• Вероятностные языковые модели •
Лекция 4.
Тематические модели локального контекста

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

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные языковые модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • 23 марта 2026

Содержание

Напоминание. Постановка задачи ARTM

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

$p(w|d) = \sum\limits_{t \in T} p(w|\cancel{d}, t)\, p(t|d) = \sum\limits_{t \in T} \phi_{wt}\theta_{td}$ с параметрами

г.у.н.

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

Критерий: максимум регуляризованного правдоподобия

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

$$\phi_{wt} \geqslant 0; \quad \textstyle\sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \textstyle\sum_t \theta_{td} = 1$$

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

Напоминание. Основная теорема ARTM

Критерий максимума регуляризованного правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}, \quad R(\Phi,\Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta)$$

EM-алгоритм: метод простой итерации для системы уравнений

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

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

где $\underset{t\in T}{\text{norm}}(x_t) = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — операция нормирования вектора.

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

Напоминание. Рациональный EM-алгоритм 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 $i = 1, \ldots, i_{\max}$

  $n_{wt} := 0$ 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 $w \in d$

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

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

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

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

K.V.Vorontsov, A.A.Potapenko. Additive Regularization of Topic Models. MLJ 2015.

Модификация M-шага, улучшающая сходимость

В формулах M-шага вместо ϕ_{wt} и θ_{td} от предыдущей итерации можно подставлять несмещённые частотные оценки (PLSA) $\phi^*_{wt} = \frac{n_{wt}}{n_t}$ и $\theta^*_{td} = \frac{n_{td}}{n_d}$, полученные в текущей итерации:

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

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

Доказано или экспериментально установлено, что при этом

- 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.

Матричная реализация EM-алгоритма

EM-алгоритм (результат E-шага $p(t|d,w)$ встроен в M-шаг):

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\left(\phi_{wt}\sum_{d\in D} n_{dw}\frac{\theta_{td}}{(\Phi\Theta)_{wd}} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\left(\theta_{td}\sum_{w\in d} n_{dw}\frac{\phi_{wt}}{(\Phi\Theta)_{wd}} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$$

Матричная запись (norm — нормировка по столбцам):

$$\Phi := \text{norm}\big(\Phi \otimes (N \oslash \Phi\Theta)\Theta^{\mathsf{T}} + \Phi \otimes \nabla_{\Phi} R\big)$$

$$\Theta := \text{norm}\big(\Theta \otimes \Phi^{\mathsf{T}}(N \oslash \Phi\Theta) + \Theta \otimes \nabla_{\Theta} R\big)$$

где $N = (n_{dw})$ — $W \times D$-матрица исходных данных,
$\otimes$ и $\oslash$ — покомпонентное умножение и деление матриц.

Илья Ирхин. Реализация ARTM: https://github.com/ilirhin/python_artm
M.Shashanka et al. Probabilistic latent variable models as nonnegative factorizations. 2008.

Библиотека BigARTM

Ключевые возможности:

- большие данные: коллекция подгружается пакетами
- онлайновый параллельный мультимодальный EM-алгоритм
- 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++, 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)

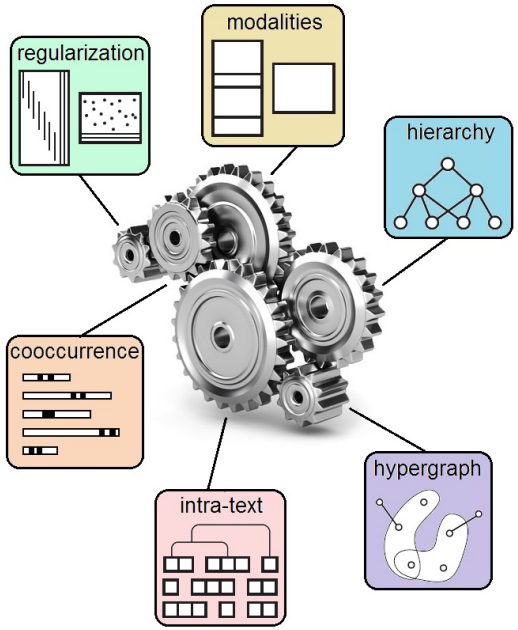
проц.	$\|T\|$	Gensim	Vowpal Wabbit	BigARTM	BigARTM асинхрон
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (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.

От BigARTM к TopicNet: ключевые возможности

BigARTM: 6 механизмов моделирования

- библиотека регуляризаторов
- мультимодальные модели
- иерархические модели
- гиперграфовые модели
- модели битермов и сети слов
- регуляризация E-шага

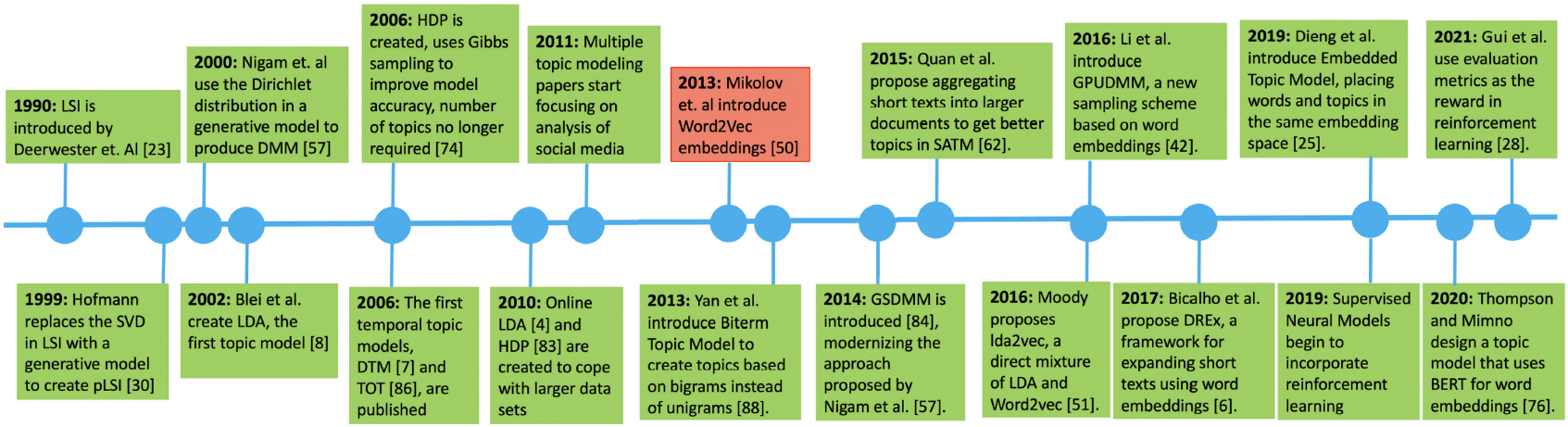


TopicNet: обёртка над BigARTM

- перебор сценариев регуляризации для выбора моделей
- автоматическое протоколирование экспериментов
- построение «банка тем» из множества моделей
- визуализация результатов тематического моделирования

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

Эволюция тематического моделирования



1999 PLSA — Probabilistic Latent Semantic Analysis
2001 LDA — Latent Dirichlet Allocation
200x мультимодальные, темпоральные, иерархические модели
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.

Нейросетевые и тематические 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* для узкого класса задач NLP/NLU
- *полнота* тематической кластерной структуры коллекции

Как «объединить лучшее от двух миров»?

Что объединяет PTM и LLM, и что их разобщает:

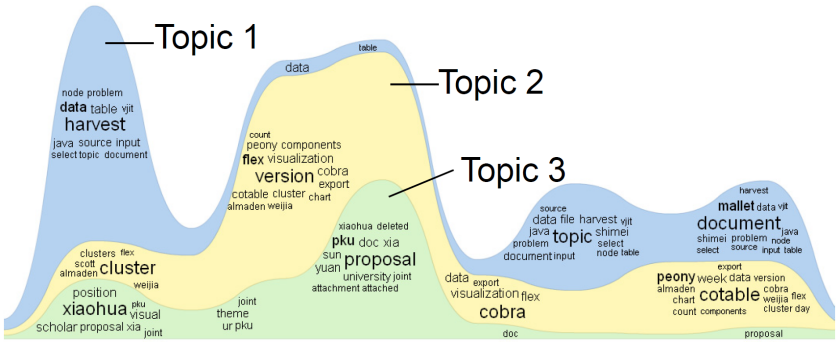
⊕ обе — вероятностные языковые модели,

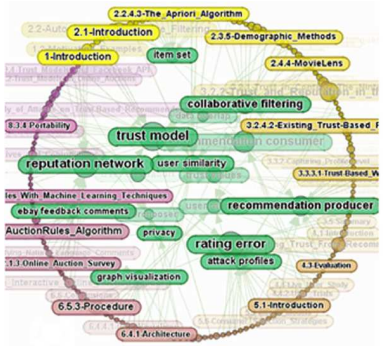
⊕ обе — автокодировщики, векторные представления текста

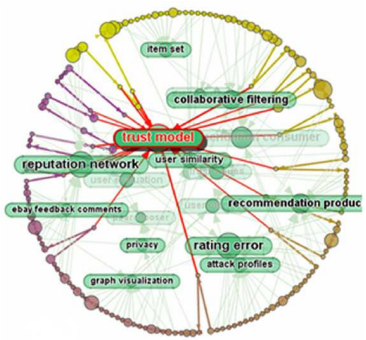
⊖ PTM: ~~мешок слов~~, архитектура матричного разложения, ~~байесовское обучение~~, трудности предобучения и др.

Мотивации: что взять от LLM в новый ARTM

- последовательность $w_1, \ldots, w_n$ вместо «мешка слов»
- локальные контексты, как в BERT, вместо документов
- быстрая тематизация любого текста или фрагмента
- быстрый поиск фрагментов, относящихся к данной теме,
- в том числе фраз для суммаризации документа или темы
- сегментация документа на тематически однородные части
- визуализация тематической структуры документа:







Идея тематизации текста за один линейный проход

Наводящие соображения:

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

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

- формула полной вероятности:

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

г.у.н.

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.

EM-алгоритм для ARTM с явным выражением Θ через Φ

Максимизация логарифма правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

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

Напоминание. Основная лемма

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

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

$$f(\Omega) \to \max_{\Omega}; \qquad \sum_{i \in I_j} \omega_{ij} = 1,\ j \in J; \qquad \omega_{ij} \geqslant 0,\ i \in I_j,\ j \in J$$

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

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

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

Численное решение системы — методом простых итераций

EM-алгоритм ARTM с объединённым M-шагом

Следствие из основной теоремы ARTM.

Точка локального экстремума (Φ, Θ) задачи

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

удовлетворяет системе уравнений

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

$$\text{M-шаг:} \quad \sum_{d,w}\sum_{t \in T} n_{dw} p_{tdw} \ln\left(\phi_{wt}\theta_{td}\right) + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

Доказательство. Применив лемму о максимизации на симплексах к

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

получим те же формулы M-шага, что и в Теореме ARTM. ■

Доказательство (по Лемме о максимизации на симплексах)

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

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

Применим Лемму к регуляризованному log-правдоподобию Q:

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

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

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

■

EM-алгоритм для ARTM с линейной тематизацией документов

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

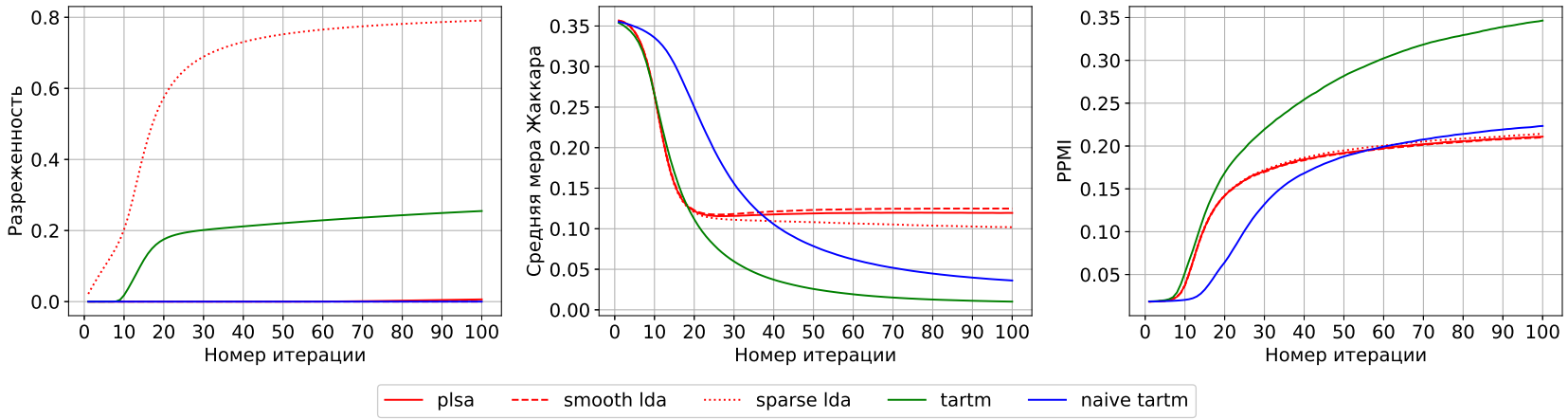
$$p'_{tdw} = p_{tdw} + \frac{\phi'_{tw}}{n_d} \left(\frac{n_{td}}{\theta_{td}} - \sum_{s \in T} \phi'_{sw} \frac{n_{sd}}{\theta_{sd}} \right)$$

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

Эксперимент. Проверка модифицированного EM-алгоритма

Коллекция NIPS, $|T| = 50$, модели:

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



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

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

https://github.com/ilirhin/python_artm

Упрощение EM-алгоритма для линейной тематизации

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

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

Контекстная тематическая модель

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 из C_i для w_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:

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

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

EM-алгоритм для контекстной тематической модели

EM-алгоритм: метод простой итерации для системы уравнений

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

$$p(t) = \sum_{w\in W}\phi_{tw}p(w) \qquad n_t = \sum_{i=1}^{n} p_{ti}$$

$$\theta_{ti} = \sum_{c\in C_i}\alpha_{ci}\phi_{tw_c} \qquad n_{tw} = \sum_{i=1}^{n} p_{ti}[w_i = w]$$

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

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

Какова сложность вычислений? Придётся ли что-то упрощать?

Доказательство (по лемме о максимизации на симплексах)

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

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

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

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

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

$$\phi_{tw}\frac{\partial Q}{\partial \phi_{tw}} = \sum_{i=1}^{n} \frac{p(w_i)}{p(w_i|C_i)} \phi_{tw} \sum_{s\in T} \frac{\partial}{\partial \phi_{tw}}\left(\frac{\phi_{sw_i}\theta_{si}}{p(s)}\right) + \phi_{tw}\frac{\partial R}{\partial \phi_{tw}} =$$

$$= \sum_{i=1}^{n} \frac{p(w_i)}{p(w_i|C_i)} \phi_{tw}\left(\frac{\theta_{ti}[w_i = w]}{p(t)} + \frac{\phi_{tw_i} q_{wi}}{p(t)} - \frac{\phi_{tw_i}\theta_{ti} p(w)}{p^2(t)}\right) + \phi_{tw}\frac{\partial R}{\partial \phi_{tw}} =$$

$$= \sum_{i=1}^{n} \underbrace{\frac{\phi_{tw_i}\theta_{ti} p(w_i)}{p(w_i|C_i)\, p(t)}}_{= p_{ti}} \left([w_i = w] + \phi_{tw}\frac{q_{wi}}{\theta_{ti}} - \phi_{tw}\frac{p(w)}{p(t)}\right) + \phi_{tw}\frac{\partial R}{\partial \phi_{tw}} =$$

$$= \underbrace{\sum_{i=1}^{n} p_{ti}[w_i = w]}_{n_{tw}} + \phi_{tw}\underbrace{\sum_{i=1}^{n} q_{wi}\frac{p_{ti}}{\theta_{ti}}}_{N_{tw}} - \phi_{tw}\underbrace{\sum_{i=1}^{n} p_{ti}}_{n_t}\frac{p(w)}{p(t)} + \phi_{tw}\frac{\partial R}{\partial \phi_{tw}}$$

■

Осмысление контекстного EM-алгоритма

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, $p(t) \neq \frac{n_t}{n}$

$q_{wi} = \sum_{c \in C_i} \alpha_{ci} [w_c = w]$ — весовая доля терма w в контексте C_i

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

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

При подстановке несмещённой оценки $\phi^*_{tw} = \frac{n_{tw}}{n_w}$ и $p(t) \approx \frac{n_t}{n}$

$$n_{tw} - \phi^*_{tw} n_t \frac{p(w)}{p(t)} \approx n_{tw} - \frac{n_{tw}}{n_w} n_t \frac{n_w}{n_t} \approx 0$$

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

Регуляризирующее воздействие локального контекста?

N_{tw} выглядит как регуляризатор R_0 такой что $\frac{\partial R_0}{\partial \phi_{tw}} = N_{tw}$; можно ли его домножать на τ? полагать $\tau = 0$?

N_{tw} увеличивает ϕ_{tw}, если терм w частый 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 терма w в контексте C_i

N_{tw} похоже 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 $k = |C_i|$ — 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»

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

$$p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{tw}\theta_{td}/p(t)\right) \qquad \theta_{td} = \sum_{w\in d} \frac{n_{dw}}{n_d}\phi_{tw}$$

$$\phi_{tw} = \underset{t\in T}{\text{norm}}\left(n_{tw} + \phi_{tw}\frac{\partial R}{\partial \phi_{tw}}\right) \qquad n_{tw} = \sum_{d\in D} n_{dw} p_{tdw}$$

$$p(t) = \sum_{w\in W} \phi_{tw} p(w)$$

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

Контекстная тематическая модель — Attentive ARTM

E-шаг — аналог self-attention в контекстной модели:

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

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

  ϕ_{tw_c} — вектор-значение (value);

  α_{ci} — коэффициент внимания;

  в роли ключа (key) — покомпонентное умножение векторов

Ещё одно наблюдение: что если $p(t)$ — внешний параметр?

- упростится (см. док-во) формула M-шага, можно ввести τ:

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

- можно обновлять на итерациях $p(t) := \sum_w \phi_{tw} p(w)$
- будет ли такая модель консистентна? будет ли сходимость?

EM-алгоритм для модели Attentive ARTM

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

инициализация ϕ_{tw}; $p(t) := 1/|T|$ для всех $w \in W$, $t \in T$;
для всех итераций $k = 1..K$ (проходов по всей коллекции)
  инициализация $(n_{tw}, N_{tw}) := 0$ 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)
      $\theta_{ti} := \mathrm{Attn}\big(p_{ti} \colon t \in T,\ i \in I_d\big)$;
      $p_{ti} := \underset{t \in T}{\mathrm{norm}}\big(p_{ti}\theta_{ti}/p(t)\big)$ для всех $t \in T$, $i \in I_d$;
    $q_{wi} := \mathrm{Attn}\big([w_i = w] \colon w \in d,\ i \in I_d\big)$;
    $N_{tw} := N_{tw} + p_{ti}q_{wi}/\theta_{ti}$ для всех $t \in T$, $w \in d$, $i \in I_d$;
    $n_{tw_i} := n_{tw_i} + p_{ti}$; для всех $t \in T$, $i \in I_d$;
  $\phi_{tw} := \underset{t \in T}{\mathrm{norm}}\left(n_{tw} + \tau \frac{n_{tw}}{n_w} N_{tw} + \frac{n_{tw}}{n_w} \frac{\partial R}{\partial \phi_{tw}} \Big|_{\phi_{tw} = \frac{n_{tw}}{n_w}}\right)$ для всех $t \in T$, $w \in W$;
  $p(t) := \sum_w \phi_{tw} p(w)$ для всех $t \in T$;

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

Как быстро вычислять взвешенные средние по контексту

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

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

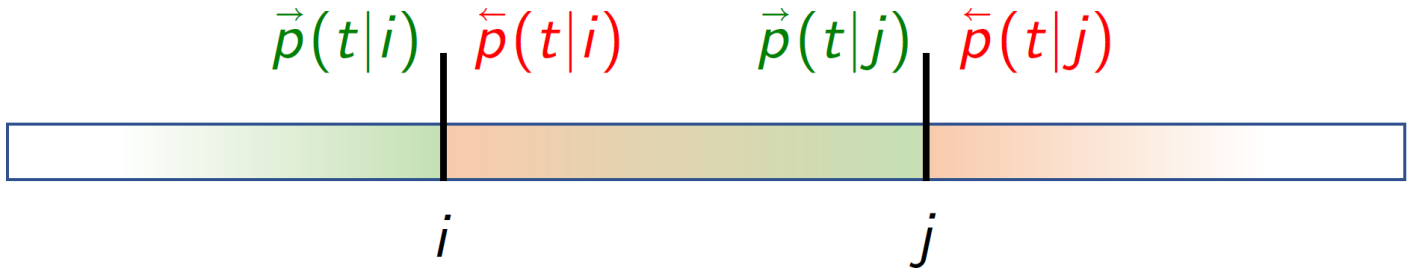
где $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — коэффициенты сглаживания в позиции i

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

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

Входные векторы слов (эмбединги)

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

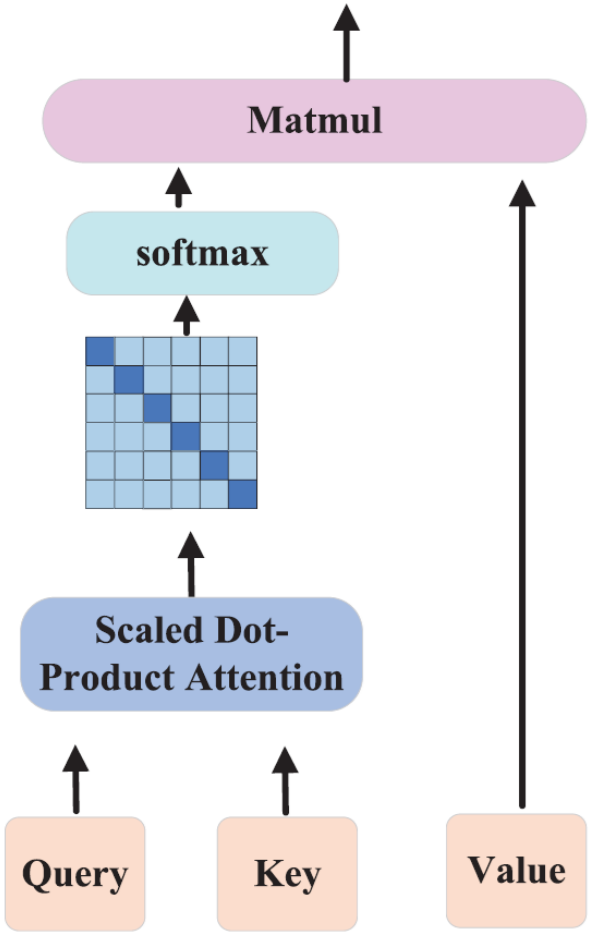
трансформируются в векторы слов, зависящие от контекстов C_i:

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

Модель внимания (self-attention):

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

$W_v x_c$ — 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 термов w_c из его контекста,
- наиболее схожих с ним по тематике

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:

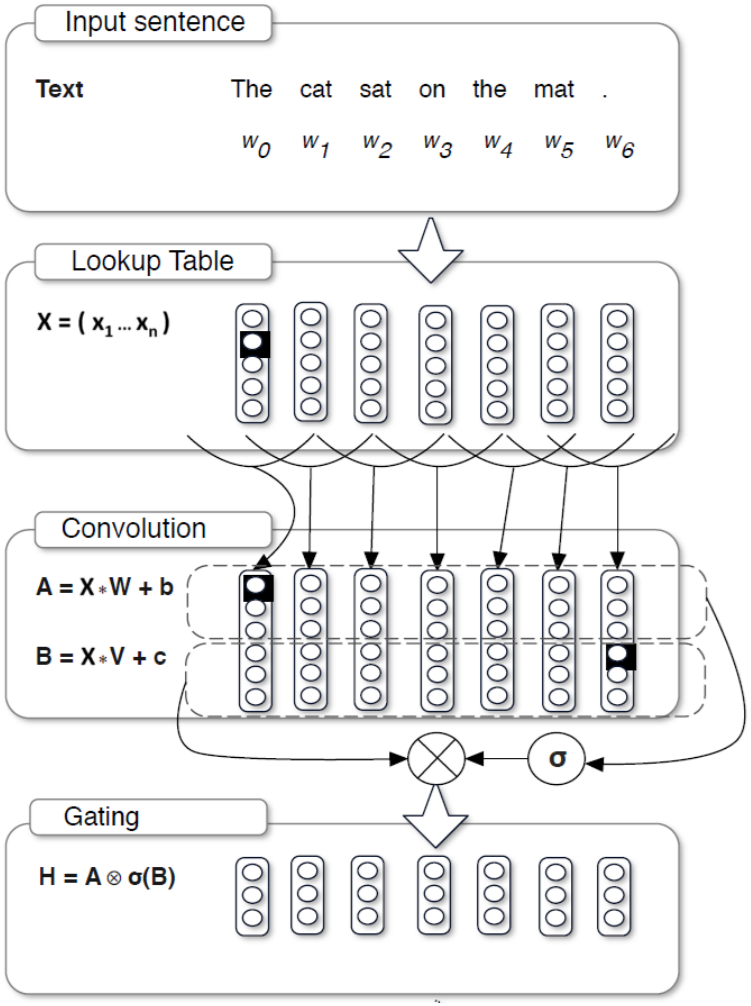
$$h_i = a_i \otimes \sigma(b_i), \quad \text{где}$$

$a_i = \sum_{c \in C_i} W_c x_c$ — 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:

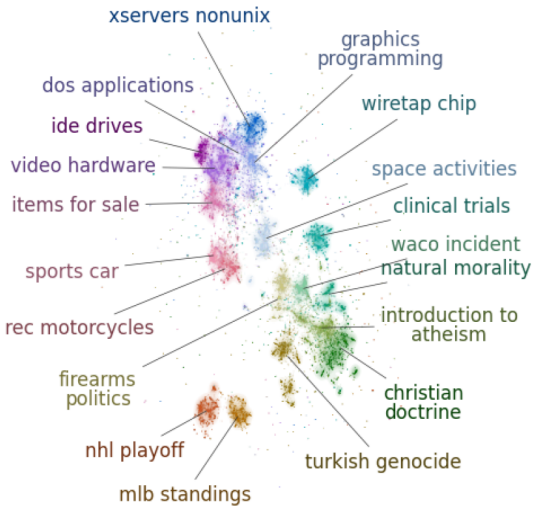
- вектор терма w_i трансформируется в контекстный вектор
- путём усреднения векторов ϕ_{w_c} его контекста,
- семантически схожих с вектором терма w_i, фильтруемых адамаровым умножением на неотрицательный вектор

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 — 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. 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

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 и встроить в ARTM:

- модель внимания, локальные контексты вместо документов
- отбор релевантных фраз и n-грамм по каждой теме
- именование и суммаризация тем на основе этих фраз
- инициализация тем по предобученным эмбедингам BERT, чтобы обеспечить качество тем даже на малых коллекциях
- автоматическое определение числа тем
- разбиение документа на монотематичные сегменты

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). Старьё?

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:

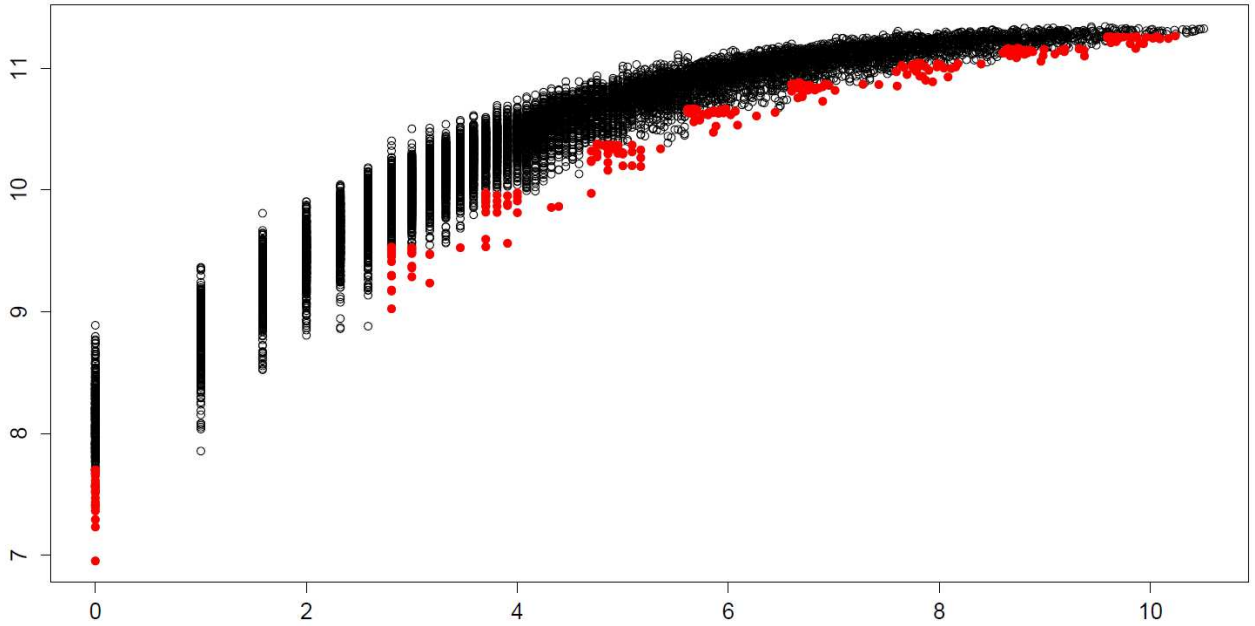
1. выделить «тематичные» слова с узкими контекстами
2. кластеризовать узкие контексты (найти кластеры-темы)
3. разбить документы на однородные сегменты (абзацы)
4. отнести каждый сегмент к ближайшей теме
5. $p(t|d)$ = доля сегментов темы t в документе

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 $\log_2 N_w$



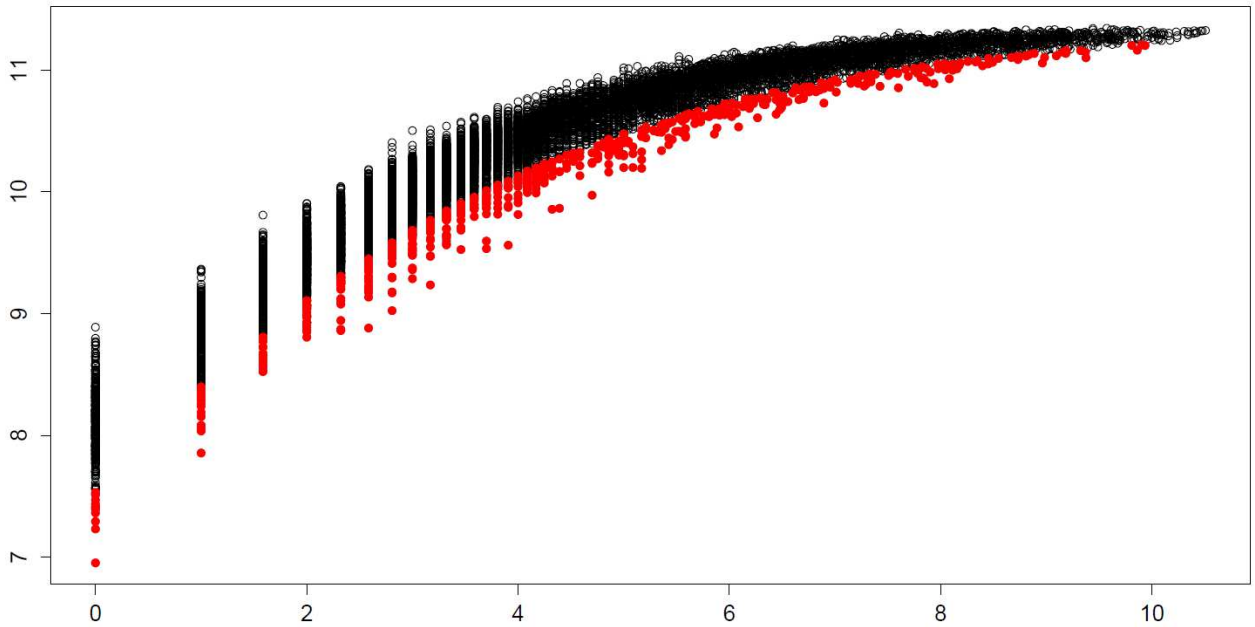
Недостаток: из-за разбиения на интервалы значительная часть узких контекстов пропускается (предвзятый отбор)

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.

Резюме

Цель и путь — «Make Topic Modeling Great Again», а именно, создать новый стандарт тематического моделирования:

1. от байесовского обучения — к аддитивной регуляризации
2. от мешка слов — к локальным контекстам
3. от BigARTM — к новой эффективной библиотеке ARTM

Что хорошего сохранить от BigARTM:

1. гибкость: регуляризации, модальности, иерархии, транзакции
2. технологичность: батчи, параллельность, скорость, метрики

Что нового привнести в ARTM от Attention и LLM:

1. параметризация модели внимания
2. каждая тема должна уметь «рассказать о себе»
3. гарантированная интерпретируемость тем
4. даже в случае тематически несбалансированной коллекции

Задания по курсу

Задача-минимум: научиться решать задачи анализа текстов с использованием тематического моделирования

Задача-максимум: получить новый научный результат

виды деятельности	оценка
теоретическая задача	X
теоретическая задача*	$2X$
теоретическая задача**	$3X$
решение прикладной задачи	$10X$
обзор по последним PTM/NTM	$10X$
участие в проекте	$20X$
работа над открытой проблемой	$25X$

где X — оценка за вид деятельности по 5-балльной шкале.
score — суммарная оценка по всем видам деятельности.

Итоговая оценка: $\min\left(5, \lfloor \text{score}/20 \rfloor\right)$ по 5-балльной шкале.

Задания к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Биграммная модель коллекции: $p(w|v) = \xi_{wv}$,
где v — слово, идущее в тексте перед w.
Найти параметры модели ξ_{wv}.

2. Биграммная модель документов: $p(w|v,d) = \xi_{dwv}$.
Найти параметры модели ξ_{dwv}.

Подсказка: применить условия ККТ или основную лемму.

3*. Творческое задание (возможны разные решения).
Предложите модель, разделяющую роли слов в текстах:
— тематические слова
— специфичные слова документа (шум)
— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т}, \text{ш}, \text{ф}\}$.
Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.
Подсказка 3: можно использовать документную частоту слов.

Задания к лекции 2

4. Пользуясь основной леммой, докажите, что регуляризатор битермов эквивалентен добавлению псевдодокументов d_u в исходную коллекцию (см. слайд 13)

Прикладная исследовательская задача:
автоматическое выделение научных терминов (ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  метод ATE на основе комбинирования ARTM и TopMine;
  обоснование, что синтаксический анализ не нужен;
  зависимость качества ATE от объёма коллекции
- Критерий:
  качество ATE (Prec, Rec, F1) на размеченных данных

Теоретическое задание к лекции 3

Выведете EM-алгоритм для тематической языковой модели:

5. $p(w|d) = \sum_t \phi_{wt}\theta_{td}$, используя в качестве исходных данных последовательность $(d_i, w_i)_{i=1}^n$ вместо счётчиков n_{dw}.
Докажите эквивалентность обычному EM-алгоритму ARTM.

6. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$,
где $\phi_{tw} = p(t|w)$, $\theta_{td} = p(t|d)$ — параметры модели.

7. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$,
где $\phi_{tw} = p(t|w)$ — параметры модели, $\theta_{td} = \sum_w \frac{n_{dw}}{n_d}\phi_{tw}$.

8*. Введение $p(t)$ как внешнего параметра упрощает выкладки, но может нарушать условия целостности модели:

$$p(t) = \sum_t \phi_{tw}p(w), \qquad p(t) = \sum_t \theta_{td}p(d).$$

Как обеспечить выполнение этих условий в EM-алгоритме?

9. Докажите, что необходимым условием максимума

$$\sum_{i=1}^{n} \ln \sum_{t \in T} p(w_i, t|i, \Omega) \;\to\; \max_{\Omega}$$

для языковой модели со скрытыми переменными $t \in T$ (не обязательно темами) и параметрами $\Omega = (\omega_{kj})$ — набором неотрицательных нормированных векторов, является система

$$\text{E-шаг:} \quad p(t|w_i, i) = \operatorname*{norm}_{t \in T} p(w_i, t|i, \Omega)$$

$$\text{M-шаг:} \quad \omega_{kj} = \operatorname*{norm}_{k} \left(\sum_{i=1}^{n} \sum_{t \in T} p(t|w_i, i)\, \omega_{kj} \frac{\partial}{\partial \omega_{kj}} \ln p(w_i, t|i, \Omega) \right)$$

Выведете EM-алгоритм для известных вам частных случаев $p(w, t|i, \Omega) = \phi_{wt}\theta_{td}$ и $p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_c \alpha_{ci} \phi_{tw_c}$.

10. Творческое задание (возможны разные решения).** Предложите «какую-нибудь разумную» параметризацию для тематической модели внимания. Используя «основную лемму», получите уравнения для новых параметров модели.

Исследовательская задача к лекции 4

Реализуйте EM-алгоритм для модели локального контекста (или воспользуйтесь чужой готовой реализацией)

Исследуйте зависимость метрик качества модели

- перплексия
- целостность: выполнение тождеств
$$p(t) = \sum_t \phi_{tw} p(w), \qquad p(t) = \sum_t \theta_{td} p(d)$$
- разреженность, различность, когерентность тем

от номера итерации и от параметров модели:

- $|T|$ — число тем
- L — число проходов
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, β — баланс левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, глав
- опция «исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}$, $\overleftarrow{\theta}_{ti}$»

Примеры датасетов для практических заданий по курсу

1. Открытые датасеты (английский): 20NG, NIPS, KOS
2. Ранжированные результаты поиска научных статей (по данным eLibrary, arXiv, PubMed)
3. Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
4. Техноблоги: Хабр (русский), TechCrunch (английский)
5. Данные социальных сетей: VK, Twitter, Telegram,...
6. Статьи по Complexity Sciences (для хронокарты науки)

- Википедия
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

---

http://bigartm.org
http://drive.google.com/drive/folders/1PPnw6aZOJAJoLRYuwdGm437RssV-XQx0

Проекты

- «Тематизатор» для социо-гуманитарных исследований:
  — пользователь задаёт грубый фильтр текстового потока;
  — задача: «классифицировать иголки в стоге сена»,
  — разделив темы на информативные и мусорные,
  — выделив аспекты и тональности в каждой теме;
  — конечная цель: кол./кач. анализ предметной области,
  — реализация данного сценария как модуля в среде Orange

- «Мастерская знаний» для научного поиска:
  — пользователь строит тематические подборки статей,
  — поисковая выдача формируется моделью SciRus;
  — задача: показать пользователю тематику подборки;
  — понадобится: автоматическое выделение терминов,
  — выделение тематических фраз из документов,
  — автоматическое именование и суммаризация тем;
  — конечная цель: помочь в понимании предметной области

Открытые проблемы тематического моделирования

1. Тематические модели внимания последовательного текста
2. Проблема несбалансированности тем в коллекции
3. Измерение интерпретируемости тем (когерентность)
4. Обеспечение 100%-й интерпретируемости тем
5. Автоматическое именование и суммаризация тем
6. Калибровка моделей тематической фильтрации
7. Согласование тем с предобученными эбмедингами LLM
8. Статистические оценки состоятельности тем
9. Обнаружение новых тем или трендов в потоке текстов
10. Обеспечение устойчивости и полноты множества тем
11. Автоматический подбор гиперпараметров, AutoML
12. Гиперграфовые тематические модели для RecSys