм
- naïve 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

Тематическая модель локального контекста Attentive ARTM

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

$C_i \subset \{1, \dots, n\}$ — локальный контекст (окружение) терма w_i

Найти: $\phi_{tw} = p(t|w)$ — параметры тематической модели и α_{ci} — коэффициенты внимания, веса термов w_c из C_i для w_i

$$p(w|C_i) = \sum_{t \in T} p(w|t)p(t|C_i) = \sum_{t \in T} p(t|w) \frac{p(w)}{p(t)} p(t|C_i)$$

$$p(t|C_i) \equiv \theta_{ti} = \sum_{c \in C_i} \alpha_{ci} p(t|w_c), \quad \sum_{c \in C_i} \alpha_{ci} = 1, \quad \alpha_{ci} \geq 0$$

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

$$\sum_{i=1}^n \ln \sum_{t \in T} \phi_{tw_i} \frac{p(w_i)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} + R(\Phi) \rightarrow \max_{\Phi, A}$$

Ilya Diyakov, Konstantin Vorontsov. Attentive additively regularized topic modeling for local context. 2026 (статья в доработке)

ЕМ-алгоритм для тематической модели локального контекста

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\begin{aligned}
 \text{Е-шаг:} & \left\{ \begin{aligned} \theta_{ti} &= \sum_{c \in C_i} \alpha_{ci} \phi_{twc} & p_{ti} &= \text{norm}_{t \in T}(\phi_{twi} \theta_{ti} / p(t)) \end{aligned} \right. \\
 \text{М-шаг:} & \left\{ \begin{aligned} q_{wi} &= \sum_{c \in C_i} \alpha_{ci} [w_c = w] & N_{tw} &= \sum_{i=1}^n q_{wi} \frac{p_{ti}}{\theta_{ti}} \\ n_{tw} &= \sum_{i=1}^n p_{ti} [w_i = w] & n_t &= \sum_{i=1}^n p_{ti} \\ \phi_{tw} &= \text{norm}_{t \in T} \left(n_{tw} + \phi_{tw} N_{tw} - \phi_{tw} n_t \frac{p(w)}{p(t)} + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} \right) \\ p(t) &= \sum_{w \in W} \phi_{tw} p(w) \end{aligned} \right.
 \end{aligned}$$

Какова сложность вычислений? Придётся ли что-то упрощать?

Доказательство (по основной лемме)

$$Q(\Phi) = \sum_{i=1}^n \ln \left(\sum_{s \in T} \phi_{sw_i} \frac{p(w_i)}{p(s)} \theta_{si} \right) + R(\Phi) \rightarrow \max_{\Phi}$$

$$\frac{\partial \phi_{sw_i}}{\partial \phi_{tw}} = [w_i = w][t = s]$$

$$\left(\frac{ab}{c} \right)' = \frac{a'b}{c} + \frac{ab'}{c} - \frac{abc'}{c^2}$$

$$\frac{\partial \theta_{si}}{\partial \phi_{tw}} = \frac{\partial}{\partial \phi_{tw}} \left(\sum_{c \in C_i} \alpha_{ci} \phi_{sw_c} \right) = \sum_c \alpha_{ci} [w_c = w][t = s] = q_{wi}[t = s]$$

$$\frac{\partial p(s)}{\partial \phi_{tw}} = \frac{\partial}{\partial \phi_{tw}} \left(\sum_{v \in W} \phi_{sv} p(v) \right) = p(w)[t = s]$$

$$\begin{aligned} \phi_{tw} \frac{\partial Q}{\partial \phi_{tw}} &= \sum_{i=1}^n \frac{p(w_i)}{p(w_i|C_i)} \phi_{tw} \sum_{s \in T} \frac{\partial}{\partial \phi_{tw}} \left(\frac{\phi_{sw_i} \theta_{si}}{p(s)} \right) + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} = \\ &= \sum_{i=1}^n \frac{p(w_i)}{p(w_i|C_i)} \phi_{tw} \left(\frac{\theta_{ti}[w_i = w]}{p(t)} + \frac{\phi_{tw_i} q_{wi}}{p(t)} - \frac{\phi_{tw_i} \theta_{ti} p(w)}{p^2(t)} \right) + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} = \\ &= \sum_{i=1}^n \underbrace{\frac{\phi_{tw_i} \theta_{ti} p(w_i)}{p(w_i|C_i) p(t)}}_{= p_{ti}} \left([w_i = w] + \phi_{tw} \frac{q_{wi}}{\theta_{ti}} - \phi_{tw} \frac{p(w)}{p(t)} \right) + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} = \\ &= \underbrace{\sum_{i=1}^n p_{ti} [w_i = w]}_{n_{tw}} + \phi_{tw} \underbrace{\sum_{i=1}^n q_{wi} \frac{p_{ti}}{\theta_{ti}}}_{N_{tw}} - \phi_{tw} \underbrace{\sum_{i=1}^n p_{ti} \frac{p(w)}{p(t)}}_{n_t} + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} \end{aligned}$$

■

Осмысление М-шага в Attentive ARTM

Интерпретация вспомогательных переменных:

$p(t) = \sum_w p(t|w)p(w) = \sum_w \phi_{tw}p(w)$ — доля темы t в коллекции

n_{tw} — сколько раз терм w относился к теме t в коллекции

n_t — сколько термов отнесены к теме t в коллекции, $\frac{n_t}{n} \approx p(t)$

$q_{wi} = \sum_{c \in C_i} \alpha_{ci}[w_c = w]$ — суммарный вес термина w в контексте C_i

$N_{tw} = \sum_{i=1}^n q_{wi} \frac{p_{ti}}{\theta_{ti}}$ — суммарный вес термина w в контекстах темы t

$$\phi_{tw} = \text{norm}_{t \in T} \left(n_{tw} + \phi_{tw} N_{tw} - \phi_{tw} n_t \frac{p(w)}{p(t)} + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} \right)$$

При подстановке несмещённых оценок $\phi_{tw}^* = \frac{n_{tw}}{n_w}$ и $p^*(t) = \frac{n_t}{n}$

$$\phi_{tw}^* n_t \frac{p(w)}{p^*(t)} = \frac{n_{tw}}{n_w} n_t \frac{n_w}{n_t} = n_{tw}$$

Что представляет из себя оставшийся член $\phi_{tw} N_{tw}$?

Регуляризирующее воздействие локального контекста

N_{tw} тем больше, чем чаще w встречается в контекстах темы t :

$$N_{tw} = \sum_{i=1}^n q_{wi} \frac{p_{ti}}{\theta_{ti}} = \sum_{i=1}^n q_{wi} \frac{p(t|C_i, w_i)}{p(t|C_i)} = \sum_{i=1}^n q_{wi} \frac{p(w_i|t)}{p(w_i|C_i)}$$

$q_{wi} = \sum_c \alpha_{ci} [w_c = w]$ — весовая доля термина w в контексте C_i

N_{tw} похож на регуляризатор R_0 такой что $\frac{\partial R_0}{\partial \phi_{tw}} = N_{tw}$;
он сближает ϕ_{tw} термов, встречающихся рядом, аналогично
моделям дистрибутивной семантики word2vec, моделям
BitermTM и WordNetworkTM, регуляризаторам когерентности

N_{tw} повышает когерентность? $\Rightarrow$ интерпретируемость тем?

N_{tw} вычисляется за $O(k^2 |T|)$, где $k = |C_i|$ — длина контекстов,
при этом остальные операции с документом занимают $O(k |T|)$

Возможно ли уйти от трудоёмкого вычисления N_{tw} ?

Частный случай: контекст равен документу, «мешок слов»

$I_d = \{i_d^{\text{beg}}, \dots, i_d^{\text{end}}\}$ — термы документа d в сквозной нумерации

$C_i = I_d \Leftrightarrow i \in I_d$ — локальный контекст = весь документ

$\alpha_{ci} = \frac{1}{n_d}[c \in I_d]$ — внимание равномерно по документу

Тогда $\theta_{ti} = \theta_{td} = p(t|d)$, $q_{wi} = \frac{n_{dw}}{n_d} = \hat{p}(w|d)$ — не зависят от i ,

$N_{tw} = \sum_d \frac{n_{dw}}{n_d} (\sum_{i \in I_d} p_{ti}) \frac{n_d}{n_{td}} = n_w$, значит, $\phi_{tw}^* N_{tw} = n_{tw}$

ЕМ-алгоритм с линейной тематизацией документов

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

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

Эвристическое упрощение М-шага в Attentive ARTM

Второй член равен первому, если контекст равен документу

Третий член равен первому при подстановке $\phi_{tw}^*, p^*(t)$

Третий член равен нулю, если $p(t)$ — внешний параметр

$$\phi_{tw} = \text{norm}_{t \in T} \left(n_{tw} + \underbrace{\phi_{tw} N_{tw}}_{\approx n_{tw}} - \underbrace{\phi_{tw} n_t \frac{p(w)}{p(t)}}_{\approx n_{tw} \text{ или } = 0} + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} \right)$$

Гипотеза: второй и третий члены можно отбросить

$$\begin{aligned} \text{Е-шаг:} & \begin{cases} p_{ti} = \text{norm}_{t \in T} (\phi_{tw_i} \theta_{ti} / p(t)) & \theta_{ti} = \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} \end{cases} \\ \text{М-шаг:} & \begin{cases} \phi_{tw} = \text{norm}_{t \in T} \left(n_{tw} + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} \right) & n_{tw} = \sum_{i=1}^n p_{ti} [w_i = w] \\ p(t) = \sum_{w \in W} \phi_{tw} p(w) \end{cases} \end{aligned}$$

Осмысление Е-шага в Attentive ARTM

Е-шаг, один проход документа — аналог Self-attention:

- $p(t|w_i) = \phi_{tw_i}$ — контекстно-независимый вектор терма w_i
- $p(t|C_i) = \theta_{ti} = \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c}$ — вектор его контекста
- $p(t|C_i, w_i) = p_{ti}$ — контекстно-зависимый вектор терма w_i

$$p_{ti} = \text{norm}_{t \in T} \left(\phi_{tw_i} \frac{1}{p(t)} \theta_{ti} \right) = \text{norm}_{t \in T} \left(\sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} \frac{1}{p(t)} \phi_{tw_i} \right),$$

$\frac{1}{p(t)} \phi_{tw_i}$ — вектор-запрос (query),

ϕ_{tw_c} — вектор-значение (value),

α_{ci} — коэффициент внимания,

в роли ключа (key) — покомпонентное умножение векторов

Е-шаг, L проходов — аналог Looped Transformer:

$p(t|C_i, w_i)$ используется в качестве $p(t|w_i)$ следующего прохода

Гипотеза: это аналогично L необучаемым блокам внимания

ЕМ-алгоритм для Attentive ARTM

Вход: $\{w_1, \dots, w_n\}$; пакеты $B_1 \sqcup \dots \sqcup B_M = \{1, \dots, n\}$; параметры $|T|$, K , L , τ ;

Выход: векторы термов: $\Phi = (\phi_{tw})_{T \times W}$, контекстные $(p_{ti})_{T \times n}$;

инициализация $\phi_{tw} := \text{norm}_t(\text{rand})$, $p(t) := 1/|T|$, для всех $t \in T$, $w \in W$;

для всех итераций $k = 1 \dots K$ (проходов по всей коллекции)

инициализация $n_{tw} := 0$, $N_{tw} := 0$, для всех $t \in T$, $w \in W$;

для всех пакетов B_m , $m = 1, \dots, M$

$p_{ti} := \phi_{tw_i}$, для всех $t \in T$, $i \in B_m$;

для всех $l = 1 \dots L$ (по аналогии с Looped Transformer)

$\theta_{ti} := \sum_{c \in C_i} \alpha_{ci} p_{tc}$, для всех $t \in T$, $i \in B_m$;

$p_{ti} := \text{norm}_{t \in T}(p_{ti} \theta_{ti} / p(t))$, для всех $t \in T$, $i \in B_m$;

$N_{tw} := N_{tw} + \sum_{i \in B_m} \frac{p_{ti}}{\theta_{ti}} \sum_{c \in C_i} [w_c = w] \alpha_{ci}$, для всех $t \in T$, $w \in W$;

$n_{tw} := n_{tw} + \sum_{i \in B_m} [w_i = w] p_{ti}$; $\phi_{tw}^* := \text{norm}_{t \in T}(n_{tw})$, для всех $t \in T$, $w \in W$;

$\phi_{tw} := \text{norm}_{t \in T}\left(n_{tw} + \tau \phi_{tw}^* N_{tw} + \phi_{tw}^* \frac{\partial R(\Phi^*)}{\partial \phi_{tw}^*}\right)$, для всех $t \in T$, $w \in W$;

$p(t) := \sum_{w \in W} \phi_{tw} p(w)$, для всех $t \in T$;

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):

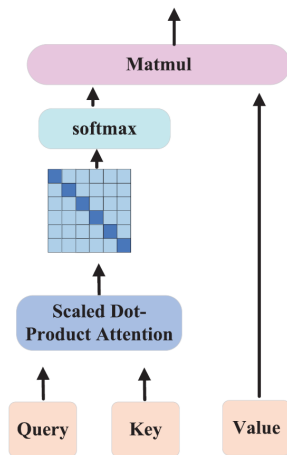
$$h_i = \sum_{c \in C_i} W_v x_c \underbrace{\text{SoftMax}_{c \in C_i} \langle W_k x_c, W_q x_i \rangle}_{\alpha_{ci}}$$

$W_v x_c$ — вектор-значение (value)

$W_k x_c$ — вектор-ключ (key)

$W_q x_i$ — вектор-запрос (query)

W_q, W_k, W_v — обучаемые параметры



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

Аналогия Attentive ARTM с моделью само-внимания

Контекстный тематический вектор на выходе Е-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \text{norm} \left(\sum_{c \in C_i} \phi_{tw_c} \alpha_{ci} \frac{1}{p(t)} \phi_{tw_i} \right)$$

Контекстный вектор на выходе модели само-внимания:

$$h_i = \sum_{c \in C_i} W_v x_c \alpha_{ci} = \sum_{c \in C_i} W_v x_c \text{SoftMax}_{c \in C_i} \langle W_k x_c, W_q x_i \rangle$$

Сходство:

- вектор терма w_i трансформируется в контекстный вектор
- путём усреднения векторов термов w_c из его контекста,
- наиболее схожих с ним по тематике

Отличия локализованного Е-шага:

- адамарово умножение вектора ϕ_{w_c} 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

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:

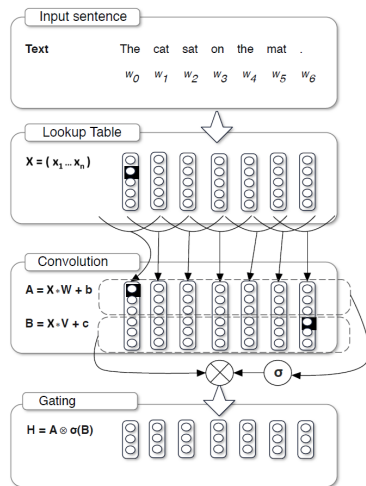
$$h_i = a_i \otimes \sigma(b_i), \text{ где}$$

$$a_i = \sum_{c \in C_i} W_c x_c \text{ — свёртка-контекст,}$$

$$b_i = \sum_{c \in C_i} V_c x_c \text{ — свёртка-фильтр,}$$

W_c, V_c — 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 Attentive ARTM с моделью GCNN

Контекстный тематический вектор на выходе Е-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \text{norm} \left(\sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} \frac{1}{p(t)} \phi_{tw_i} \right)$$

Контекстный вектор на выходе модели GCNN:

$$h_i = \left(\sum_{c \in C_i} W_c x_c \right) \otimes \sigma \left(\sum_{c \in C_i} V_c x_c \right)$$

Сходство:

- вектор термина w_i трансформируется в контекстный вектор
- путём усреднения векторов ϕ_{w_c} его контекста,
- семантически схожих с вектором термина w_i , фильтруемых адамаровым умножением на неотрицательный вектор

Отличия локализованного Е-шага:

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

Быстрое вычисление коэффициентов внимания

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

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

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

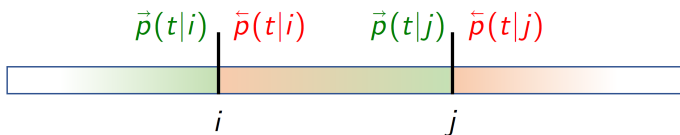
где $\vec{\gamma}_i, \bar{\gamma}_i$ — коэффициенты сглаживания в позиции i

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

Как задать $\vec{\gamma}_i, \bar{\gamma}_i$ — коэффициенты сглаживания в позиции 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 w_i
- γ_i можно умножать на оценку важности термина w_i

Использование двунаправленных векторов контекста



Через *двунаправленные тематические векторы* определяется:

- $\vec{p}_{ti} = \vec{p}(t|i)$ — тематика левого контекста термина w_i
- $\vec{p}_{ti} = \vec{p}(t|i)$ — тематика правого контекста термина w_i
- $p_{ti} = \frac{1}{2}(\vec{p}_{ti} + \vec{p}_{ti})$ — тематика двустороннего контекста w_i
- $\max_i \|\vec{p}_{ti} - \vec{p}_{ti}\|$ — граница i между сегментами
- при различных γ_i — короткие и длинные контексты
- $p(t|i..j) = \frac{1}{2}(\vec{p}_{ti} + \vec{p}_{tj})$ — тематика сегмента $[i..j]$
- $\vec{p}_{ti} \approx \vec{p}_{tj}$ — однородность тематики сегмента $[i..j]$
- возможность быстрого отбора фраз по заданной теме

Обучение весов внимания — Adaptive Attentive ARTM

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

$$\sum_{i=1}^n \ln \sum_{t \in T} \phi_{tw_i} \frac{p(w_i)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c} + R(\Phi, A) \rightarrow \max_{\Phi, A}$$

Идея: $\alpha_{ci} = \beta_{w_i, c-i}$ привяжем к термам словаря $w \in W$, чтобы терм $w_i = w$ образовал контекст $C_i = \{i - K, \dots, i + K\}$,
 $\beta_{wk} \geq 0$, $\sum_{k=-K}^K \beta_{wk} = 1$ — распределение весов в контексте терма w .

Необходимые условия экстремума (по основной лемме):

$$\beta_{wk} = \underset{k \in \{-K, \dots, K\}}{\text{norm}} \left(\beta_{wk} \sum_{i=1}^n \sum_{t \in T} p_{ti} \frac{\phi_{tw_{i+k}}}{\theta_{ti}} [w_i = w] + \beta_{wk} \frac{\partial R}{\partial \beta_{wk}} \right)$$

Эксперименты: 3 коллекции, 5 метрик качества, 9 моделей

Dataset	Train docs	Test docs	Classes	Vocabulary	Train tokens	Test tokens
20 Newsgroups	10,975	7,290	20	9,348	835,008	513,395
AG News	120,000	7,600	4	10,483	2,664,355	167,299
DBpedia14	559,975	69,997	14	11,329	12,887,091	1,609,248

Метрики качества:

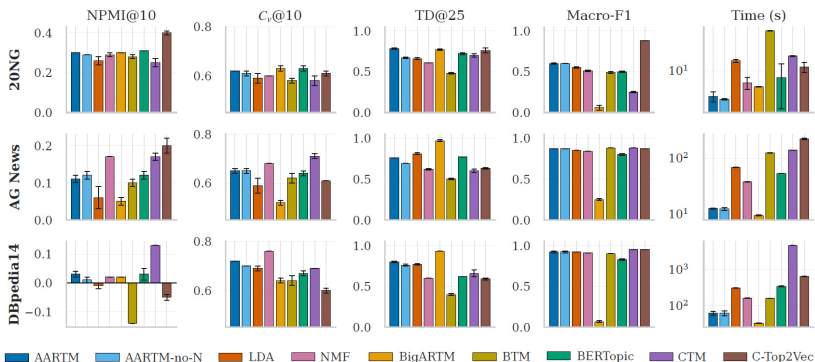
- NPMI@10, $C_v@10$ — две разновидности когерентности
- TD@25 — различность тем
- macro-F1 — классификация d по $p(t|d)$, логрегрессия
- time(s) — время обучения модели, без предобработки

Модели:

- AARTM, AARTM-no-N — две разновидности AARTM
- PTMs: LDA, BTM (Biterm TM), BigARTM (PLSA, $R = 0$)
- NTMs: BERTopic, CTM (Combined TM), Contextual-Top2Vec
- NMF: non-negative matrix factorization

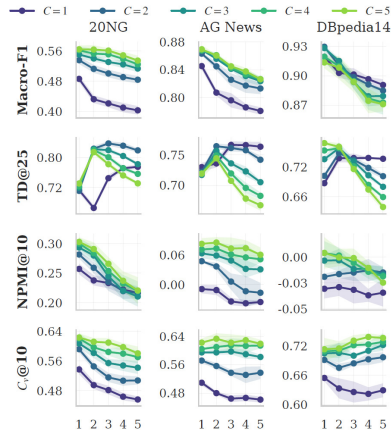
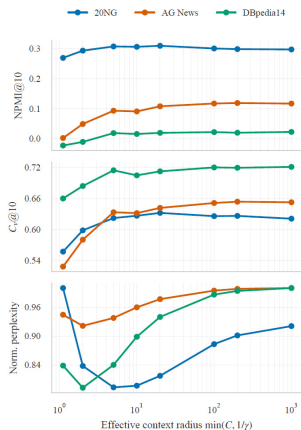
Эксперименты: основное сравнение

$|T| = 100$ тем, $K = 50$ итераций, $|C_i| = 100$, $L = 1$, $\tau \in \{0, 1\}$



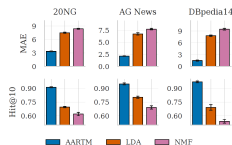
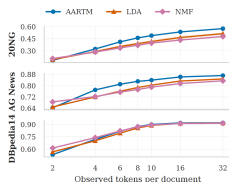
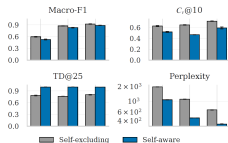
- высокая скорость при сопоставимых оценках качества
- отбрасывание N_{tw} ($\tau = 0$) почти не ухудшает AARTM
- странный провал качества классификации у BigARTM :-)

Эксперименты: влияние параметров $C = |C_i|$, L



- длины контекстного окна 10, как правило, достаточно
- $L = 1$ достаточно, дополнительные проходы не нужны
- лучше увеличивать контекст C , чем число проходов L

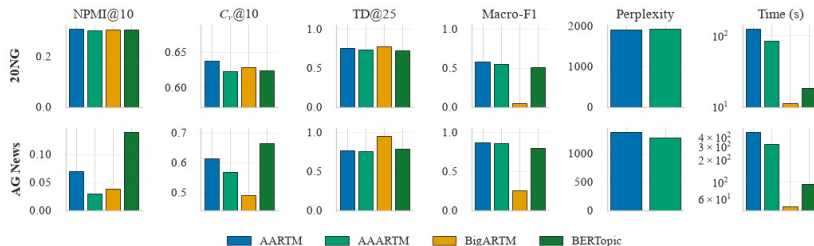
Эксперименты: влияние параметров $C = |C_i|$ и L



- self-aware — включение термина в свой контекст $i \in C_i$ ведёт к переобучению и ухудшает все метрики качества на тестовых данных
- AARTM немного улучшает качество классификации документов по их коротким фрагментам
- AARTM гораздо успешнее определяет место склейки сегментов текста, взятых из разных классов

Обучение весов внимания — Adaptive Attentive ARTM

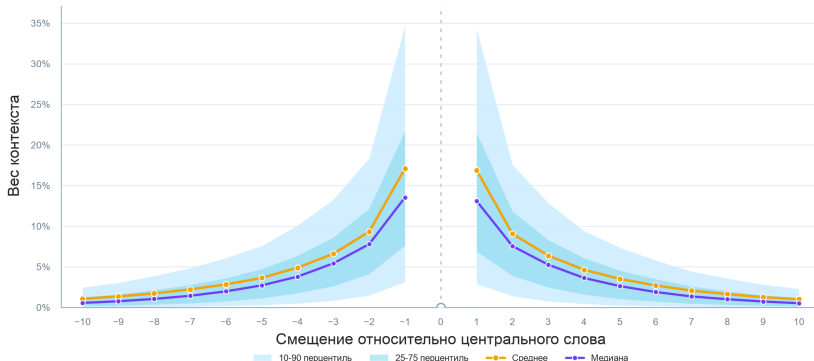
Сравнение: 2 коллекции, 6 метрик качества, 4 модели



- A²RTM и A³RTM примерно сопоставимы по качеству
- Удалось добиться эффективности A³RTM по времени
- Добиться заметных преимуществ по качеству путём обучения весов внимания пока не удалось — видимо, эта модель слишком проста, мало отличается от A²RTM

Эксперименты: обучение весов внимания

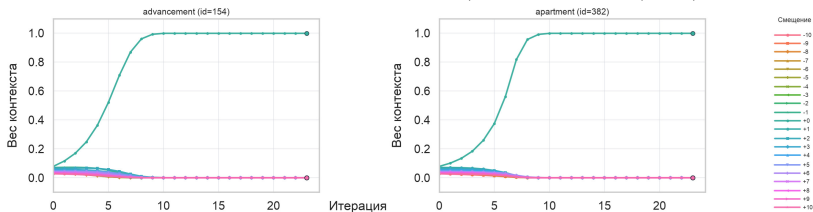
Распределение весов внимания термов β_{wk} , $k = -10, \dots, 10$



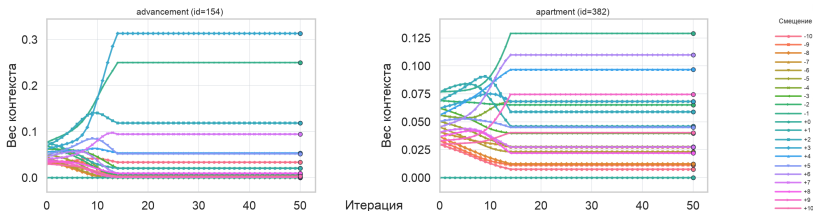
- в среднем β_{wk} быстро убывают как влево, так и вправо
- для отдельных термов монотонность может нарушаться
- целесообразна параметризация, особенно для редких термов

Эксперименты: обучение весов внимания

Сходимость весов внимания термов β_{wk} , $k = -10, \dots, 10$



• при $i \in C_i$ веса внимания переобучены и вырождены



Выводы по экспериментам с тематическим вниманием

- проблема « $i \in C_i$ » в обычных РТМ не осознавалась, хотя на коротких текстах приводила к переобучению
- пока опробованы самые простые модели внимания
- оптимальные длинные контексты сами собой не возникли
- нужны более сложные и содержательные модели внимания
- планируется усложнять внимание постепенно, сохраняя вычислительную эффективность ЕМ-алгоритма ARTM
- идея следующего эксперимента:

$\alpha_{ci} = \beta_{w_i, c-i}$ — терм w распределяет веса по контексту;

$\alpha_{ci} = \text{norm}_{c \in C_i}(\gamma_{wc} \beta_{wc, i-c})$ — терм w воздействует

на контексты своих соседей с силой β_{wk} ,

где γ_w — распределение влиятельности термов $w \in W$.



Серия «Классический учебник МГУ»

Воронцов К. В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. — М.: URSS, 2025. — 224 с. ISBN 978-5-9710-9933-8

Обзор ARTM (68 публикаций на 01-05-2026)

Воронцов К. В. Аддитивная регуляризация тематических моделей: теория и приложения. — 40 с.



Дальнейшая цель — «Make PTMs Great Again» — задать новый стандарт вероятностного тематического моделирования:

- от байесовского обучения — к аддитивной регуляризации
- от мешка слов — к локальным контекстам и вниманию
- от BigARTM — к новой эффективной библиотеке ARTM
- разобраться, почему темы бывают неинтерпретируемыми