

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

Константин Вячеславович Воронцов
k.vorontsov@iaai.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

1 Тематическая модель «мешка слов»

- EM-алгоритм
- Подходы к ускорению EM-алгоритма
- Библиотека BigARTM

2 Тематическая модель локальных контекстов

- Быстрая однопроходная тематизация текста
- Тематическая модель локального контекста
- Быстрое вычисление векторов контекста

3 Сравнение с другими контекстными моделями

- Модель внимания и трансформер
- Свёрточная сеть GCNN
- Контекстные модели Contextual-Top2Vec и CDC

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

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

- $\phi_{wt} = p(w|t)$ — 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

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

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

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

Напоминание. Основная теорема ARTM

Критерий максимума регуляризованного правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}, \quad R(\Phi, \Theta) = \sum_k \tau_k R_k(\Phi, \Theta)$$

EM-алгоритм: метод простой итерации для системы уравнений

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

где $\mathop{\text{norm}}_{t \in T}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$ — 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 $i = 1, \dots, i_{\max}$

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

для всех документов $d \in D$

$n_{td} := 0$ для всех $t \in T$;

для всех термов $w \in d$

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

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

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

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

Модификация 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:

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

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

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):

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

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

Матричная запись (norm — нормировка по столбцам):

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

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

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
- онлайн-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

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

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

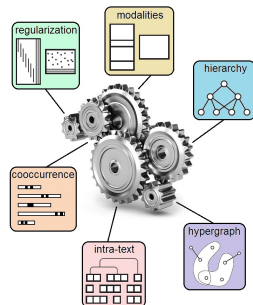
D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov.

Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

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

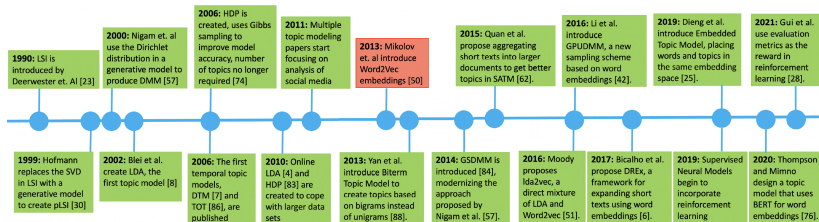


TopicNet: обёртка над BigARTM

- 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.**

Идея тематизации текста за один линейный проход

Наводящие соображения:

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

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

- формула полной вероятности:

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

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:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

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:

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

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω . Если ω_j — вектор локального экстремума задачи $f(\Omega) \rightarrow \max$ 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

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

удовлетворяет системе уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} \equiv p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \sum_{d,w} \sum_{t \in T} n_{dw} p_{tdw} \ln(\phi_{wt} \theta_{td}) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta} \end{cases}$$

Доказательство. Применив лемму о максимизации на симплексах к

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

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 $\Theta(\Phi)$:

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

Применим Лемму к регуляризованному log-правдоподобию Q :

$$\begin{aligned} \phi_{wt} \frac{\partial Q}{\partial \phi_{wt}} &= \sum_{d \in D} n_{dw} p_{tdw} + \sum_{d,s,u} n_{du} p_{sdu} \frac{\phi_{wt}}{\theta_{sd}} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} + \phi_{wt} \sum_{d,s} \frac{\partial R}{\partial \theta_{sd}} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} = \\ &= \sum_{d \in D} n_{dw} \left(p_{tdw} + \frac{1}{n_{dw}} \sum_{s \in T} \frac{\phi_{wt}}{\theta_{sd}} \underbrace{\left(\sum_{u \in d} n_{du} p_{sdu} + \theta_{sd} \frac{\partial R}{\partial \theta_{sd}} \right)}_{n_{sd}} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} \right) + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} = \\ &= \sum_{d \in D} n_{dw} \underbrace{\left(p_{tdw} + \frac{1}{n_{dw}} \sum_{s \in T} \frac{n_{sd}}{\theta_{sd}} \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} \right)}_{p'_{tdw}} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}. \quad \blacksquare \end{aligned}$$

EM-алгоритм для ARTM с линейной тематизацией документов

$$\theta_{td}(\Phi) = \sum_{w \in D} \frac{n_{dw}}{n_d} \operatorname{norm}_{t \in T}(\phi_{wt} p_t) \Rightarrow \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} = \frac{n_{dw}}{n_d} \tilde{\phi}_{tw} (\delta_{st} - \tilde{\phi}_{sw})$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\tilde{\phi}_{tw} \equiv p(t|w) = \operatorname{norm}_{t \in T}(\phi_{wt} n_t); \quad \theta_{td} = \sum_{w \in D} \frac{n_{dw}}{n_d} \tilde{\phi}_{tw}$$

$$p_{tdw} \equiv p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}); \quad n_t = \sum_{d \in D} \sum_{w \in D} n_{dw} p_{tdw}$$

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

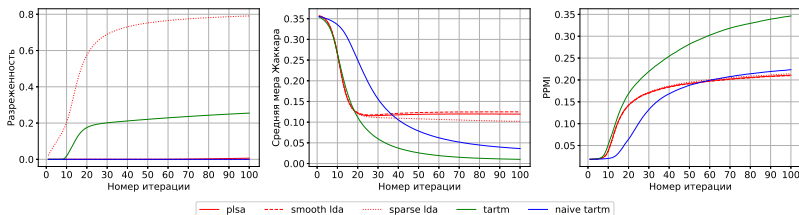
$$p'_{tdw} = p_{tdw} + \frac{\tilde{\phi}_{tw}}{n_d} \left(\frac{n_{td}}{\theta_{td}} - \sum_{s \in T} \tilde{\phi}_{sw} \frac{n_{sd}}{\theta_{sd}} \right)$$

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

Эксперимент. Проверка модифицированного EM-алгоритма

Коллекция NIPS, $|T| = 50$, модели:

- TARTM (Θ less 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

$$\begin{aligned} \tilde{\phi}_{tw} &= \operatorname{norm}_{t \in T}(\phi_{wt} n_t); & \theta_{td} &= \sum_{w \in D} \frac{n_{dw}}{n_d} \tilde{\phi}_{tw}; \\ p_{tdw} &= \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}); & n_t &= \sum_{d \in D} \sum_{w \in D} n_{dw} p_{tdw}; \\ \phi_{wt} &= \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right). \end{aligned}$$

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

Контекстная тематическая модель

Дано: коллекция текстовых документов, w_1, \dots, w_n
 $C_i \subset \{1, \dots, n\}$ — локальный контекст (окружение) термина w_i
 α_{ci} — коэффициент внимания, вес термина w_c из C_i для w_i

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

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

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

Критерий: максимум \log правдоподобия с регуляризатором R :

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

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

$$p_{ti} \equiv p(t|C_i, w_i) = \operatorname{norm}_{t \in T} (\phi_{tw_i} \theta_{ti} / p(t))$$

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

$$n_t = \sum_{i=1}^n p_{ti}$$

$$\theta_{ti} = \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c}$$

$$n_{tw} = \sum_{i=1}^n p_{ti} [w_i = w]$$

$$q_{wi} = \sum_{c \in C_i} \alpha_{ci} [w_c = w]$$

$$N_{tw} = \sum_{i=1}^n q_{wi} \frac{p_{ti}}{\theta_{ti}}$$

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

Какова сложность вычислений? Придётся ли что-то упрощать?

Доказательство (по лемме о максимизации на симплексах)

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

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

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

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

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

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

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

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

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

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

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} 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 термина w в контексте C_i

N_{tw} выглядит как регуляризатор R_0 такой что $\frac{\partial R_0}{\partial \phi_{tw}} = N_{tw}$

N_{tw} похож на регуляризаторы когерентности, BitermTM, WNTM, которые также сближают ϕ_{tw} слов, встречающихся рядом; возможно, он взаимозаменяем с ними или даже не обязателен

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

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

Возможно ли уйти от трудоёмкого вычисления N_{tw} ?

Частный случай: контекст равен документу, «мешок слов»

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

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

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

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

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

EM-алгоритм с линейной тематизацией документов

$$p_{tdw} = \operatorname{norm}_{t \in T} (\phi_{tw} \theta_{td} / p(t))$$

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

$$\phi_{tw} = \operatorname{norm}_{t \in T} \left(n_{tw} + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} \right)$$

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

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

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

Упрощение M-шага

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

Третий член равен первому, если контекст равен документу:

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

Гипотеза: второй и третий члены можно отбросить

EM-алгоритм с линейной тематизацией локальных контекстов

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

Контекстная тематическая модель — Attentive ARTM

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

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

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

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

ϕ_{tw_c} — вектор-значение (value),

α_{ci} — коэффициент внимания,

в роли ключа (key) — покомпонентное умножение векторов

Е-шаг, L проходов документа — аналог Transformer:

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

Гипотеза: это аналогично L необучаемым блокам внимания

EM-алгоритм для модели Attentive ARTM

вход: текстовая коллекция, число тем $|T|$, параметры K, L, τ ;

выход: матрица Φ , векторы термов документов $p_{ti}, t \in T, i = 1, \dots, n$;

инициализация $\phi_{tw}; p(t) := 1/|T|$ 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)

$\theta_{ti} := \text{Attn}((p_{ti})_{t \in T}: i \in I_d)$;

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

$q_{wi} := \text{Attn}([w_i = w]_{w \in d}: i \in I_d)$;

$N_{tw} := N_{tw} + p_{ti}q_{wi}/\theta_{ti}$ для всех $t \in T, w \in d, i \in I_d$;

$n_{tw_i} := n_{tw_i} + p_{ti}$; для всех $t \in T, i \in I_d$;

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

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

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

Как быстро вычислять взвешенные средние по контексту

за линейное время вместо квадратичного по длине контекста

функция Attn: *двунаправленное экспоненциальное скользящее среднее:*

вход: $(x_i \in \mathbb{R}^H : i \in I_d)$ — последовательность n_d векторов;

выход: $(y_i \in \mathbb{R}^H : i \in I_d)$ — последовательность n_d векторов;

$$\vec{y}_i := \vec{\gamma}_i x_i + (1 - \vec{\gamma}_i) \vec{y}_{i-1} \quad \text{для всех } i = 1 \dots n, \quad \vec{\gamma}_1 = 1;$$

$$\tilde{y}_i := \tilde{\gamma}_i x_i + (1 - \tilde{\gamma}_i) \tilde{y}_{i+1} \quad \text{для всех } i = 1 \dots n, \quad \tilde{\gamma}_n = 1;$$

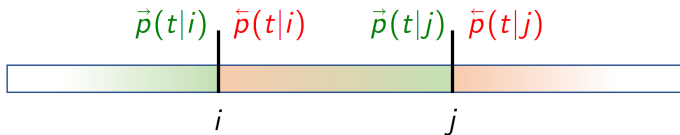
$$y_i := \beta \vec{y}_i + (1 - \beta) \tilde{y}_i \quad \text{для всех } i = 1 \dots n;$$

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

Как задать $\vec{\gamma}_i, \tilde{\gamma}_i$ — коэффициенты сглаживания в позиции i :

- $\gamma_i \approx \frac{1}{h}$, где h — ширина окна, размер контекста
- $\gamma_i = 1$, если надо забыть контекст, сменить документ
- $\gamma_i = 0$, если надо проигнорировать терм w_i
- γ_i можно умножать на оценку важности термина w_i

Использование двунаправленных векторов контекста



Через *двунаправленные тематические векторы* определяется:

- $\vec{p}_{ti} = \vec{p}(t|i)$ — тематика левого контекста термина w_i
- $\vec{p}_{ti} = \vec{p}(t|i)$ — тематика правого контекста термина w_i
- $p_{ti} = \frac{1}{2}(\vec{p}_{ti} + \vec{p}_{ti})$ — тематика двустороннего контекста w_i
- $p(t|i \dots j) = \frac{1}{2}(\vec{p}_{ti} + \vec{p}_{tj})$ — тематика сегмента $[i \dots j]$
- $\vec{p}_{ti} \approx \vec{p}_{tj}$ — однородность тематики сегмента $[i \dots j]$
- $\max_i \|\vec{p}_{ti} - \vec{p}_{ti}\|$ — 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)

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

трансформируются в векторы слов,
зависящие от контекстов C_i :

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

Модель внимания (self-attention):

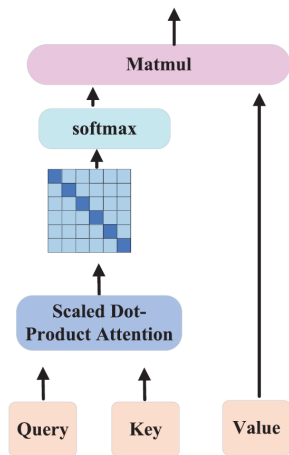
$$h_i = \sum_{c \in C_i} W_v x_c \text{SoftMax}_{c \in C_i} \langle W_k x_c, W_q x_i \rangle$$

$W_v x_c$ — 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



Аналогия Attentive ARTM с моделью само-внимания

Контекстный тематический вектор на выходе E-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \operatorname{norm}_{t \in T} \left(\sum_{c \in C_i} \phi_{twc} \alpha_{ci} \frac{1}{p(t)} \phi_{twi} \right)$$

Контекстный вектор на выходе модели само-внимания:

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

Сходство:

- вектор термина w_i трансформируется в контекстный вектор
- путём усреднения векторов терминов 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

Свёрточная нейросеть GCNN (Gated Convolutional Network)

Входные векторы слов (эмбединги)

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

трансформируются в векторы слов, зависящие от контекстов C_i :

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

через адамарово произведение:

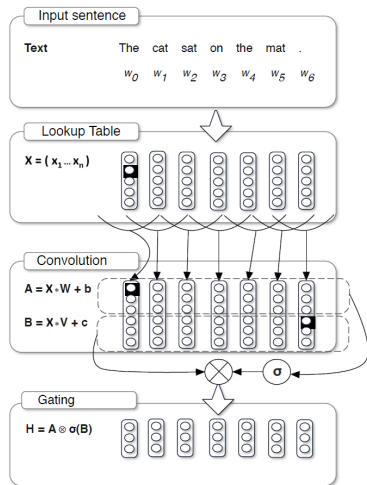
$$h_i = a_i \otimes \sigma(b_i), \text{ где}$$

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

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

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

$\sigma(x) = \frac{1}{1+e^{-x}}$ — 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:

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

Контекстный вектор на выходе модели GCNN:

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

Сходство:

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

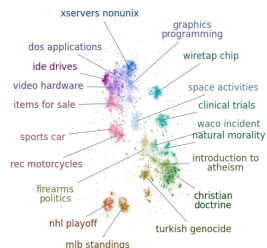
Отличия локализованного E-шага:

- 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 РТМ — конвейер 8 технологий:

- 1 векторизация токенов (Sentence-BERT)
- 2 векторизация предложений скользящим окном в 50 токенов (mean pooling)
- 3 понижение размерности векторов (UMAP)
- 4 иерархическая кластеризация (hDbSCAN), автоматическое определение числа тем
- 5 иерархическое укрупнение тем слиянием мелких кластеров с ближайшими соседями (Top2Vec)
- 6 разбиение документа на монотематические сегменты
- 7 $p(t|d)$ = доля векторов данной темы в документе
- 8 именованые 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 n -грамм и релевантных фраз по каждой теме
- именованное 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).

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.

Сходства и отличия CDC [2004] и Contextual-Top2Vec [2024]

Сходство:

- основаны на общих лингвистических допущениях:
 - гипотеза дистрибутивной семантики
 - каждая фраза — это сообщение о взаимодействии двух агентов [Тён ван Дейк], значит, относится к 1–2 темам
 - фразы одного абзаца имеют общую основную тему
- локальные контексты вместо документов
- тема — это кластер сегментов или их эмбедингов
- документ разбивается на монотематичные сегменты
- $p(t|d)$ = доля сегментов данной темы в документе
- легко отбирать фразы для именованя и суммаризации тем

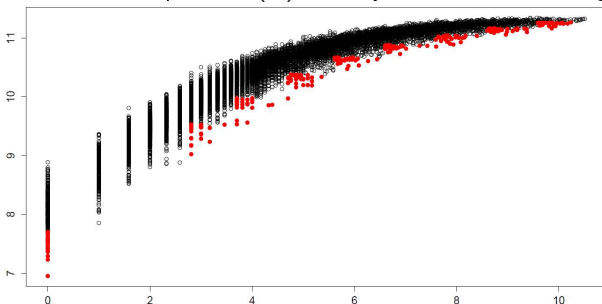
Отличия Contextual-Top2Vec:

- использование предобученных эмбедингов BERT
- есть понижение размерности
- нет отбора узких тематичных контекстов

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 $\log_2 N_w$



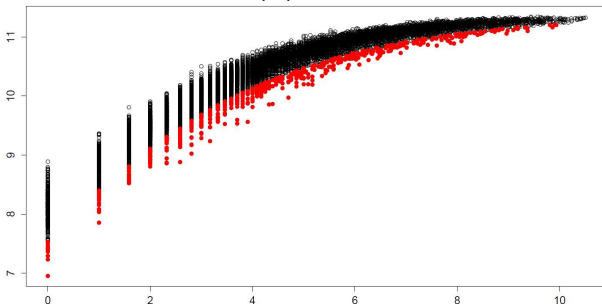
Недостаток: из-за разбиения на интервалы значительная часть узких контекстов пропускается (предвзятый отбор)

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

Выделение слов, имеющих узкие контексты

Закон Хипса $\Rightarrow H(w)$ зависит от $\log_2 N_w$ логарифмически
Отбор локальных контекстов 5%-квантильной регрессией:

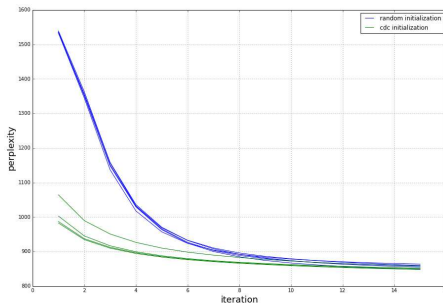
Зависимость энтропии $H(w)$ от документной частоты $\log_2 N_w$



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.

Правда ли, что CDC — по сути тематическая модель?

Зависимость перплексии от числа итераций (коллекция MMPO) при инициализации тематической модели с помощью CDC:



- инициализация CDC ускоряет сходимость,
- но она не идеальна, и EM-алгоритм заметно улучшает её

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

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 модель тематического внимания, затем её параметризация
- 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: $\min(5, \lfloor \text{score}/20 \rfloor)$ 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: $p(w|v, d) = \xi_{dvw}$.

Найти параметры модели ξ_{dvw} .

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 $p(r|w)$, $r \in \{\text{т, ш, ф}\}$.

Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.

Подсказка 3: можно использовать документную частоту слов.

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 (АТЕ)

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

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

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

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

$$p(t) = \sum_w \phi_{tw} p(w), \quad p(t) = \sum_d \theta_{td} p(d).$$

Как обеспечить выполнение этих условий в EM-алгоритме?

9. Докажите, что необходимым условием максимума

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

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

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

10. Выведите отсюда EM-алгоритм для частных случаев:

$$1) p(w, t|i, \Omega) = \phi_{wt} \theta_{td_i}$$

$$2) p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{w \in d_i} \frac{n_{d_i w}}{n_{d_i}} \phi_{tw};$$

$$3) p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c}.$$

11**. **Творческое задание.** Предложите способ ввести обучаемые параметры в тематическую модель внимания.

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

Исследуйте зависимость метрик качества модели

- перплексия: $\mathcal{P} = \exp\left(-\frac{1}{n} \sum_{i=1}^n p(w|C_i)\right)$
- разреженность, различность, когерентность тем
- дефекты целостности модели:

$$\|p(t) - \frac{n_t}{n}\|, \quad \|p(t) - \sum_t \phi_{tw} p(w)\|, \quad \|p(t) - \sum_t \theta_{td} p(d)\|$$

от номера итерации и от параметров модели:

- $|T|$ — число тем
- L — число проходов
- τ — вес N_{tw} в формуле M-шага, особый случай $\tau = 0$
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i, \beta$ — баланс левого и правого контекста
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, секций
- опция « $i \in C_i$ или $i \notin C_i$ »

- 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
 - Викиновости (1.5М статей, проект закрыт 30/03/2026)
 - Данные кадровых агентств: резюме + вакансии
 - Транзакции клиентов Sberbank DSD 2016
 - Акты арбитражных судов РФ

- «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

- 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 именованное и суммаризация тем
- 6 Калибровка моделей тематической фильтрации
- 7 Согласование тем с предобученными эмбедингами 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