Тематические и нейросетевые вероятностные языковые модели: курс на сближение

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

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

МГУ Институт искусственного интеллекта

Проблемы искусственного интеллекта — совместный научный семинар Российской ассоциации искусственного интеллекта и ФИЦ «Информатика и управление» РАН
• 28 января 2026 •

Аннотация

Доклад состоит из трёх частей.

1. Обзор вероятностных моделей языка: частотных, тематических, дистрибутивных, глубоких нейросетевых.
Цели и задачи тематического моделирования, и почему оно остаётся актуальным в эпоху больших языковых моделей.

2. Теория аддитивной регуляризации тематических моделей (ARTM) концептуально приближает EM-алгоритм обучения вероятностных тематических моделей к градиентным методам обучения дистрибутивных и нейросетевых моделей языка.

3. Новая тематическая модель локальных контекстов позволяет окончательно уйти от гипотезы «мешка слов» — наиболее критикуемого допущения в тематическом моделировании.
По сути, это простейший вариант модели внимания без обучаемых параметров, с богатыми возможностями развития.

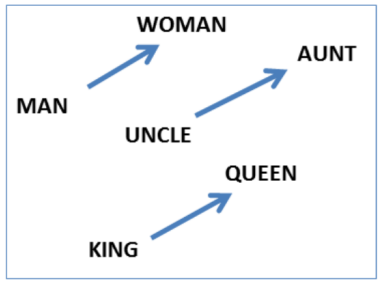
Содержание

Эволюция подходов машинного обучения в анализе текстов

Декомпозиция задач по уровням пирамиды NLP

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

Вероятностные модели языка на основе векторных представлений слов и матричных разложений



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

Нейросетевые модели локальных контекстов

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

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

Простейшая частотная вероятностная языковая модель

Дано:

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

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

Найти:

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

Критерий: максимум логарифма правдоподобия:

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

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

Решение (из условий Каруша–Куна–Таккера):

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

Вероятностная языковая модель коллекции документов

Дано:

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

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

Найти:

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

Критерий: максимум логарифма правдоподобия:

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

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

Решение (из условий Каруша–Куна–Таккера):

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

Вероятностная тематическая модель коллекции документов

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 языковую модель (в.п. $D \times W \times T$)

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

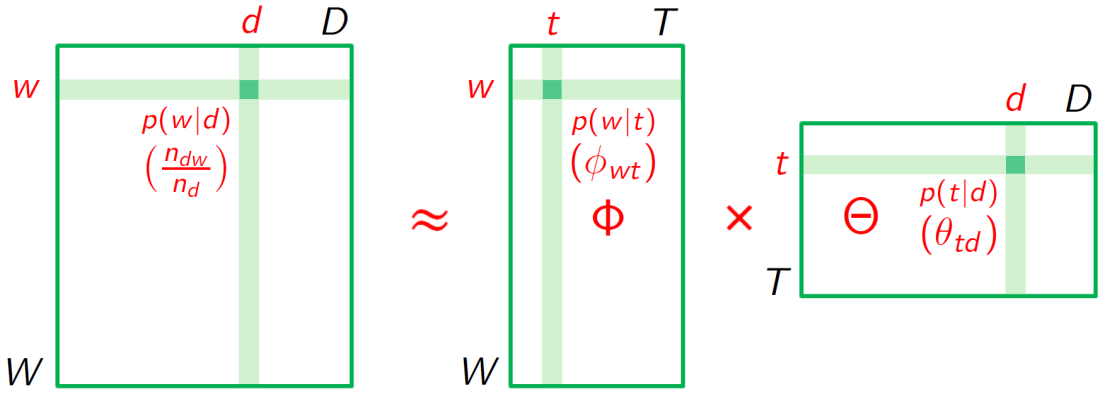
Критерий: правдоподобие предсказания слов w в документах d

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

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

Три интерпретации задачи тематического моделирования

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



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

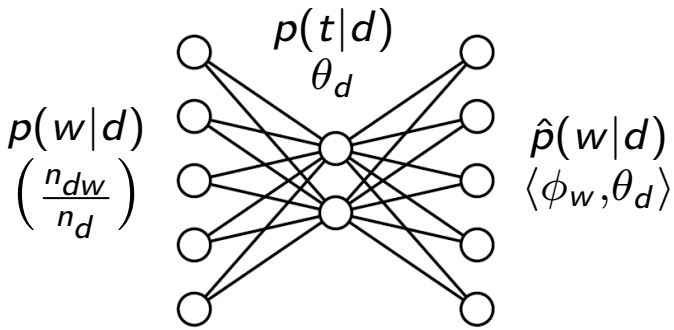
3. **Автокодировщик** документов в тематические эмбединги:

— 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

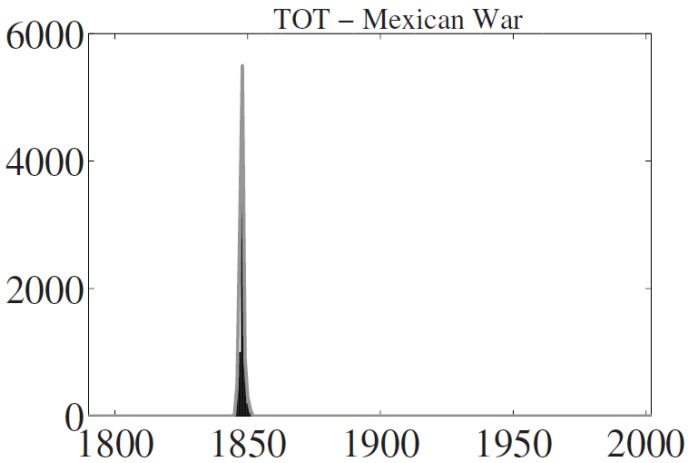
Коллекция 1000 статей конференций ММРО, ИОИ на русском

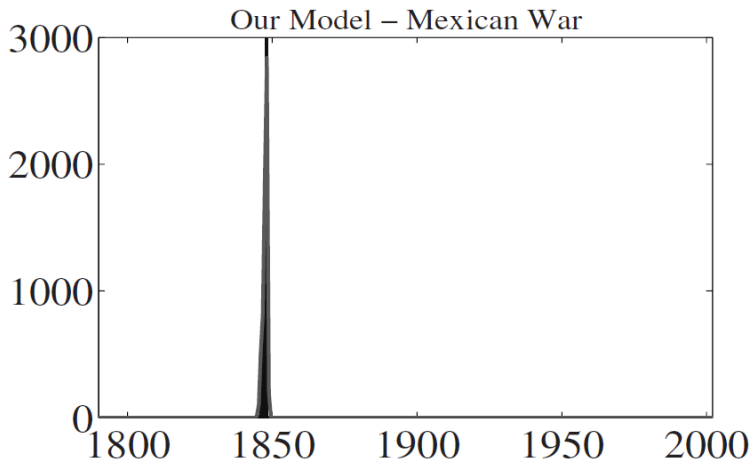
распознавание образов в биоинформатике		теория вычислительной сложности	
униграммы	биграммы	униграммы	биграммы
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	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.

Пример 3. Совмещение темпоральной и n-граммной модели

Коллекция еженедельных выступлений президентов США





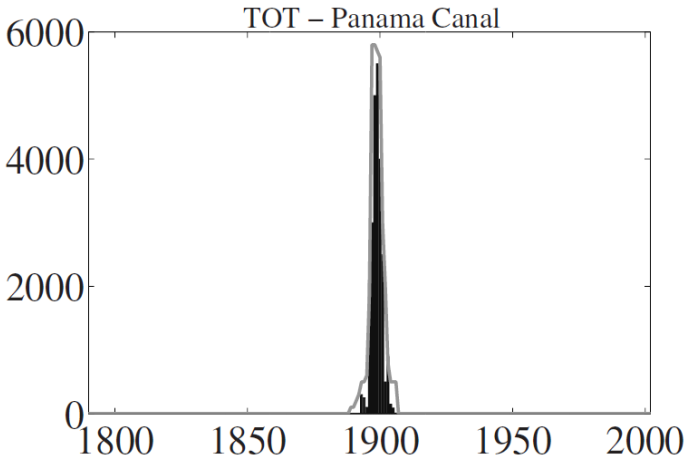
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

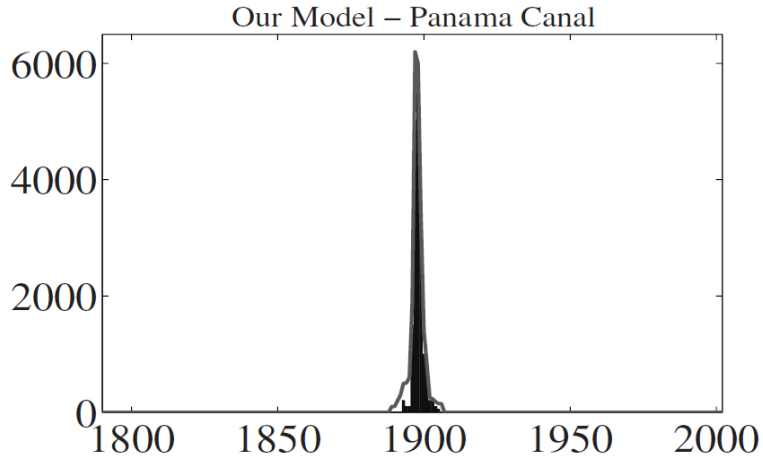
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

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

Пример 3. Совмещение темпоральной и n-граммной модели

Коллекция еженедельных выступлений президентов США





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. 2013.

Цели и не-цели тематического моделирования

Цели:

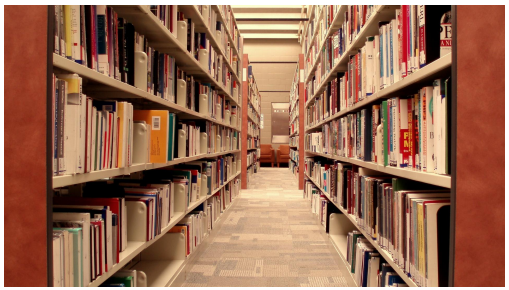
- выявлять тематическую кластерную структуру текстовой коллекции (сколько в ней тем, и о чём они), представляя результат в удобной для человека форме
- получать *интерпретируемые* тематические векторы (эмбединги) слов $p(t|w)$, слов-в-контексте $p(t|d,w)$, документов $p(t|d)$, фрагментов $p(t|s)$, объектов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

Не-цели:

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

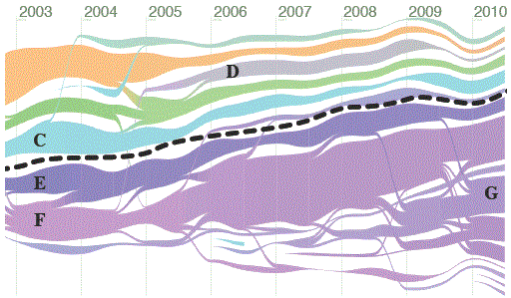
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

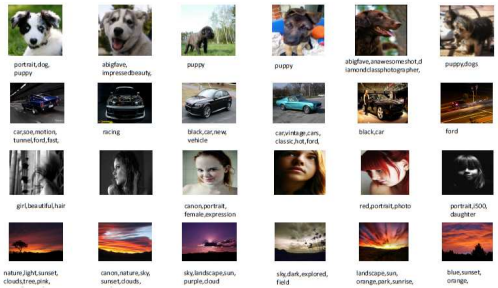


поиск тематических сообществ в соцсетях

выявление и отслеживание цепочек новостей

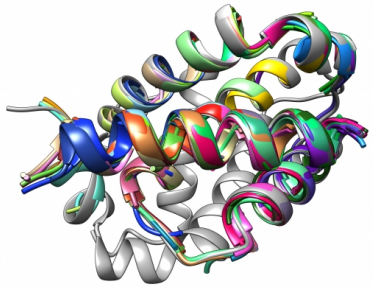


мультимодальный поиск текстов и изображений



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.

Дистрибутивная гипотеза и виды семантической близости слов

Как закодировать вектором не только темы, но смысл слова?
Смысл слова есть множество всех контекстов его употребления

- Words that occur in the same contexts tend to have similar meanings [Harris, 1954].
- You shall know a word by the company it keeps [Firth, 1957].

Синтагматическая близость слов:
сочетаемость слов в одном контексте
(здание–строитель, кран–вода, функция–точка)

Парадигматическая близость слов:
взаимозаменяемость слов в одном контексте
(здание–дом, кран–смеситель, функция–отображение)

Z.Harris. Distributional structure. 1954.
J.R.Firth. A synopsis of linguistic theory 1930-1955. Oxford, 1957.
P.Turney, P.Pantel. From frequency to meaning: vector space models of semantics. 2010.

Обучение векторных представлений слов (word2vec)

Дано: $\{w_1, \ldots, w_n\}$ — текст, последовательность слов (токенов)
$C_i = \{\ldots, w_i, \ldots\}$ — контекст слова w_i, например, $\pm k$ слов

Найти: векторы u_w, v_w (embedding), кодирующие смысл слов
$p(w|w_i) = \underset{w \in W}{\mathrm{SoftMax}} \langle u_w, v_{w_i} \rangle$ — языковая модель *Skip-gram*

Критерий: log-loss для бинарной классификации пар слов:

$$\sum_{i=1}^{n} \sum_{w \in C_i} \Big(\log p(+1|w, w_i) + \log p(-1|\bar{w}, w_i) \Big) \to \max_{U,V}$$

$p(y|w, w_i) = \sigma\big(y \langle u_w, v_{w_i} \rangle\big)$ — модель классификации, $y = \pm 1$;
$y = +1$, если w находится в контексте слова w_i;
$y = -1$, если w не находится в контексте слова w_i;
$\bar{w}$ сэмплируется из $p(w)^{3/4}$ (skip-gram negative sampling, SGNS)

T.Mikolov et al. Efficient estimation of word representations in vector space, 2013.

Связь word2vec с матричными разложениями

d — размерность векторов слов v_w и u_w словаря W
$V = (v_w)_{W \times d}$ — матрица предсказывающих векторов слов
$U = (u_w)_{W \times d}$ — матрица предсказываемых векторов слов

SGNS строит матричное разложение $P \approx UV^{\mathsf{T}}$ матрицы Shifted PMI (Point-wise Mutual Information):

$$P_{ab} = \ln \frac{n_{ab} n}{n_a n_b} - \ln k,$$

n_{ab} — частота пары слов a, b в окне $\pm k$ слов,
n_a, n_b — число пар с участием слова a и b соответственно,
n — число всех пар слов в коллекции.

В качестве эвристики используют также Shifted Positive PMI:

$$P^+_{ab} = \left(\ln \frac{n_{ab} n}{n_a n_b} - \ln k \right)_+ .$$

O.Levy, Y.Goldberg. Neural word embedding as implicit matrix factorization. 2014.

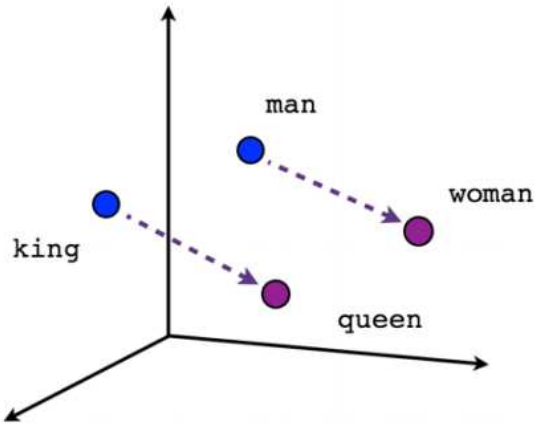
Проверка на задачах семантической близости и аналогии слов

Задача семантической близости слов:

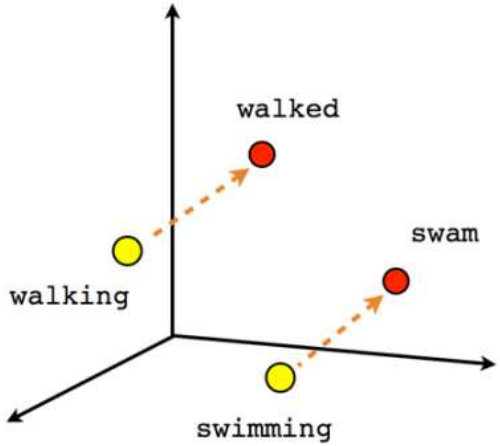
по выборке пар слов (a, b) оценивается корреляция Спирмена между $\cos(v_a, v_b)$ и экспертными оценками близости слов

Задача семантической аналогии слов:

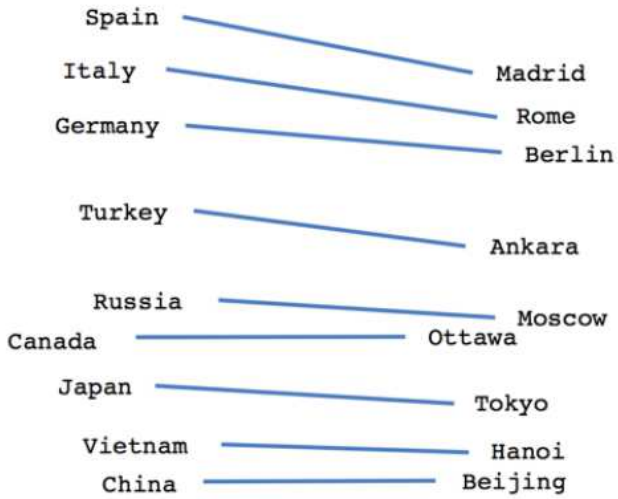
по трём словам угадать четвёртое



Male-Female



Verb tense



Country-Capital

Модель векторных представлений FastText

Идея: векторное представление слова w определяется как сумма векторов всех его буквенных n-грамм $G(w)$:

$$u_w = \sum_{g \in G(w)} u_g$$

В Skip-gram вместо векторов слов u_w обучаются векторы u_g

Пример: $G(\text{дармолюб}) = \{<\text{да}, \text{арм}, \text{рмо}, \text{мол}, \text{олю}, \text{люб}, \text{юб}>\}$

Преимущества:

- Это решает проблемы новых слов и слов с опечатками
- Подходит для обработки текстов социальных медиа
- Словарь 2- и 3-грамм много меньше словаря слов
- Существует много предобученных моделей

Bojanowski et al. Enriching word vectors with subword information. 2016.

Модели векторных представлений для текстов и графов

word2vec: *эмбединги* (векторные представления) слов
T.Mikolov et al. Efficient estimation of word representations in vector space. 2013.

paragraph2vec: эмбединги фрагментов или документов
Q.Le, T.Mikolov. Distributed representations of sentences and documents. 2014.

sent2vec: эмбединги предложений
M.Pagliardini et al. Unsupervised learning of sentence embeddings using compositional n-gram features. 2017.

FastText: эмбединги символьных *n*-грамм
`https://github.com/facebookresearch/fastText`

node2vec: эмбединги вершин графа
A.Grover, J.Leskovec. Node2vec: scalable feature learning for networks. 2016.

graph2vec: более общие эмбединги на графах
A.Narayanan et al. Graph2vec: learning distributed representations of graphs. 2017.

StarSpace: эмбединги чего угодно от Facebook AI Research
L.Wu, A.Fisch, S.Chopra, K.Adams, A.B.J.Weston. StarSpace: embed all the things! 2018.

BERT: эмбединги фраз и предложений от Google AI Language
J.Devlin et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018.

GPT-3: эмбединги, предобученные по 570Gb текстов от OpenAI
T.B.Brown et al. Language Models are Few-Shot Learners. 2020.

Трасформер для машинного перевода

Трасформер (transformer) — это нейросетевая архитектура для трансформации векторов слов в контекстно-зависимые

Схема преобразований данных в машинном переводе:

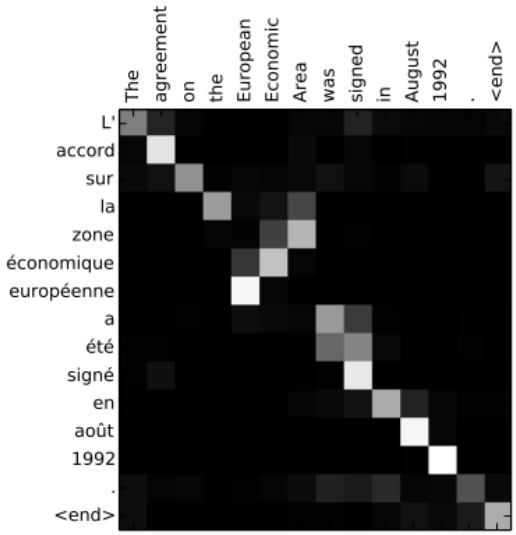
- $S = (w_1, \ldots, w_n)$ — слова предложения на входном языке
  $\downarrow$ обучаемая или пред-обученная векторизация слов
- $X = (x_1, \ldots, x_n)$ — векторы слов входного предложения
  $\downarrow$ трансформер-кодировщик
- $Z = (z_1, \ldots, z_n)$ — контекстно-зависимые векторы слов
  $\downarrow$ трансформер-декодировщик, похож на кодировщика
- $Y = (y_1, \ldots, y_m)$ — векторы слов выходного предложения
  $\downarrow$ генерация слов из построенной языковой модели
- $\tilde{S} = (\tilde{w}_1, \ldots, \tilde{w}_m)$ — слова предложения на выходном языке

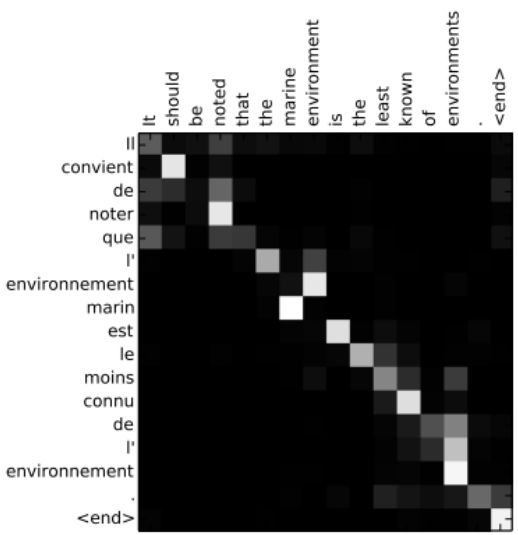
Vaswani et al. (Google) Attention is all you need. 2017.

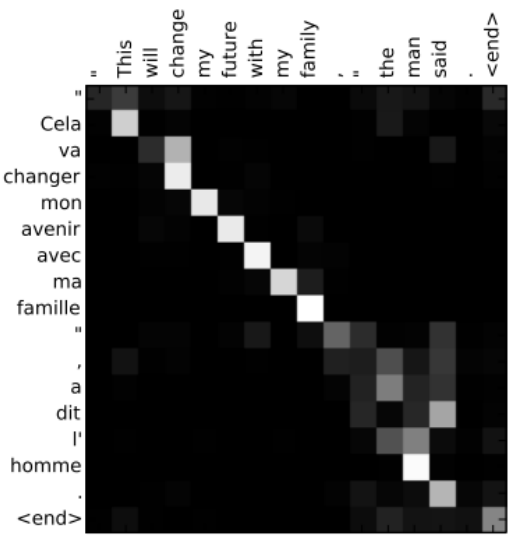
Модели внимания для машинного перевода

$X = (x_1, \ldots, x_n)$ — векторы слов входного предложения
$Y = (y_1, \ldots, y_m)$ — векторы слов выходного предложения

Модель внимания оценивает матрицу семантического сходства $A_{ti} = a(x_i, y_t)$ — насколько входное слово x_i важно (требует внимания) для обработки выходного слова y_t







Bahdanau et al. Neural machine translation by jointly learning to align and translate. 2015.

Модель внимания Query–Key–Value

q — *вектор-запрос* для трансформации в вектор-контекст z
$K = (k_1, \ldots, k_n)$ — *векторы-ключи*, сравниваемые с запросом
$X = (x_1, \ldots, x_n)$ — *векторы-значения*, образующие контекст

Модель внимания — трёхслойная сеть, вычисляющая z как выпуклую комбинацию векторов x_i, релевантных запросу q:

$$z = \text{Attn}(q, K, X) = \sum_i x_i \operatorname*{SoftMax}_i a(k_i, q),$$

где $a(k, q)$ — *оценка релевантности* ключа k запросу q,
например $a(k, q) = k^\mathsf{T} q$ или $k^\mathsf{T} W q$ с матрицей параметров W

Модель внутреннего внимания (самовнимания, self-attention):

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

трансформирует входную последовательность $X = (x_1, \ldots, x_n)$
в выходную последовательность векторов контекста $(z_1, \ldots, z_n)$

Архитектура трансформера-кодировщика

1. Добавляются позиционные векторы p_i:
$h_i = x_i + p_i, \quad H = (h_1, \ldots, h_n)$ $d = \dim x_i, p_i, h_i = 512$, $\dim H = 512 \times n$

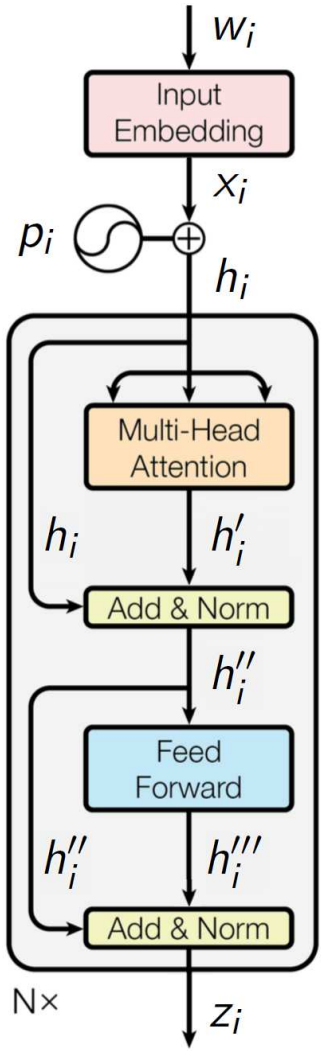
2. J голов самовнимания:
$h_i^j = \text{Attn}(W_q^j h_i, W_k^j H, W_v^j H)$ $j = 1, \ldots, J = 8$, $\dim h_i^j = 64$, $\dim W_q^j, W_k^j, W_v^j = 64 \times 512$

3. Конкатенация (multi-head attention):
$h'_i = \text{MH}_j(h_i^j) \equiv \left[h_i^1 \cdots h_i^J\right]$ $\dim h'_i = 512$

4. Сквозная связь + нормировка уровня:
$h''_i = \text{LN}(h'_i + h_i; \mu_1, \sigma_1)$ $\dim h''_i, \mu_1, \sigma_1 = 512$

5. Полносвязная 2х-слойная сеть FFN:
$h'''_i = W_2\,\text{ReLU}(W_1 h''_i + b_1) + b_2$ $\dim W_1 = 2048 \times 512$, $\dim W_2 = 512 \times 2048$

6. Сквозная связь + нормировка уровня:
$z_i = \text{LN}(h'''_i + h''_i; \mu_2, \sigma_2)$ $\dim z_i, \mu_2, \sigma_2 = 512$



Несколько дополнений и замечаний

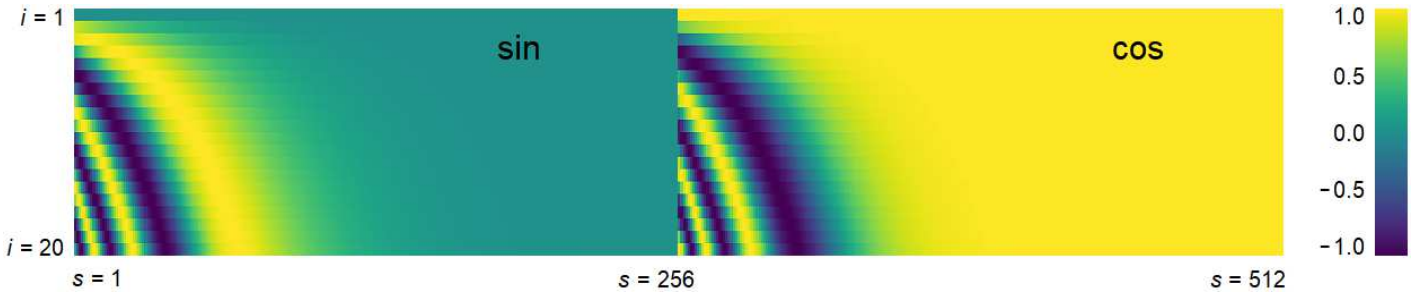
- $N = 6$ блоков $h_i \to \square \to z_i$ соединяются последовательно
- эмбединги слов $x_i \in \mathbb{R}^d$ — обучаемые или пред-обученные
- нормировка уровня (Layer Normalization), $x, \mu, \sigma \in \mathbb{R}^d$:

$$\mathrm{LN}_s(x; \mu, \sigma) = \sigma_s \frac{x_s - \bar{x}}{\sigma_x} + \mu_s, \quad s = 1, \ldots, d,$$

$\bar{x} = \frac{1}{d}\sum_s x_s$ и $\sigma_x^2 = \frac{1}{d}\sum_s (x_s - \bar{x})^2$ — среднее и дисперсия x

- Позиции слов i кодируются векторами p_i, $i = 1, \ldots, n$;
чем больше $|i - j|$, тем больше $\|p_i - p_j\|$, n не ограничено:

$$p_{is} = \sin(i\, 10^{-8\frac{s}{d}}), \quad p_{i,s+\frac{d}{2}} = \cos(i\, 10^{-8\frac{s}{d}}), \quad s = 1, \ldots, \tfrac{d}{2}$$

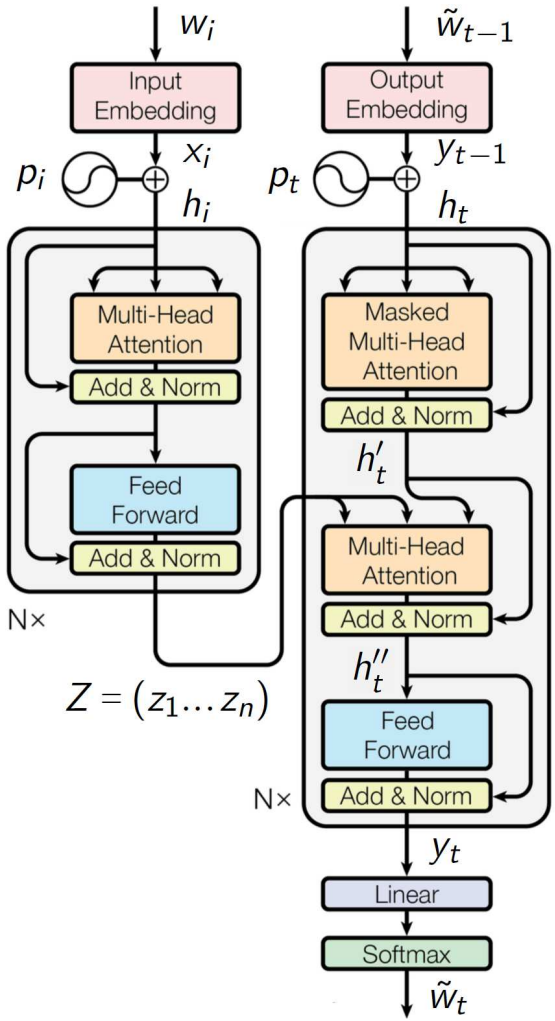


Архитектура трансформера декодировщика

Авторегрессионный синтез последовательности:
$y_0 = \langle \text{BOS} \rangle$ — вектор символа начала;
для всех $t = 1, 2, \ldots$:

1. Маскирование «данных из будущего»:
   $h_t = y_{t-1} + p_t; \quad H_t = (h_1, \ldots, h_t)$
2. Многомерное самовнимание:
   $h'_t = \text{LN} \circ \text{MH}_j \circ \text{Attn}(W^j_q h_t, W^j_k H_t, W^j_v H_t)$
3. Многомерное внимание на кодировку Z:
   $h''_t = \text{LN} \circ \text{MH}_j \circ \text{Attn}(\tilde{W}^j_q h'_t, \tilde{W}^j_k Z, \tilde{W}^j_v Z)$
4. Двухслойная полносвязная сеть:
   $y_t = \text{LN} \circ \text{FFN}(h''_t)$
5. Линейный предсказывающий слой:
   $p(\tilde{w}|t) = \underset{\tilde{w}}{\text{SoftMax}}(W_y y_t + b_y)$

генерация $\tilde{w}_t = \arg\max_{\tilde{w}} p(\tilde{w}|t)$ **пока** $\tilde{w}_t \neq \langle \text{EOS} \rangle$



Vaswani et al. (Google) Attention is all you need. 2017.

Критерии обучения и валидации для машинного перевода

Критерий для обучения параметров нейронной сети W по обучающей выборке предложений S с переводом $\tilde{S}$:

$$\sum_{(S,\tilde{S})} \sum_{\tilde{w}_t \in \tilde{S}} \ln p(\tilde{w}_t | t, S, W) \to \max_W$$

Критерий оценивания моделей (недифференцируемые) по выборке пар предложений «перевод S, эталон S_0»:

BiLingual Evaluation Understudy:

$$\text{BLEU} = \min\left(1, \frac{\Sigma \text{len}(S)}{\Sigma \text{len}(S_0)}\right) \underset{(S_0,S)}{\text{mean}} \left(\prod_{n=1}^{4} \frac{\#n\text{-грамм из } S, \text{ входящих в } S_0}{\#n\text{-грамм в } S} \right)^{\frac{1}{4}}$$

Word Error Rate:

$$\text{WER} = \underset{(S_0,S)}{\text{mean}} \left(\frac{\#\text{вставок} + \#\text{удалений} + \#\text{замен}}{\text{len}(S)} \right)$$

Vaswani et al. (Google) Attention is all you need. 2017.

BERT (Bidirectional Encoder Representations from Transformers)

Трансформер BERT — это кодировщик без декодировщика, предобучаемый на большой текстовой коллекции для решения широкого класса задач автоматической обработки текста

Схема преобразования данных в задачах NLP:

- $S = (w_1, \ldots, w_n)$ — токены предложения входного текста
  $\downarrow$ обучение эмбедингов вместе с трансформером
- $X = (x_1, \ldots, x_n)$ — эмбединги токенов входного предложения
  $\downarrow$ трансформер кодировщика
- $Z = (z_1, \ldots, z_n)$ — трансформированные эмбединги
  $\downarrow$ дообучение на конкретную задачу
- Y — выходной текст / разметка / классификация и т.п.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language)
BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Критерий MLM (masked language modeling) для обучения BERT

Критерий маскированного языкового моделирования MLM, строится автоматически по текстам (self-supervised learning):

$$\sum_{S} \sum_{i \in M(S)} \ln p(w_i | i, S, W) \to \max_{W},$$

где $M(S)$ — подмножество маскированных токенов из S,

$$p(w|i, S, W) = \underset{w \in V}{\text{SoftMax}}(W_z z_i(S, W_T) + b_z)$$

— языковая модель, предсказывающая i-й токен предложения S;

$z_i(S, W_T)$ — контекстный эмбединг i-го токена предложения S на выходе трансформера-кодировщика с параметрами W_T;

$W = (W_T, W_z, b_z)$ — все параметры языковой модели

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language) BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Критерий NSP (next sentence prediction) для обучения BERT

Критерий предсказания связи между предложениями NSP, строится автоматически по текстам (self-supervised learning):

$$\sum_{(S,S')} \ln p\big(y_{SS'}|S,S',W\big) \to \max_W,$$

где $y_{SS'} = [\text{за } S \text{ следует } S']$ — классификация пары предложений,

$$p(y|S,S',W) = \underset{y\in\{0,1\}}{\text{SoftMax}}\big(W_y\,\text{th}(W_s z_0(S,S',W_T) + b_s) + b_y\big)$$

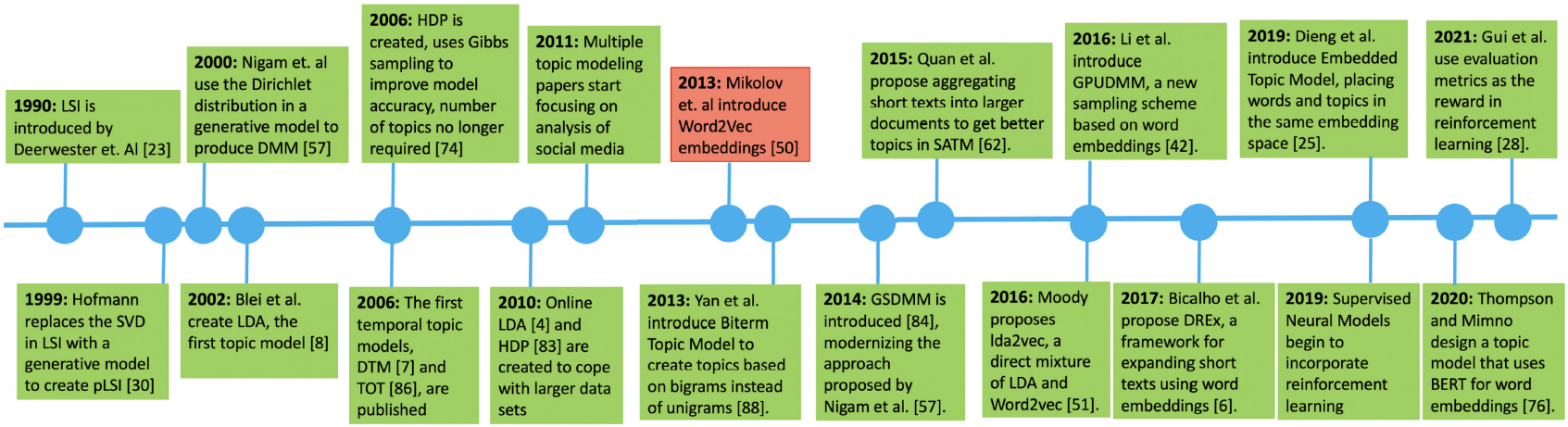
— вероятностная модель бинарной классификации пар (S,S'),
$z_0(S,S',W_T)$ — контекстный эмбединг токена $\langle\texttt{CLS}\rangle$ для пары предложений, записанной в виде $\langle\texttt{CLS}\rangle\ S\ \langle\texttt{SEP}\rangle\ S'\ \langle\texttt{SEP}\rangle$

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language)
BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Ещё несколько замечаний про трансформеры

- **Fine-tuning:** для дообучения на задаче задаётся модель $f(Z(S, W_T), W_f)$, выборка $\{S\}$ и критерий $\mathscr{L}(S, f) \to \max$
- **Multi-task learning:** для дообучения на наборе задач $\{t\}$ задаются модели $f_t(Z(S, W_T), W_t)$, выборки $\{S\}_t$ и сумма критериев $\sum_t \lambda_t \sum_S \mathscr{L}_t(S, f_t) \to \max$
- *GLUE*, *SuperGLUE*, *Russian SuperGLUE*, *MERA*, *SLAVA* — наборы тестовых задач на понимание и генерацию языка
- Трансформеры обычно строятся не на словах, а на токенах, получаемых BPE (Byte-Pair Encoding) или WordPiece
- Первый трансформер: $N = 6$, $d = 512$, $J = 8$, весов 65M
- $\text{BERT}_{\text{BASE}}$, GPT1: $N = 12$, $d = 768$, $J = 12$, весов 110M
- $\text{BERT}_{\text{LARGE}}$: $N = 24$, $d = 1024$, $J = 16$, весов 340M

Neural Topic Models — эволюция PTM в сторону LLM



Как «объединить лучшее от двух миров»?

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

Что объединяет PTM и LLM, и что их разобщает:

⊕ обе — вероятностные языковые модели,

⊕ обе — автокодировщики, векторные представления текста

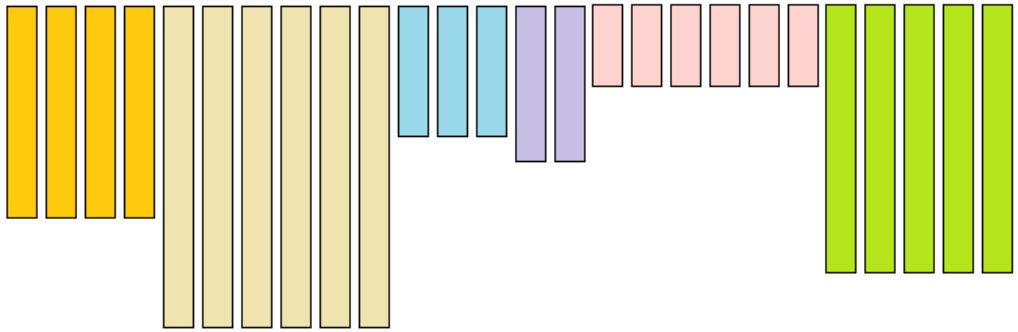
⊖ PTM: байесовское обучение, архитектура MF, мешок слов

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

Максимизация функции на единичных симплексах

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

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



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

Доказательство леммы о максимизации на симплексах

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

Функция Лагранжа:

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

Условия Каруша–Куна–Таккера для вектора ω_j:

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

Умножим обе части первого равенства на ω_{ij}:

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

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

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

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

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

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

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

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

А.Н.Тихонов
(1906–1993)

Аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

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

Доказательство (по лемме о максимизации на симплексах)

Применим лемму к log-правдоподобию с регуляризатором:

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

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

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

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

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

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, LDA: первые и 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:

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

Thomas Hofmann

LDA: latent Dirichlet allocation (латентное размещение Дирихле):

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

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

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

David Blei

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

Модификация M-шага, улучшающая сходимость

В формулах M-шага вместо ϕ_{wt} и θ_{td} можно подставлять несмещённые частотные оценки (PLSA) $\hat{\phi}_{wt} = \frac{n_{wt}}{n_t}$ и $\hat{\theta}_{td} = \frac{n_{td}}{n_d}$:

$$\phi_{wt} = \underset{w \in W}{\text{norm}} \left(n_{wt} + \hat{\phi}_{wt} \frac{\partial R(\hat{\Phi}, \hat{\Theta})}{\partial \phi_{wt}} \right)$$

$$\theta_{td} = \underset{t \in T}{\text{norm}} \left(n_{td} + \hat{\theta}_{td} \frac{\partial R(\hat{\Phi}, \hat{\Theta})}{\partial \theta_{td}} \right)$$

Доказано, что в результате такой модификации

- увеличивается значение регуляризованного правдоподобия
- монотонный рост регуляризованного правдоподобия начинается быстрее — как правило, 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.

От байесовского вывода к аддитивной регуляризации

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

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

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

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

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

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

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

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

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*

Библиотека BigARTM

Ключевые возможности:

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

Сообщество:

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

Лицензия и среда разработки:

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

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

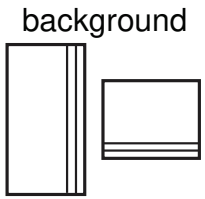
Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

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)

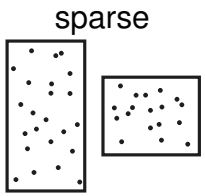
D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov. Fast and modular regularized topic modelling. FRUCT ISMW, 2017.

Регуляризаторы для улучшения интерпретируемости тем



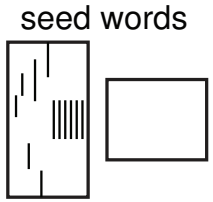
Сглаживание фоновых тем $B \subset T$:

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

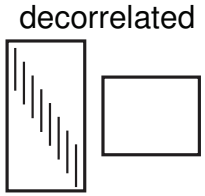


Разреживание предметных тем $S = T \setminus B$:

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

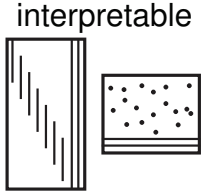


Сглаживание для выделения релевантных тем
с помощью словаря «затравочных» ключевых слов



Декоррелирование для повышения различности тем:

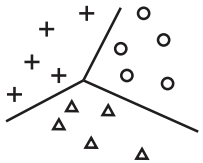
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

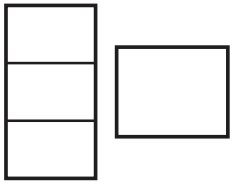
Регуляризаторы для мультимодальных тематических моделей

supervised



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

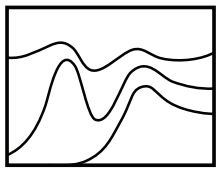
multilanguage



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

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

temporal



Темпоральные модели с модальностью времени i:

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

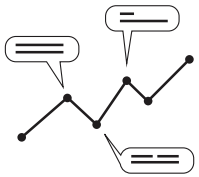
geospatial

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

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

Регуляризаторы для учёта взаимосвязей и зависимостей

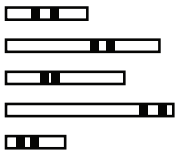
regression



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

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

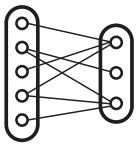
biterm



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

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

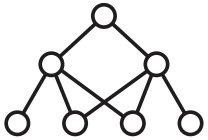
relational



Связи или ссылки между документами:

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

hierarchy



Связи родительских тем t с дочерними подтемами s:

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

Тематические модели дистрибутивной семантики

Тематическая модель битермов (Biterm Topic Model),
n_{uv} — частота битерма (слов u, v в общем контексте):

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

Тематическая модель сети слов (Word Network Topic Model),
d_u — псевдо-документ, объединение всех контекстов слова u:

$$R(\Phi, \Theta) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} \phi_{wt} \theta_{td_u}$$

Регуляризатор когерентности:

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

Xiaohui Yan, et al. A Biterm Topic Model for Short Texts. WWW 2013.
Yuan Zuo et al. Word Network Topic Model: a simple but general solution... 2014.
Mimno D. et al. Optimizing semantic coherence in topic models. EMNLP 2011.

Когерентность — мера интерпретируемости тем

Когерентность (согласованность) темы t по k топовым словам:

$$\text{coh}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{PMI}(w_i, w_j)$$

где w_i — i-е слово в порядке убывания ϕ_{wt},

$\text{PMI}(u,v) = \ln \frac{P_{uv}}{P_u P_v}$ — *поточечная взаимная информация* (pointwise mutual information),

P_{uv} — доля документов, в которых слова u, v хотя бы один раз встречаются рядом (в одном предложении или в окне 10 слов),

P_u — доля документов, в которых u встретился хотя бы 1 раз,

P_{uv}, P_u можно вычислять по другой коллекции (Википедии).

Когерентность модели = средняя когерентность всех тем.

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Эксперимент. Связь когерентности и интерпретируемости

Измерялась ранговая корреляция Спирмена экспертных оценок с каждой из 15 мер интерпретируемости.

PMI — лучшая метрика.

Gold-standard — средняя корреляция Спирмена между оценками разных экспертов.

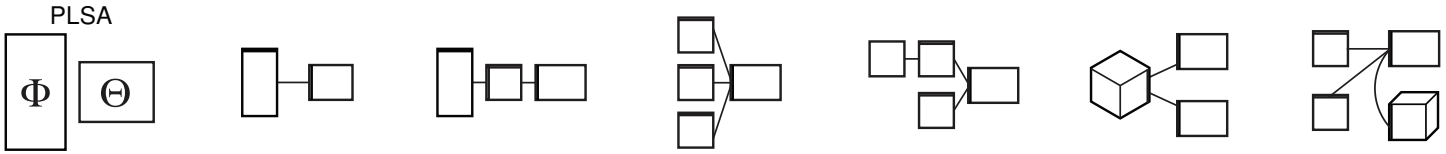
Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	−0.20	0.19
	LCH	−0.31	−0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	−0.08	0.27
	WUP	0.41	0.26
Wikipedia	RACO	0.62	0.69
	MIW	0.68	0.70
	DOCSIM	0.59	0.60
	PMI	**0.74**	**0.77**
Google	TITLES	0.51	
	LOGHITS	−0.19	
Gold-standard	IAA	0.82	0.78

Вывод: когерентность близка к «золотому стандарту».

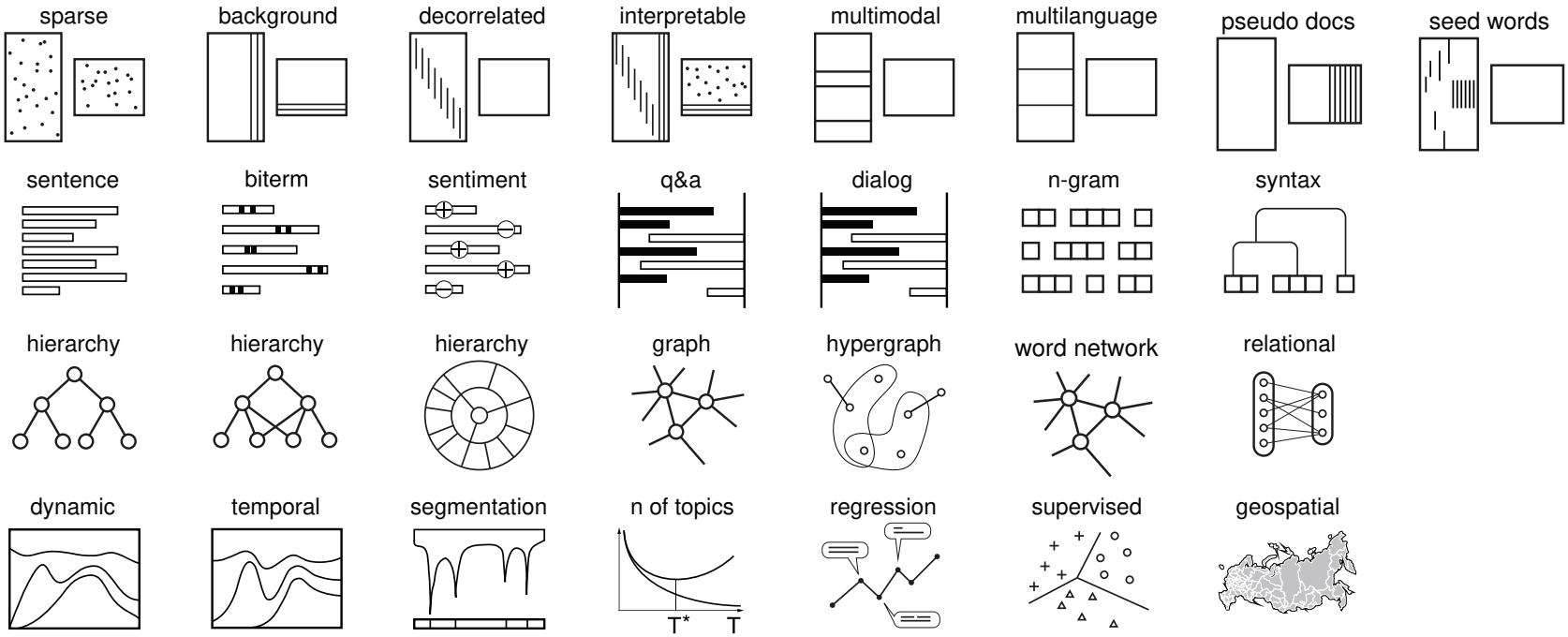
Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Язык пиктограмм для описания комбинаций моделей

Структуры матричных разложений в вероятностных моделях:



Регуляризаторы — дополнительные критерии и ограничения (в т.ч. взаимосвязи между термами, документами, темами):



Стратегии регуляризации (управление коэффициентами τ_i)

1. Регуляризация ведёт итерационный процесс к матричному разложению с требуемыми свойствами, но даёт смещённые оценки матриц Φ, Θ. На последней итерации имеет смысл вернуться к несмещённым оценкам (PLSA, $R(\Phi, \Theta) = 0$):

$$\phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt}\right)$$

$$\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td}\right)$$

2. Коэффициенты регуляризации τ_i можно менять в итерациях

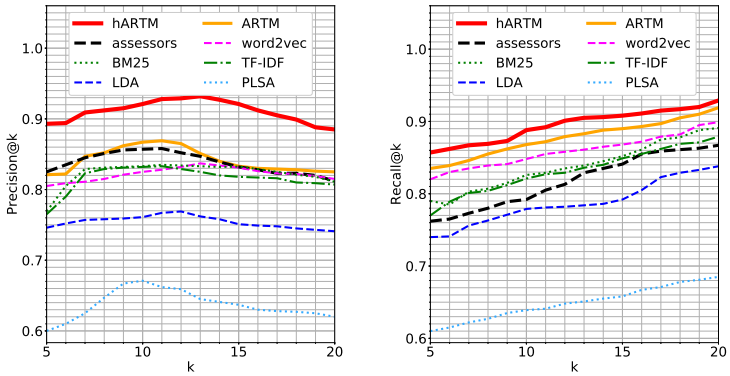
3. Регуляризаторы можно включать в определённом порядке

4. Регуляризаторы можно отключать по достижению эффекта

5. Одни регуляризаторы могут выполнять подготовительную работу для применения следующих регуляризаторов

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 → 1400 (Habr.ru), 475 → 2800 (TechCrunch.com).

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

Поиск и рубрикация научных публикаций на 100 языках

Цель: 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.

Поиск обсуждений болезней в социальных сетях

Цель: «поиск и классификация иголок в стоге сена» — сообщений в Twitter о болезнях, симптомах, способах лечения, побочных эффектах

Регуляризаторы:

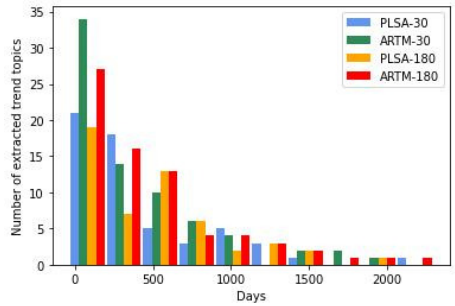
$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{seed words}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) +$$
$$+ R\left(\text{hierarchy}\right) + R\left(\text{temporal}\right) + R\left(\text{geospatial}\right) \to \max$$

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

M.J.Paul, M.Dredze. Discovering Health Topics in Social Media Using Topic Models, 2014

Выявление трендов в коллекции научных публикаций

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



Регуляризаторы:

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

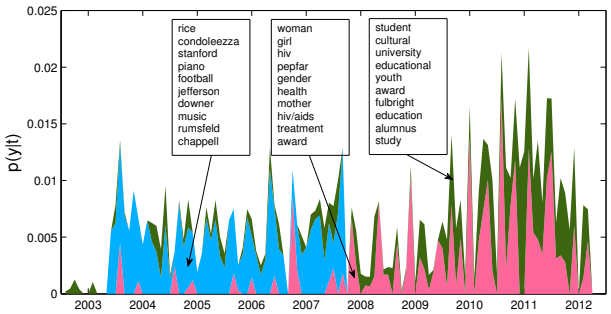
Результаты:

- 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:

$$\mathscr{L}\left(\Phi,\Theta\right) + R\left(\Phi,\Theta\right) + R\left(\Phi,\Theta\right) + R\left(\Phi,\Theta\right) + R\left(\Phi,\Theta\right) + R\left(\Phi,\Theta\right) \to \max$$

PLSA · interpretable · temporal · multimodal · n-gram · multilanguage

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:

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

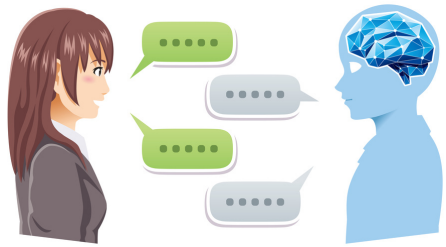
Результаты:

- 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.

Выявление намерений клиентов для построения чат-ботов

Цель: выявить тематику и интенты (намерения) клиентов по коллекции обращений в контактный центр. Построить рубрикатор интентов для последующей разметки диалогов.



Регуляризаторы:

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

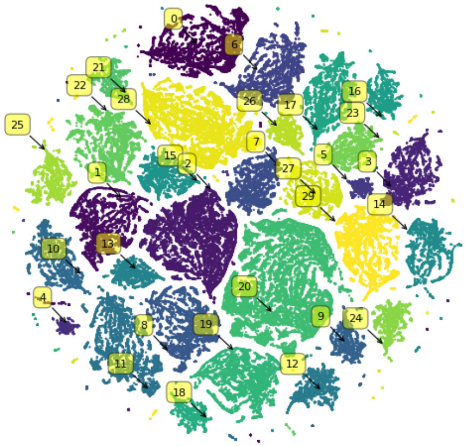
Результаты: точность классификации интентов 60% → 66%.

A.Popov, V.Bulatov, D.Polyudova, E.Veselova. Unsupervised dialogue intent detection via hierarchical topic model. RANLP, 2019.

Тематическая модель банковских транзакционных данных

Цель: Выявление паттернов потребительского поведения клиентов банка, причём

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



Регуляризаторы:

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

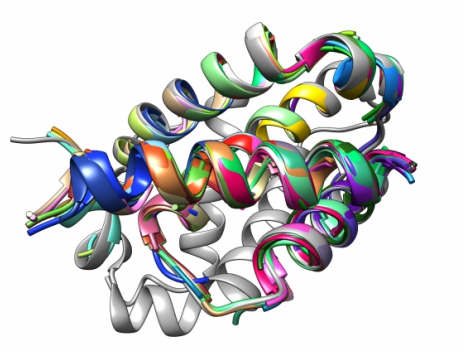
Результаты:

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

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

Обработка последовательностей нуклеотидов или аминокислот

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



Регуляризаторы (гипотеза):

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

Такая модель легко реализуема в BigARTM.

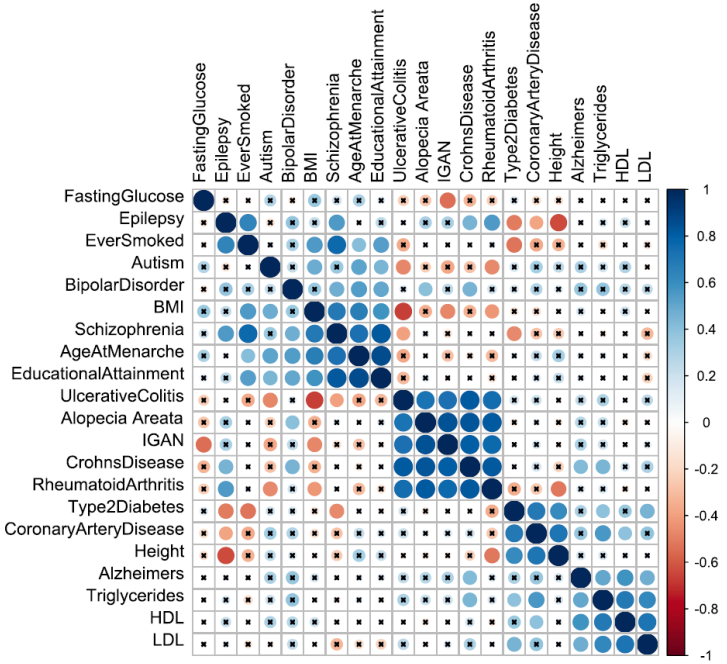
J.B.Gutierrez, K.Nakai. A study on the application of topic models to motif finding algorithms. 2016.

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

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

Функциональная аннотация генома человека

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



Регуляризаторы (гипотеза):

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

Такая модель легко реализуема в BigARTM.

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

«Make PTM Great Again» — почему это разумная цель

Почему рано отказываться от PTM, когда вокруг LLM

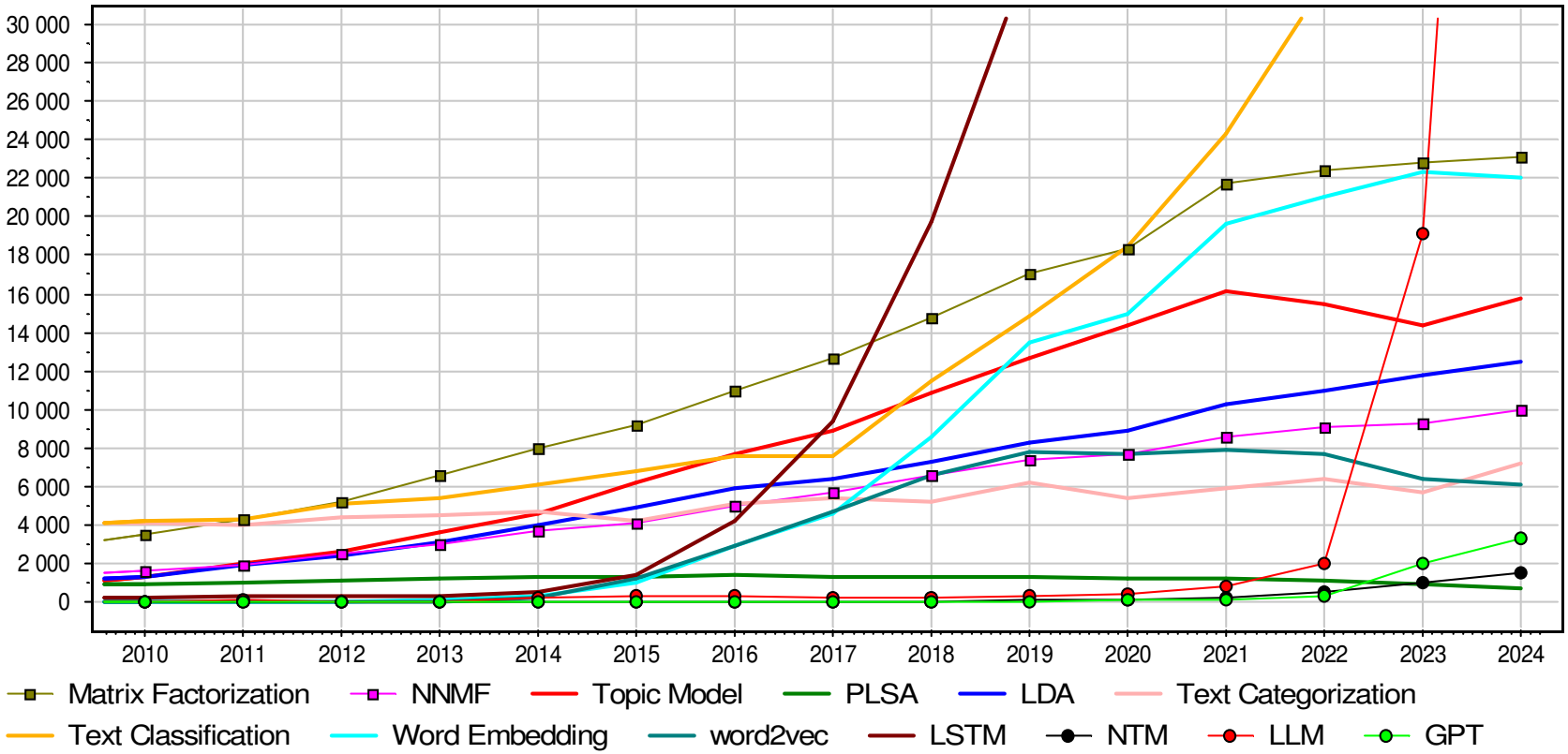
- разведочный тематический анализ и фильтрация тем по-прежнему нужны в социогуманитарных исследованиях, научном поиске, анализе трендов, социальных сетей
- PTM решают узкий класс задач лучше и быстрее, чем LLM благодаря интерпретируемости, простоте и полноте
- PTM — это мягкая кластеризация за линейное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

Как наводить мост через пропасть между PTM и LLM

- ушли от байесовского вывода к комбинированию моделей
- уходим от «мешка слов» к контекстному вниманию
- переходим от документов к локальным контекстам слов
- будем: устранять недостатки интерпретируемости тем,
- генерировать заголовки и аннотации тем с помощью LLM,
- параметризовать контекстное внимание в стиле LLM.

Научные тренды: PTM, LLM и смежные с ними

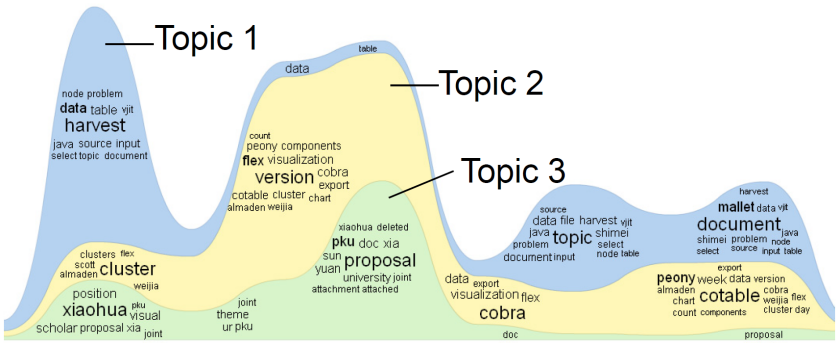
Динамика цитирования (по данным Google Scholar):
Topic Modeling и смежные области исследований:

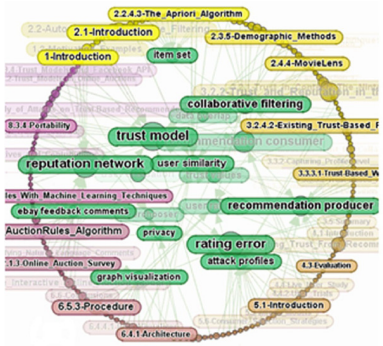


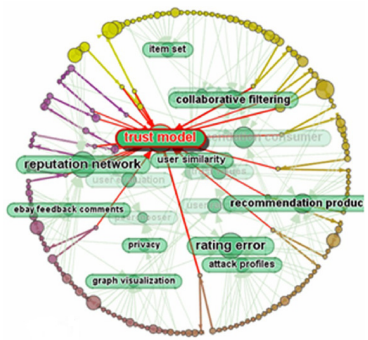
Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.
He Zhao et al. Topic Modelling Meets Deep Neural Networks: A Survey. 2021

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

- вместо «мешка слов» — последовательность $w_1, \ldots, w_n$
- вместо документов — локальные контексты слов
- **хотим:** 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.







Контекстная тематическая модель Attentive ARTM

Дано: коллекция текстовых документов, $w_1, \ldots, w_n$
$C_i \subset \{1, \ldots, n\}$ — локальный контекст (окружение) терма w_i
α_{ci} — коэффициент внимания, вес терма w_c в контексте C_i

Найти: $\phi_{tw} = p(t|w)$ — параметры тематической модели

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

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

Критерий: максимум log правдоподобия с регуляризатором R:

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

EM-алгоритм для модели Attentive ARTM

EM-алгоритм: метод простой итерации для системы уравнений

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

$$N_{tw} = \sum_{i=1}^{n} q_{wi}\frac{p_{ti}}{\theta_{ti}} \qquad q_{wi} = \sum_{c\in C_i} \alpha_{ci}[w_c = w]$$

$$\phi_{tw} = \underset{t\in T}{\text{norm}}\left(n_{tw} + \phi_{tw}N_{tw} + \phi_{tw}\frac{\partial R}{\partial \phi_{tw}}\right) \qquad n_{tw} = \sum_{i=1}^{n} p_{ti}[w_i = w]$$

Интерпретация вспомогательных переменных:

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

q_{wi} — суммарный вес всех вхождений терма w в контекст C_i

N_{tw} — суммарный вес терма w в контекстах темы t

$p(t) = \frac{n_t}{n} = \frac{1}{n}\sum_{i=1}^{n} p_{ti}$ — «массовая» доля темы t в коллекции

Регуляризирующее воздействие локального контекста

Добавка N_{tw} похожа на регуляризатор R_0 т. ч. $\frac{\partial R_0}{\partial \phi_{tw}} = N_{tw}$, можно домножить на τ, полагая по умолчанию $\tau = 0$

N_{tw} увеличивает $\phi_{tw} = p(t|w)$, если w часто встречается в локальных контекстах C_i, где центральный терм w_i относится к теме t с большой вероятностью:

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

N_{tw} похожа на дистрибутивные модели языка типа word2vec, их тематические аналоги BitermTM, WordTM, WordNetworkTM, регуляризаторы когерентности, сближающие ϕ_{tw} близких слов

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

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

Частный случай: контекст равен документу, «мешок слов»

$I_d = \left[i_d^{\text{beg}}, \ldots, i_d^{\text{end}}\right]$ — 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 $\theta_{ti} = \theta_{td}$, $q_{wi} = \frac{n_{dw}}{n_d} = \hat{p}(w|d)$ — не зависят от i,
$N_{tw} = n_w$ — не зависит от t, $\phi^*_{tw} N_{tw} = n_{tw}$

$$p_{tdw} = \underset{t\in T}{\text{norm}}\big(\phi_{tw}\theta_{td}/p(t)\big);$$

$$\phi_{tw} = \underset{t\in T}{\text{norm}}\left(2n_{tw} + \phi_{tw}\frac{\partial R}{\partial \phi_{tw}}\right);$$

$$\theta_{td} = \sum_{w\in d}\frac{n_{dw}}{n_d}\phi_{tw};$$

$$n_{tw} = \sum_{d\in D} n_{dw} p_{tdw};$$

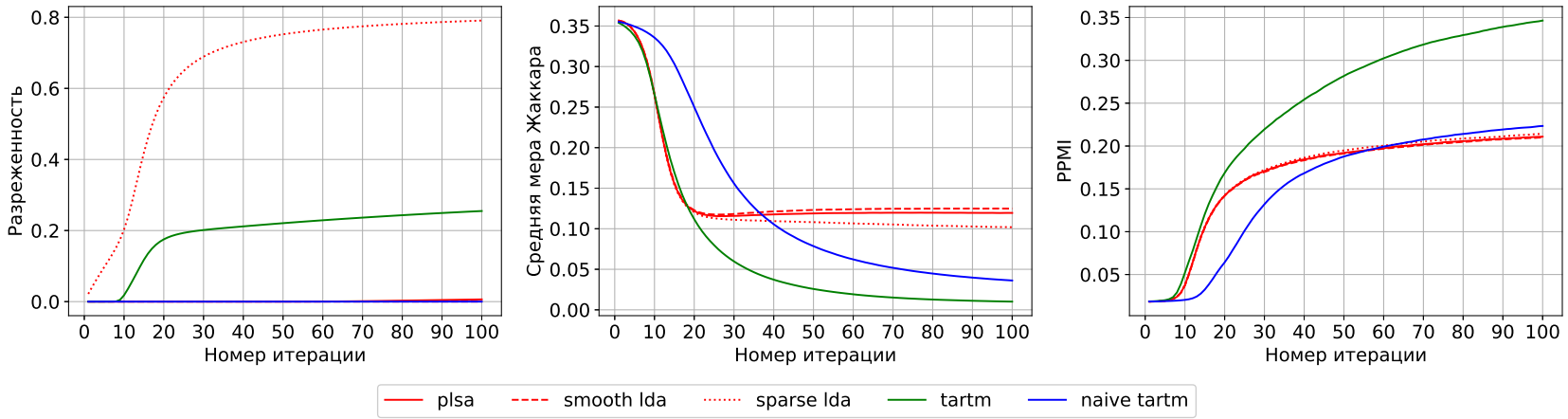
$$p(t) = \sum_{w\in W}\frac{n_{dw}}{n} p_{tdw}.$$

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

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

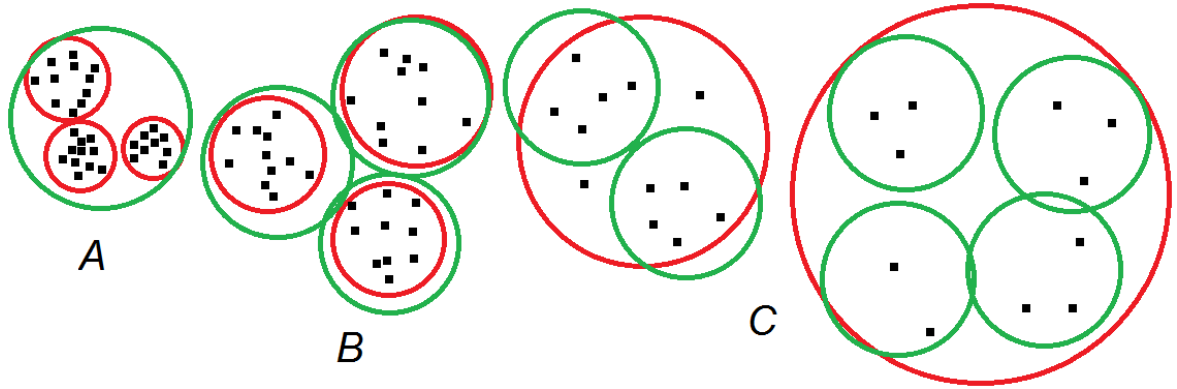
И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста, 2020.

https://github.com/ilirhin/python_artm

Проблема дефицита интерпретируемости и её связь с $p(t)$

В случаях, когда коллекция тематически не сбалансирована, появляются мусорные темы и темы-дубликаты

Причина: EM стремится равномерно распределить $p(t)$



Тема — кластер с центром $p(w|t)$ и точками $p(w|t,d)$, $d \in D$
(A) тяжёлые кластеры разбиваются на темы-дубликаты
(C) лёгкие кластеры сливаются в разнородные мусорные темы

Что делать: детектировать дубликаты и мусорные темы; после слияний и разделений поддерживать целевое $p(t)$

Как быстро вычислять взвешенные средние по контексту

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

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

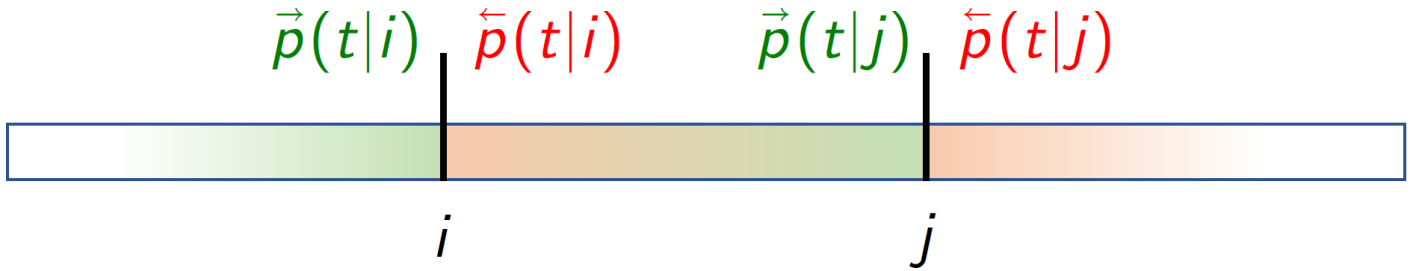
где $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — коэффициенты сглаживания в позиции i

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

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

EM-алгоритм для модели Attentive ARTM

Вход: текстовая коллекция, число тем $|T|$, параметры K, L;
Выход: матрица Φ, векторы термов документов p_{ti}, $t \in T$, $i = 1, \ldots, n$;

инициализация ϕ_{tw}; $n_t := 1$ для всех $w \in W$, $t \in T$;
для всех итераций $k = 1..K$ (проходов по всей коллекции)
  инициализация $(n_{tw}, N_{tw}, \tilde{n}_t) := 0$ для всех $w \in W$, $t \in T$;
  для всех документов $d \in D$
    $p_{ti} := \phi_{tw_i}$ для всех $t \in T$, $i \in I_d$;
    для всех $l = 1..L$ (аналог L блоков внимания в трансформере)
      $\theta_{ti} := \mathrm{Attn}\big(p_{ti} : t \in T,\ i \in I_d\big)$;
      $p_{ti} := \underset{t \in T}{\mathrm{norm}}\big(p_{ti}\theta_{ti}/n_t\big)$ для всех $t \in T$, $i \in I_d$;
    $q_{wi} := \mathrm{Attn}\big([w_i = w] : w \in d,\ i \in I_d\big)$;
    $N_{tw} := N_{tw} + q_{wi}p_{ti}/\theta_{ti}$ для всех $t \in T$, $w \in d$, $i \in I_d$;
    $n_{tw_i} := n_{tw_i} + p_{ti}$; $\tilde{n}_t := \tilde{n}_t + p_{ti}$ для всех $t \in T$, $i \in I_d$;
  $\phi_{tw} := \underset{t \in T}{\mathrm{norm}}\left(n_{tw} + \frac{n_{tw}}{n_w}N_{tw} + \frac{n_{tw}}{n_w}\frac{\partial R}{\partial \phi_{tw}}\Big|_{\phi_{tw}=\frac{n_{tw}}{n_w}}\right)$ для всех $t \in T$, $w \in W$;
  $n_t := \tilde{n}_t$ для всех $t \in T$;

$y_{hi} := \mathrm{Attn}\big(x_{hi} : h \in H,\ i \in I_d\big)$ означает $y_{hi} := \sum_{c \in C_i} \alpha_{ci} x_{hc}$

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

Входные векторы слов (эмбединги)

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

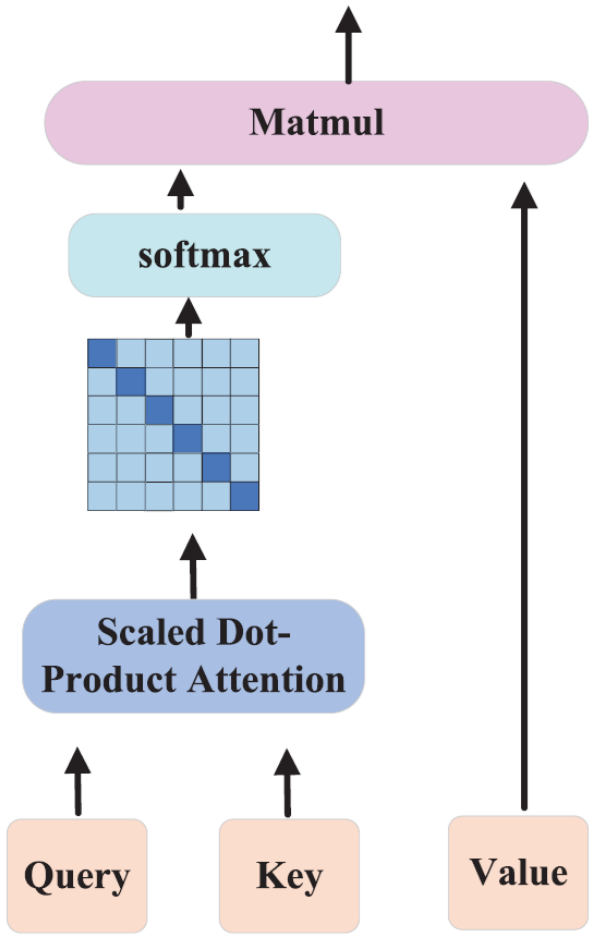
трансформируются в векторы слов, зависящие от контекстов C_i:

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

Модель внимания (self-attention):

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

$W_v x_c$ — 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 E-шага:

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

Контекстный вектор на выходе модели само-внимания:

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

Сходство:

- вектор терма w_i трансформируется в контекстный вектор
- путём усреднения векторов ϕ_{w_c} из контекста терма w_i,
- наиболее (семантически) схожих с вектором терма w_i

Отличия локализованного E-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

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:

$\phi_{tw_i} = p(t|w_i)$ — бесконтекстный вектор терма

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

$\theta_{ti} = \beta \vec{p}(t|i) + (1-\beta)\overleftarrow{p}(t|i)$ — век. 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 $p_{ti} = p(t|C_i, w_i)$ на следующем проходе берётся вместо его бесконтекстного вектора $\phi_{tw_i} = p(t|w_i)$

L итераций аналогичны L необучаемым блокам внимания

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:

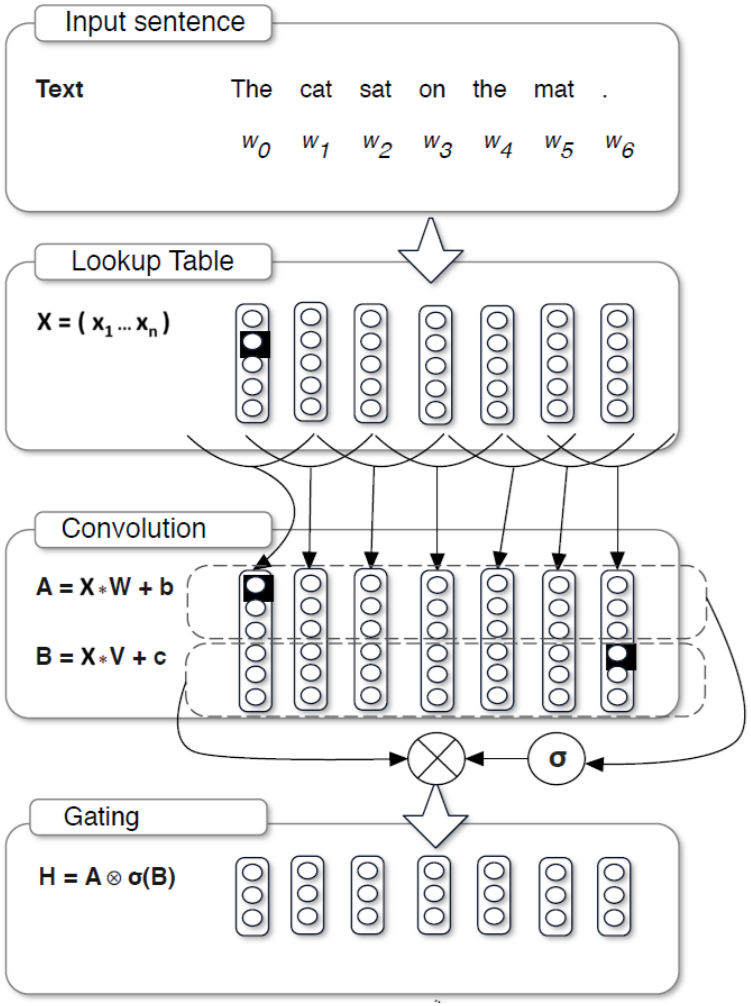
$$h_i = a_i \otimes \sigma(b_i), \quad \text{где}$$

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

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

W_c, V_c — матрицы размера $d \times T$, обучаемые параметры модели,

$\sigma(x) = \frac{1}{1+e^{-x}}$ — 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 E-шага:

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

Контекстный вектор на выходе модели GCNN:

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

Сходство:

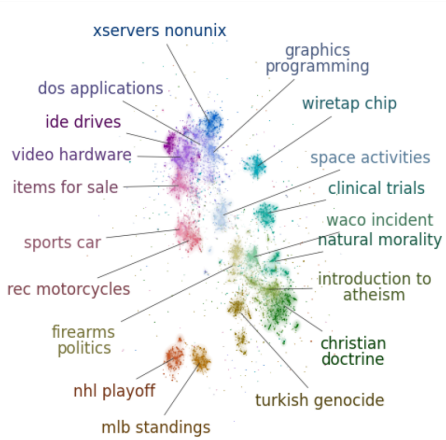
- вектор терма w_i трансформируется в контекстный вектор
- путём усреднения векторов ϕ_{w_c} его контекста,
- семантически схожих с вектором терма w_i, фильтруемых адамаровым умножением на неотрицательный вектор

Отличия локализованного E-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

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 с Attentive ARTM:

- 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.

Контекстная документная кластеризация (CDC)

n_{uw} — частота сочетания пары слов u, w в некотором окне

$p(u|w) = \frac{n_{uw}}{n_w}$ — контекст слова w

$H(w) = -\sum_u p(u|w) \log p(u|w)$ — энтропия контекста слова w

Узкий контекст — контекст с низкой энтропией, аналог темы, слова u, неслучайно часто встречающиеся рядом со словом w

Метод CDC — Contextual Document Clustering:

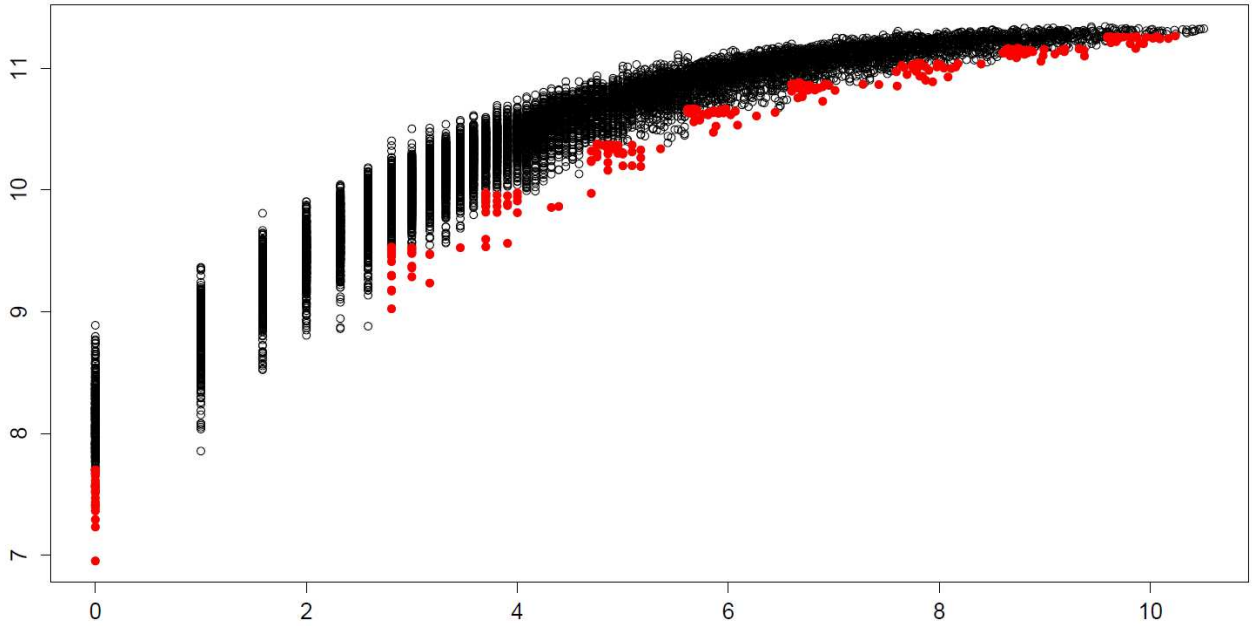
1. выделить «тематичные» слова с узкими контекстами
2. кластеризовать узкие контексты (найти кластеры-темы)
3. разбить документы на однородные сегменты (абзацы)
4. отнести каждый сегмент к ближайшей теме
5. $p(t|d)$ = доля сегментов темы t в документе

Vladimir Dobrynin, D.Patterson, N.Rooney. Contextual document clustering. 2004.
D.Patterson, N.Rooney, V.Dobrynin, M.Galushka. SOPHIA: A novel approach for textual case-based reasoning. 2005.

Выделение слов, имеющих узкие контексты

Оригинальный CDC: диапазон $\log_2 N_w$ разбивается на интервалы, в каждом интервале отбираются слова с наименьшими $H(w)$:

Зависимость энтропии $H(w)$ от документной частоты $\log_2 N_w$



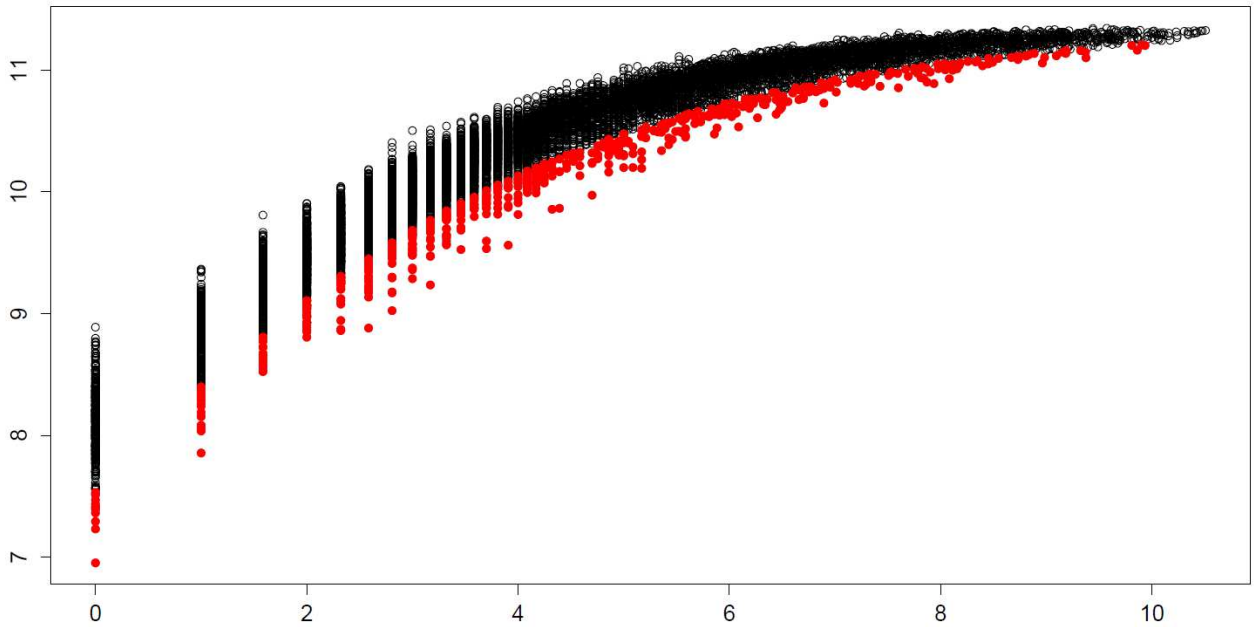
Недостаток: из-за разбиения на интервалы значительная часть узких контекстов пропускается (предвзятый отбор)

V.Dobrynin, D.Patterson, N.Rooney. Contextual document clustering. ECIR, 2004.

Выделение слов, имеющих узкие контексты

Закон Хипса $\Rightarrow$ зависимость $H(w)$ от $\log_2 N_w$ логарифмическая
Более аккуратный отбор локальных контекстов
с помощью квантильной регрессии (отсекаем 5% снизу).

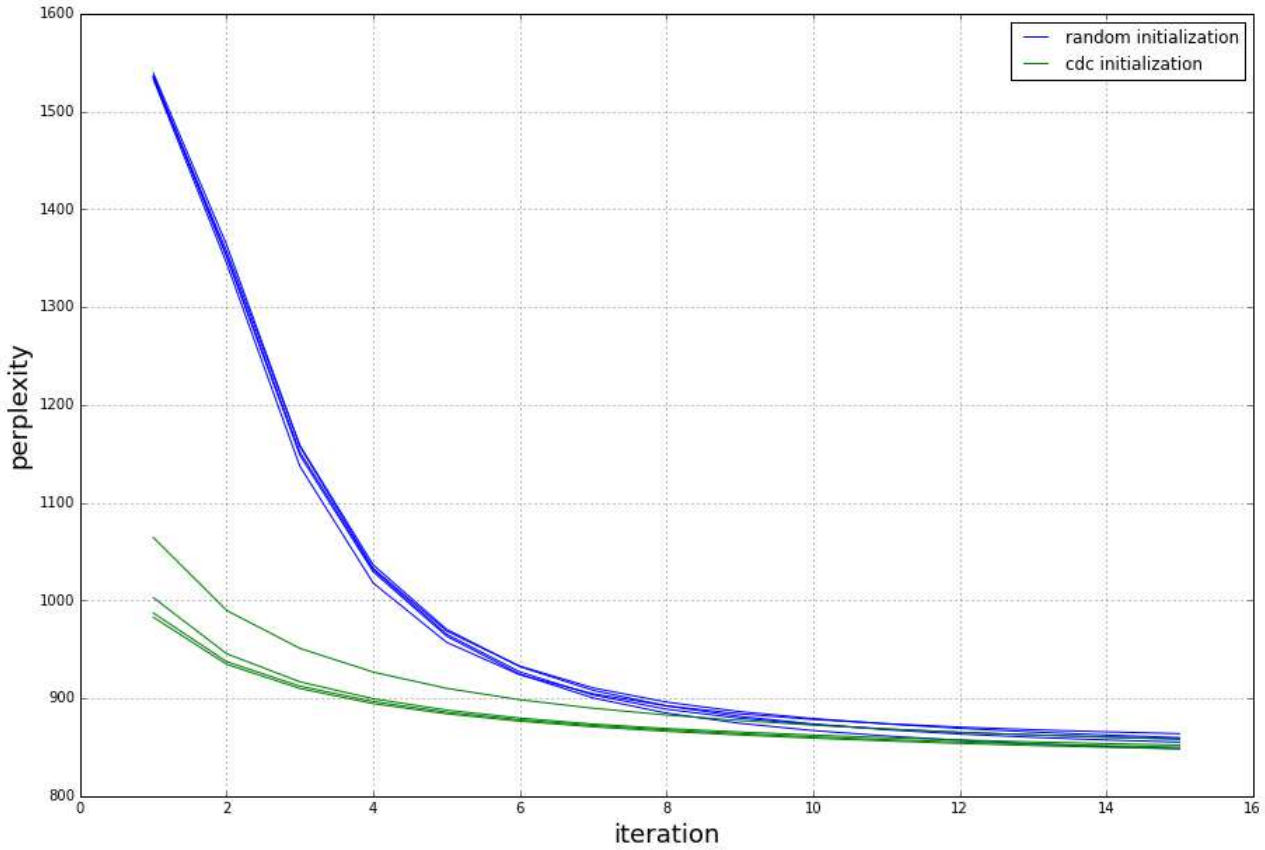
Зависимость энтропии $H(w)$ от документной частоты $\log_2 N_w$



V.Dobrynin, D.Patterson, N.Rooney. Contextual document clustering. ECIR, 2004.
Алексей Гринчук. Использование контекстной документной кластеризации для улучшения качества тематических моделей // ВКР бакалавра, МФТИ. 2015.

Инициализация тематической модели с помощью CDC

Зависимость перплексии от числа итераций (коллекция MMRO)



Алексей Гринчук. Использование контекстной документной кластеризации для улучшения качества тематических моделей // ВКР бакалавра, МФТИ. 2015.

Открытые проблемы и семейство моделей A*RTM

1. оптимизировать структуру контекста ($\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ и др.)
2. гипотеза: можно опустить трудоёмкое вычисление N_{tw}
3. гипотеза: улучшается интерпретируемость тем
4. гипотеза: управление $p(t)$ позволит балансировать темы

Семейство моделей A*RTM: Attentive, Apprehensive, Aware, Adaptive, Automated, Available, Additively Regularized TM

1. ARTM: модальности, связи, иерархии, транзакции,...
2. формирование «рассказа о себе» для каждой темы: релевантные фразы, фрагменты, название, суммаризация
3. согласование $p(t|w)$ с предобученными эмбедингами LLM
4. введение обучаемых параметров в Attentive ARTM
5. проверка налету стат-гипотез о согласии распределений
6. настройка гиперпараметров в потоке данных (AutoML)

Несколько методологических замечаний

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

Воронцов К.В., Потапенко А.А. Регуляризация, робастность и разреженность вероятностных тематических моделей. 2012.
Воронцов К.В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. 2014.
Vorontsov K. V. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization. 2023.

Книга в серии «Учебник ФПМИ МФТИ»

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

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

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

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

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

---

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