

Вероятностные тематические модели

Лекция 2.

Онлайновый ЕМ-алгоритм и модели локальных контекстов

Константин Вячеславович Воронцов
k.v.vorontsov@phystech.edu

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

1 Алгоритмическая реализация ARTM

- Рациональный EM-алгоритм
- Онлайновый EM-алгоритм
- Пакетный онлайновый EM-алгоритм

2 Линейная тематизация текста

- Однопроходный E-шаг
- Локализованный E-шаг
- Двухнаправленные векторы контекста

3 Аналогии с нейросетевыми моделями языка

- Свёрточная нейросеть GCNN
- Модель внимания и трансформер
- Нейросетевая тематическая модель

Напоминание. Постановка задача и основная теорема ARTM

Дано: коллекция текстовых документов, $\hat{p}(w|d) = n_{dw}/n_d$

Найти: параметры тематической модели $p(w|d) = \sum_t \phi_{wt}\theta_{td}$

Критерий: максимум **регуляризованного** log-правдоподобия:

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

Теорема. Локальный экстремум (Φ, Θ) удовлетворяет системе уравнений, которая решается численно EM-алгоритмом:

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

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

Рациональный ЕМ-алгоритм

Идея: Е-шаг встраивается внутрь М-шага для каждого $d \in D$, чтобы не хранить трёхмерный массив значений n_{dwt} .

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

Выход: матрицы термов тем Φ и тем документов Θ ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D, w \in W, t \in T$;

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

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

для всех документов $d \in D$ и 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 $t \in T$;

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

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

Онлайновый EM-алгоритм

Вход: коллекция D , число тем $|T|$, параметры $j_{\max}$, γ ;

Выход: матрицы термов тем Φ и тем документов Θ ;

инициализировать $n_{wt} := 0$; $\tilde{n}_{wt} := 0$; $\phi_{wt} := \text{random}$;

для всех документов $d \in D$

инициализировать $\theta_{td} := \frac{1}{|T|}$;

для всех $j = 1, \dots, j_{\max}$ (итерации по документу)

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

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

$\tilde{n}_{wt} := \tilde{n}_{wt} + n_{tdw}$ для всех $w \in d$;

если пора обновить матрицу Φ **то**

$n_{wt} := \gamma n_{wt} + \tilde{n}_{wt}$; $\tilde{n}_{wt} := 0$;

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

Пакетный онлайновый EM-алгоритм в BigARTM

Коллекция D разбивается на пакеты D_b , $b = 1, \dots, B$, которые могут обрабатываться параллельно и/или распределённо.

Вход: коллекция документов D , число тем $|T|$,
параметры $\delta \equiv \text{decay_weight}$, $\alpha \equiv \text{apply_weight}$;

Выход: матрица Φ ;

инициализировать $n_{wt} := 0$, $\tilde{n}_{wt} := 0$, $\phi_{wt} := \text{random}$;

для всех пакетов D_b , $b = 1, \dots, B$

($\tilde{n}_{wt}$) := ($\tilde{n}_{wt}$) + **ProcessBatch**(D_b, Φ);
если пора обновить матрицу Φ **то**
 $n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}$, $\tilde{n}_{wt} := 0$;
 $\phi_{wt} := \text{norm}_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$;

Oleksandr Frei, Murat Apishev. Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.

Пакетный онлайновый EM-алгоритм: функция ProcessBatch

Функция **ProcessBatch** обрабатывает пакет документов D_b , не меняя матрицу Φ , и выдаёт счётчики термов в темах $\tilde{n}_{wt}$.

Вход: пакет D_b , матрица $\Phi = (\phi_{wt})$, параметр $j_{\max}$;

Выход: матрица счётчиков $(\tilde{n}_{wt})_{W \times T}$;

инициализировать $\tilde{n}_{wt} := 0$;

для всех $d \in D_b$

инициализировать $\theta_{td} := \frac{1}{|T|}$;

для всех $j = 1, \dots, j_{\max}$ (итерации по документу)

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

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

$\tilde{n}_{wt} := \tilde{n}_{wt} + n_{dw} p_{tdw}$;

Oleksandr Frei, Murat Apishev. Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.

Сравнение оффлайн и онлайн алгоритмов

Оффлайн EM-алгоритм:

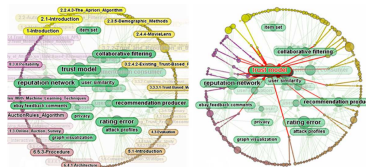
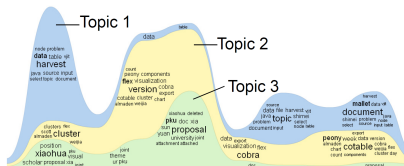
- 1 многократное итерирование по коллекции
- 2 однократный проход по документу
- 3 хранение матрицы Θ
- 4 обновление Φ в конце каждого прохода по коллекции
- 5 применяется при обработке небольших коллекций

Онлайн EM-алгоритм:

- 1 однократный проход по коллекции
- 2 многократное итерирование по каждому документу
- 3 нет необходимости хранить матрицу Θ
- 4 обновление Φ через заданное число пакетов
- 5 применяется при потоковой обработке больших коллекций

Мотивации. Что хотим:

- вместо «мешка слов» — последовательность $w_1, \dots, w_n$
- вместо документов — локальные контексты слов
- определять тематику любого фрагмента текста
- быстро находить фрагменты, относящиеся к данной теме
- в том числе фразы для суммаризации документа или темы
- разделять документ на тематически однородные сегменты
- визуализировать тематическую структуру документа



Идея тематизации текста за один проход

Дано: s — фрагмент текста d , Φ — тематическая модель

Найти: $p(t|s)$ — тематический вектор фрагмента текста

Проблемы:

- как не переобучить вектор $p(t|s)$, если текст s короткий?
- как согласовать $p(t|s)$ с объемлющим контекстом $p(t|d)$?
- как согласовать $p(t|s)$ с $p(t|w) = \phi_{wt} \frac{p(t)}{p(w)}$ термов $w \in s$?

Наводящие соображения:

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

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

- формула полной вероятности + гипотеза условн. независ.:

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

ЕМ-алгоритм для ARTM с явным выражением Θ через Φ

Максимизация логарифма правдоподобия:

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

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

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

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

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

Напоминание. Лемма о максимизации на единичных симплексах

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

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

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

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

Численное решение системы — методом простых итераций

Напоминание 2. Объединённый М-шаг

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

$$\begin{aligned} \text{Е-шаг: } & \left\{ p_{tdw} \equiv p(t|d, w) = \frac{\text{norm}(\phi_{wt} \theta_{td})}{\sum_{t \in T} \phi_{wt} \theta_{td}} \right. \\ \text{М-шаг: } & \left\{ \sum_{d,w} \sum_{t \in T} n_{dw} p_{tdw} \ln(\phi_{wt} \theta_{td}) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta} \right. \end{aligned}$$

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 М-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 :

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

■

ЕМ-алгоритм для ARTM с линейной тематизацией документов

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

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

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

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

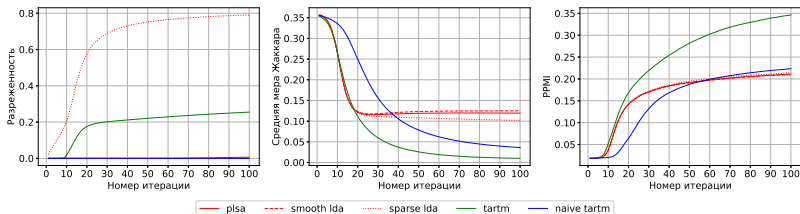
$$p'_{tdw} = p_{tdw} + \frac{\phi'_{tw}}{n_d} \left(\frac{n_{td}}{\theta_{td}} - \sum_{s \in T} \phi'_{sw} \frac{n_{sd}}{\theta_{sd}} \right)$$

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

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 ЕМ-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 М-шага приводит к упрощению: $p'_{tdw} = p_{tdw}$

ЕМ-алгоритм: метод простой итерации для системы уравнений

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

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

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

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

Линейная тематизация: от документа к локальным контекстам

Тематизация документа $d = (w_1, \dots, w_{n_d})$ за один проход:

$$\theta_{td}(\Phi) \equiv p(t|d) = \frac{1}{n_d} \sum_{i=1}^{n_d} p(t|w_i) = \frac{1}{n_d} \sum_{i=1}^{n_d} \phi'_{tw_i}$$

Тематизация *локального контекста* $C_i = (\dots, w_i, \dots)$ термина w_i :

$$\theta_{ti}(\Phi) \equiv p(t|C_i) = \frac{1}{|C_i|} \sum_{u \in C_i} p(t|u) = \frac{1}{|C_i|} \sum_{u \in C_i} \phi'_{tu}$$

Введём *коэффициенты внимания* (attention score) α_{ui} :

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

Локализованная тематическая модель:

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

ЕМ-алгоритм с локализованным E-шагом (Attentive ARTM)

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

C_i — локальный контекст (окружение) терма w_i

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

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

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

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

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

Быстрое вычисление коэффициентов внимания

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

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

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

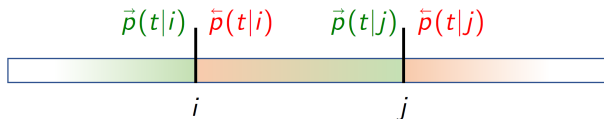
где $\vec{\gamma}_i, \bar{\gamma}_i$ — коэффициенты сглаживания в позиции i

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

Несколько соображений, как распоряжаться выбором $\vec{\gamma}_i, \bar{\gamma}_i$:

- $\gamma_i \approx \frac{1}{h}$, 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 термина w_i
- $\tilde{p}(t|i)$ — тематика правого контекста термина w_i
- $\frac{1}{2}(\vec{p}(t|i) + \tilde{p}(t|i))$ — тематика двустороннего контекста w_i
- $p(t|i \dots j) = \frac{1}{2}(\tilde{p}(t|i) + \vec{p}(t|j))$ — тематика сегмента $[i \dots j]$
- $\tilde{p}(t|i) \approx \vec{p}(t|j)$ — однородность тематики сегмента $[i \dots j]$
- $\max_i \|\vec{p}(t|i) - \tilde{p}(t|i)\|$ — граница i между сегментами
- при различных γ_i — короткие и длинные контексты

Сильно упрощённая модель внимания без обучаемых параметров с тематическими (интерпретируемыми!) эмбедингами.

Онлайновый EM-алгоритм с локализованным E-шагом

Вход: коллекция, число тем $|T|$, параметры $\beta, \vec{\gamma}_i, \tilde{\gamma}_i, \alpha, \delta$;

Выход: матрица Φ , векторы термов документов p_{ti} ;

инициализация: $n_{wt} := 0$; $\tilde{n}_{wt} := 0$; $n_t := 1$; $\phi_{wt} := \text{random}$;

для всех документов $d \in D$

$$p_{ti} := \text{norm}_t(\phi_{w_i t} n_t), \quad i = 1, \dots, n_d, \quad t \in T;$$

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

$$\tilde{\theta}_{ti} := \tilde{\gamma}_i p_{ti} + (1 - \tilde{\gamma}_i) \tilde{\theta}_{t,i+1}, \quad i = n_d, \dots, 1, \quad \tilde{\gamma}_{n_d} = 1, \quad t \in T;$$

$$p_{ti} := \text{norm}_t(\phi_{w_i t} (\beta \vec{\theta}_{ti} + (1 - \beta) \tilde{\theta}_{ti})), \quad i = 1, \dots, n_d, \quad t \in T;$$

$$\tilde{n}_{w_i t} := \tilde{n}_{w_i t} + p_{ti}; \quad n_t := n_t + p_{ti}, \quad i = 1, \dots, n_d, \quad t \in T;$$

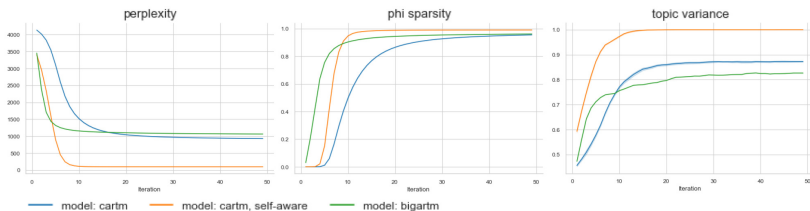
если пора обновить матрицу Φ **то**

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

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

Первая реализация и эксперименты с Attentive ARTM

Коллекция «20 Newsgroups»: $|D| = 18846$, $|W| = 107672$.



— улучшилась перплексия, разреженность Φ , различность тем

model	time@10 topics	time@30 topics	time@70 topics	time@100 topics
CARTM@CPU	4min 55s $\pm$ 1.7s	9min 20s $\pm$ 9.52s	20min 37s $\pm$ 2.36s	25min 52s $\pm$ 4.63s
BigARTM	1min 30s $\pm$ 1.6s	3min 5s $\pm$ 1.98s	4min 55s $\pm$ 653ms	6min 22s $\pm$ 5.81s

— время хуже в несколько раз, при этом реализация AttARTM на Python/JAX, тогда как ядро BigARTM на C++

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

Открытые проблемы и постановки задач

- надо ли исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}$, $\bar{\theta}_{ti}$?
 - какие другие варианты $\alpha(u|i)$ кроме скользящих средних?
 - как найти баланс β левого и правого контекста?
 - правильно ли подставлять p_{ti}/n_t вместо $\phi_{w_{it}}$ на Е-шаге?
 - имеет ли смысл увеличивать число проходов L ?
-
- как (и нужно ли) параметризовать модель внимания?
 - как обучать её параметры, разные для разных проходов?
 - как (и нужно ли) ввести аналог многих голов внимания?
-
- слишком много эвристических преобразований сделано... мы всё ещё решаем исходную оптимизационную задачу?
 - действительно ли на Е-шаге можно подвергать $p(t|d, w_i)$ всяким модификациям, почему и в каких пределах?

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 :

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

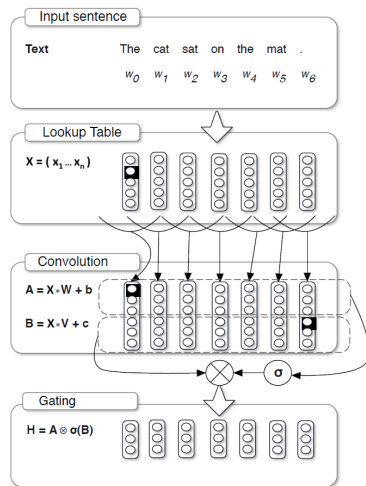
через адамарово произведение:

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

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

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

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



Аналогия локализованного Е-шага с моделью GCNN

Контекстный тематический вектор на выходе Е-шага:

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

Контекстный вектор на выходе модели GCNN:

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

Сходства:

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

Отличия:

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

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

Входные векторы слов (эмбединги)

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

трансформируются в векторы слов,
зависящие от контекстов C_i :

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

Модель внимания (self-attention):

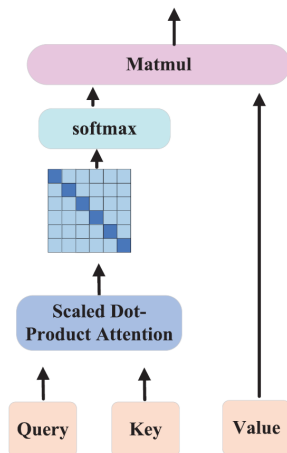
$$h_i = \sum_{u \in C_i} W_v x_u \text{ SoftMax}_{u \in C_i} \langle W_k x_u, W_q x_i \rangle$$

$W_v x_u$ — вектор-значение (value)

$W_k x_u$ — вектор-ключ (key)

$W_q x_i$ — вектор-запрос (query)

W_q, W_k, W_v — матрицы параметров



Аналогия локализованного Е-шага с моделью само-внимания

Контекстный тематический вектор на выходе Е-шага:

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

Контекстный вектор на выходе модели само-внимания:

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

Сходства:

- вектор терма w_i трансформируется в контекстный вектор
- путём усреднения векторов ϕ'_u из контекста терма w_i ,
- наиболее (семантически) схожих с вектором терма w_i

Отличия:

- адамарово умножение вектора ϕ'_u на вектор-фильтр ϕ_{w_i}
- нет обучаемых матриц W_q, W_k, W_v как у модели внимания
- проецирование итогового вектора на единичный симплекс

Аналогия локализованного Е-шага с моделью трансформера

Один проход документа аналогичен модели внимания:

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

$\phi'_{tw_i} = \text{norm}_t(\phi_{w_it} p_t)$ — бесконтекстный вектор термина $p(t|w_i)$

$\vec{p}(t|i) = \vec{\theta}_{ti}$, $\bar{p}(t|i) = \bar{\theta}_{ti}$ — векторы левого и правого контекста

$\theta_{ti} = \beta \vec{\theta}_{ti} + (1 - \beta) \bar{\theta}_{ti}$ — вектор двустороннего контекста

$p_{ti} = \text{norm}_t(\phi_{w_it} \theta_{ti})$ — контекстный вектор термина $p(t|C_i, w_i)$

Несколько таких проходов аналогичны трансформеру:

контекстный вектор термина $p_{ti} = p(t|C_i, w_i)$ с предыдущего прохода
используется вместо его бесконтекстного вектора $\phi'_{tw_i} = p(t|w_i)$

L итераций аналогичны L необучаемым блокам внимания

Онлайновый EM с многопроходным локализованным E-шагом

Вход: коллекция, число тем $|T|$, параметры $L, \beta, \vec{\gamma}_i, \tilde{\gamma}_i, \alpha, \delta$;

Выход: матрица Φ , векторы термов документов p_{ti} ;

инициализация: $n_{wt} := 0$; $\tilde{n}_{wt} := 0$; $n_t := 1$; $\phi_{wt} := \text{random}$;

для всех документов $d \in D$

$p_{ti} := \text{norm}_t(\phi_{wt} n_t)$;

для всех $l = 1, \dots, L$ (аналог L блоков внимания)

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

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

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

$\tilde{n}_{w_i t} := \tilde{n}_{w_i t} + p_{ti}; \quad n_t := n_t + p_{ti}$;

если пора обновить матрицу Φ **то**

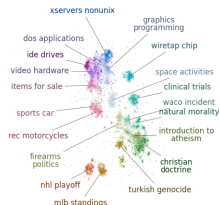
$n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}; \quad \tilde{n}_{wt} := 0$;

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

Нейросетевая тематическая модель Contextual-Top2Vec

Вместо PTM — сумма технологий:

- 1 векторизация токенов (Sentence-BERT)
- 2 векторизация предложений скользящим окном в 50 токенов (mean pooling)
- 3 понижение размерности векторов (UMAP)
- 4 иерархическая кластеризация (hDbSCAN)
с автоматическим определением числа тем
- 5 иерархическое укрупнение тем слиянием мелких кластеров
с ближайшими соседями (Top2Vec)
- 6 разбиение документа на монотематические сегменты
- 7 $p(t|d)$ = доля векторов данной темы в документе
- 8 именованение тем: поиск фраз, ближайших к центроиду темы



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

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

Нейросетевая тематическая модель Contextual-Top2Vec

Достоинства:

- модель BERT предобучена по большим внешним данным, поэтому качество тем не зависит от размера коллекции
- документ разбивается на монотематические сегменты
- автоматически определяется число тем (hDbSCAN, Top2Vec)
- тема описывается фразами, а не отдельными словами

Недостатки:

- это работает долго, особенно на больших коллекциях
- инкрементное добавление документов не предполагается

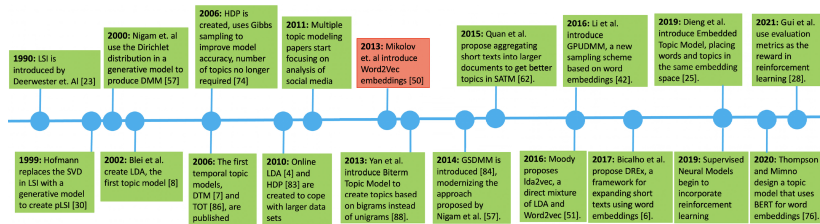
Сходство локализованного E-шага:

- обработка локальных контекстов скользящим окном
- возможно разреживать $p(t|C_i)$ до монотематичности

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

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

- Нужно больше экспериментов с AttARTM и аналогами
- LDA стала основной для создания сотен PTM моделей, сможет ли AttARTM претендовать на роль лучшей основы?
- Neural Topic Models — поток публикаций начиная с 2016
Как «объединить лучшее от двух миров»?



Xiaobao Wu, Thong Nguyen, Anh Tuan Luu. A Survey on Neural Topic Models: Methods, Applications, and Challenges. 2023

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

He Zhao et al. Topic Modelling Meets Deep Neural Networks: A Survey. 2021

Tian Tian et al. Attention-based Autoencoder Topic Model for Short Texts. 2019

Shuangyin Li et al. Recurrent Attentional Topic Model. 2017

Задача-минимум: научиться решать задачи NLP
с использованием тематического моделирования в BigARTM

Задача-максимум: сделать полезное мини-исследование

виды деятельности	оценка
теоретические задания	$\sum_i X_i$
решение прикладной задачи	5X
обзор по последним NeuralTM	5X
интеграция ARTM в pyTorch	5X
участие в одном из проектов	10X
работа над открытой проблемой	10X

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

Итоговая оценка: $\min(10, \lfloor \text{score}/5 \rfloor)$ по 10-балльной шкале.

Теоретическое задание к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Униграммная модель документов: $p(w|d) = \xi_{dw}$

Найти параметры модели ξ_{dw} .

2. Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d

Найти параметры модели ξ_w .

Подсказка: применить условия ККТ или основную лемму.

3. Творческое задание (возможны разные решения)

Предложите модель, определяющую 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. Запишите критерий логарифма правдоподобия с регуляризацией для тематической модели $p(w|d) = \sum_t \phi_{wt} \theta_{td}$, используя исходные данные $(d_i, w_i)_{i=1}^n$ вместо счётчиков n_{dw} . Выведете из него ЕМ-алгоритм, докажите его эквивалентность обычному ЕМ-алгоритму для ARTM.

5. Запишите критерий логарифма правдоподобия для локализованной тематической модели $p(w|C_i) = \sum_t \phi_{wt} p(t|C_i)$. Выведете из него ЕМ-алгоритм с локализованным Е-шагом.

Какие приближения пришлось сделать в процессе вывода?
Какие переменные удобнее оставить в модели, ϕ_{wt} или ϕ'_{tw} ?

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

Открытая проблема. Продолжить исследование Ильи Ирхина:

- Освоить код: https://github.com/ilirhin/python_artm
- Реализовать локализованный E-шаг

Исследовать зависимость метрик качества от параметров (перплексия, разреженность, различность, когерентность):

- L — число проходов
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i$ — асимметричность левого и правого контекста
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, глав
- β — баланса левого и правого контекста
- α, δ — параметры онлайн-алгоритма EM
- опция «подставлять p_{ti}/n_t вместо $\phi_{w_i t}$ на E-шаге»
- опция «исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}, \overleftarrow{\theta}_{ti}$ »

- Открытые датасеты (английский): 20 newsgroups, NIPS, KOS
- Научные статьи: eLibrary, Semantic Scholar, arXiv, PubMed
- Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
- TechCrunch (английский)
- Данные социальных сетей: VK, Twitter, Telegram,...
- Википедия
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

- «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: ускорить понимание предметной области
- «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: q&q аналитика проблемной среды

- 1 Проблема несбалансированности тем в коллекции
- 2 Обеспечение 100%-й интерпретируемости тем
- 3 Тематические модели внимания последовательного текста
- 4 Обнаружение новых тем или трендов в потоке текстов
- 5 Автоматическое именованное и аннотирование тем
- 6 Обзор подходов в нейросетевых тематических моделях
- 7 Обеспечение полноты и устойчивости множества тем
- 8 Автоматический подбор гиперпараметров, AutoML
- 9 Оптимизация гиперпараметров в потоковом режиме
- 10 Проблема несбалансированности текстов по длине
- 11 Бережное слияние моделей нескольких коллекций
- 12 Гиперграфовые тематические модели в RecSys