Тематическое моделирование

Анализ социальных сетей
Лекция 2

Михаил Пожидаев

Институт прикладной математики и компьютерных наук ТГУ

3 ноября 2026 г.

Тематическое моделирование отвечает на вопрос о выделении фактических тем в массиве документов, когда тематические категории изначально не заданы. Анализ при этом выполняется на основе множества лексем по принципу «мешка слов». При таком подходе учитывается количество лексем без учёта их взаимного расположения.

Применение

Тематическое моделирование применяют для анализа отзывов, новостей, научных публикаций, обращений в поддержку, сообщений в соцсетях и юридических документов. Оно помогает быстро понять, о чём пишут пользователи, найти повторяющиеся проблемы, сгруппировать похожие тексты и построить удобную навигацию по большим архивам.

Важно

Найденные темы не всегда являются «истинными» категориями. Их нужно интерпретировать человеку, потому что качество результата зависит от очистки текста, выбранного числа тем, языка, размера корпуса и параметров модели.

CC BY-NC 4.0

TF-IDF

Term freq — inverse document freq

TF-IDF — статистическая мера, характеризующая вес лексемы в некотором документе в составе множества документов.

Является хорошей характеристикой для **лексического** анализа текстов и корпусов текстов.

$$tfidf_{ij} = tf_{ij} idf_{ij}$$

$$tf_{ij} = \frac{c_{ij}}{n_j}$$

$$idf_{ij} = \log \frac{p}{d_i}$$

На практике часто применяется **BM25** — усовершенствованный вариант TF-IDF, который учитывает насыщение частоты термина и нормализует влияние длины документа.

Обозначения:

- i — анализируемая лексема, j документ в корпусе документов;
- c_{ij} — количество употреблений лексемы в документе, n_j — общее количество всех лексем в документе;
- p — общее количество документов, d_i — количество документов, содержащих лексему i.

title em-алгоритм SUBTITLE Назначение EM-алгоритма

EM-алгоритм (Expectation-Maximization) применяется для тематического моделирования и других задач со скрытыми переменными. В строгом смысле EM-алгоритм — это метод, используемый для нахождения максимального правдоподобия параметров вероятностных моделей.

- Важным условием его применения является существование множества скрытых (ненаблюдаемых) переменных.
- Значения скрытых переменных должны вычисляться на основе текущего приближения к искомым значениям параметров.

Функция правдоподобия

EM-алгоритм вводит функцию правдоподобия и итеративно максимизирует её значение. Он может считаться итеративной разновидностью метода максимального правдоподобия, предполагающего аналитический поиск экстремумов значений функции правдоподобия. Функция правдоподобия позволяет оценить параметры статистической модели на основе наблюдаемых данных. Она показывает, насколько вероятно получить конкретные данные при заданных параметрах модели. Основная задача состоит в поиске таких значений параметров, при которых наблюдаемые данные становятся наиболее вероятными.

Функция правдоподобия

Если параметр модели обозначен как θ, то функцию правдоподобия обычно записывают как $\phi(x|\theta)$. В аналитическом виде поиск экстремумов значений параметров проводят путём вычисления производной по θ. Для EM-алгоритма необходимо задать некоторые начальные значения искомых параметров.

Сходимость

Работа алгоритма продолжается до наступления сходимости к искомым значениям. Существование сходимости строго не гарантируется. Сходимость может зависеть от выбранных начальных значений искомых параметров, и такая чувствительность является недостатком алгоритма.

Смесь распределений

Смесь распределений

Зададим смесь распределений k случайных величин $p(x) = \sum_{j=1}^{k} w_j p_j(x)$, в которой $w_j \geq 0$ — вес компоненты с выполнением условия нормировки $\sum_{j=1}^{k} w_j = 1$.

Функция правдоподобия компонента смеси

Каждый компонент смеси распределений представляет собой заданную функцию правдоподобия $p_j(x) = \phi(x|\theta_j)$, зависящую от неизвестного и искомого параметра θ_j. Общий вид функции $\phi(x)$ для каждого компонента смеси задан одинаково, но имеет разные значения параметров. [ex] Пример Например, $\phi(x)$ может быть определена как нормальное распределение, но для разных компонентов смеси иметь разные значения математического ожидания и дисперсии.

Искомые параметры

Общее множество искомых параметров имеет вид: $\Theta = \{w_1, w_2, \ldots, w_k; \theta_1, \theta_2, \ldots, \theta_k\}$.

Обозначения

- Элементы w_j — это неизвестные веса компонентов в смеси распределений.
- Элементы θ_j — это неизвестные параметры функции правдоподобия, которая используется в качестве функции распределения соответствующей компоненты смеси.

Целевая функция максимизации

Пусть имеется m наблюдений $x_1, x_2, \ldots, x_m$. Необходимо найти такие значения Θ, чтобы выражение $\prod_{i=1}^{m} p(x_i)$ имело бы наибольшее значение.

Поскольку логарифм — это монотонно возрастающая функция, а все $p(x)$ неотрицательны, можно максимизировать логарифм выражения: $\ln \prod_{i=1}^{m} p(x_i)$ Подставляя $p(x_i)$, получаем основное выражение, которое следует максимизировать:

$$\sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j \phi(x_i|\theta_j)$$

Найти аналитически максимум логарифма суммы невозможно

Максимум ищется итерационно и численно с использованием множества скрытых переменных.

- На шаге expectation значения скрытых переменных вычисляются на основе текущего значения параметров.
- На шаге maximization значения параметров вычисляются на основе значений скрытых переменных.

Шаг expectation

Введём матрицу скрытых переменных.

Предположим, что наступило событие x_i. Вероятность того, что параметр θ_j является корректным для распределения j-той компоненты при предположении, что событие x_i действительно принадлежит этой компоненте, обозначим как $p(\theta_j|x_i)$. Применив формулу Байеса, получаем:

$$p(\theta_j|x_i) = \frac{p(x_i|\theta_j)p(\theta_j)}{p(x_i)}$$

Вычисление скрытых переменных

Значение $p(\theta_j)$ является вкладом j-той компоненты в смесь и равняется просто w_j. Значение $p(x_i|\theta_j)$ можно заменить значением функции правдоподобия $\phi(x_i|\theta_j)$, которое соответствует $p_j(x_i)$.

В результате выражение принимает вид:

$$p(\theta_j|x_i) = \frac{w_j\phi(x_i|\theta_j)}{p(x_i)}$$

Обозначим его как h_{ij}, и пусть оно является одной из введённых скрытых переменных. Общее множество скрытых переменных записывается в матрицу H размерности $m \times k$.

Шаг maximization

И критерий для остановки

Весовые коэффициенты компонентов смеси

Оптимальные веса w_j вычисляются как среднее арифметическое значение для j: $w_j = \frac{1}{m} \sum_{i=1}^{m} h_{ij}$.

Параметры распределений

Исходное выражение максимизации может быть записано как $\sum_{i=1}^{m} \ln \sum_{j=1}^{k} h_{ij} p(x_i)$, и в таком виде оно становится вычислимым.

Критерий остановки

Обычно используют проверку на достижение заданного числа итераций. Также используют малозначительное изменение логарифмической функции правдоподобия между последовательными итерациями. Ещё одним критерием является достижение максимального числа шагов без значительного улучшения оценки параметров.

Спасибо за внимание!

Всё о курсе:
https://marigostra.ru/materials/sna.html

Электронная почта:
msp@luwrain.org

Telegram:
@Marigostra



Материал можно распространять с соблюдением требований Creative Commons Attribution-NonCommercial 4.0 International License. При обнаружении любых ошибок и неточностей в материале просьба сообщить о них автору.