Тематическое моделирование

- *Тематическое моделирование (topic modeling)* – одно из современных приложений *машинного обучения к анализу текстов*
- Тематическая *модель (topic model)* коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему.
- **Вероятностная тематическая модель (ВТМ)** описывает каждую тему дискретным распределением на множестве терминов, каждый документ дискретным распределением на множестве тем.
- Предполагается, что коллекция документов – это последовательность терминов, выбранных *случайно из смеси таких распределений.*
- Поскольку документ или термин может относиться ко многим темам с различными вероятностями, говорят, что ВТМ осуществляет «мягкую» кластеризацию документов и терминов по кластерам-темам.
- Синонимы, часто употребляющиеся в схожих контекстах, с большой вероятностью попадают в одну тему. Омонимы, употребляющиеся в разных контекстах, распределяются между несколькими темами соответственно частоте употребления.

Тематические **модели применяются для**:

- **выявления трендов** в научных публикациях или новостных потоках;
- **классификации** документов и изображений;
- **семантического информационного поиска**, в том числе многоязычного;
- **обнаружения текстового спама**;
- **в рекомендательных системах.**

Применение BTM для тематического поиска научных публикаций

Документы представляются векторами, координаты которых соответствуют словам, а значения – статистическим характеристикам слов, например, частотам.

Поиск документов по коротким запросам реализуется путем поиска векторов, в которых часто встречаются слова запроса.

Вероятностная *модель коллекции документов*

- Пусть D – множество (коллекция) текстовых документов, W – множество (словарь) всех употребляемых в них терминов [слов или словосочетаний]. Каждый документ $d \in D$ представляет собой последовательность n терминов (w_1, w_2,…, w_n) из словаря W. Термин может повторяться в документе много раз.

- **Вероятностное пространство** и **гипотеза независимости**. Предполагается, что существует конечное множество тем T, и каждое употребление термина w в каждом документе d связано с некоторой неизвестной темой $t \in T$. Коллекция документов рассматривается как множество троек (d, w, t), выбранных случайно из дискретного распределения p (d, w, t), заданного на конечном множестве D x W x T. Документы и термины – наблюдаемые переменные, тема – латентная (скрытая) переменная. Гипотеза о независимости элементов выборки (гипотеза «мешка слов» (bag of words)): порядок терминов в документах не важен для выявления тематики. Порядок документов в коллекции также не имеет значения (гипотеза «мешка документов»).

- **Постановка задачи тематического моделирования.** Построить тематическую модель коллекции документов – значит найти множество тем T, распределения p (w I t) для всех тем и распределения p (t I d) для всех документов. «Мягкая» кластеризация означает, что каждый документ или термин не жестко приписывается какой-то одной теме, а распределяется по нескольким темам.

- Гипотеза условной независимости. Появление слов в документе d, относящихся к теме t, описывается общим для всей коллекции распределением p (w I t) и не зависит от документа.
- Вероятностная модель. Согласно формуле полной вероятности и гипотезе условной независимости

$$p(w \mid d) = \sum p(t \mid d)\, p(w \mid t)$$

- Алгоритм 1:
- Вход: распределения p (t I d), p (w I t).

1 для всех d

задать длину n документа d;

2 для всех i=1…n

выбрать случайную тему t из распределения p(t I d),

выбрать случайный термин w из распределения p(w I t).

Добавить в выборку пару (d, w). Тема забывается.

Выход: выборка пар (d_i, w_i), где i=1, …, n.

Частотные (выборочные) оценки вероятностей. Вероятности, связанные с наблюдаемыми переменными d и w, можно оценивать по выборке как частоты (здесь и далее выборочные оценки вероятностей p будем обозначать через $\hat{p}$):

$$\hat{p}(d,w) = \frac{n_{dw}}{n}, \qquad \hat{p}(d) = \frac{n_d}{n}, \qquad \hat{p}(w) = \frac{n_w}{n}, \qquad \hat{p}(w \mid d) = \frac{n_{dw}}{n_d}, \tag{1.3}$$

n_{dw} — число вхождений термина w в документ d;

$n_d = \sum\limits_{w \in W} n_{dw}$ — длина документа d в терминах;

$n_w = \sum\limits_{d \in D} n_{dw}$ — число вхождений термина w во все документы коллекции;

$n = \sum\limits_{d \in D} \sum\limits_{w \in d} n_{dw}$ — длина коллекции в терминах.

Вероятности, связанные со скрытой переменной t, также можно оценивать как частоты, если рассматривать коллекцию документов как выборку троек (d, w, t):

$$\hat{p}(t) = \frac{n_t}{n}, \qquad \hat{p}(w \mid t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(t \mid d) = \frac{n_{dt}}{n_d}, \qquad \hat{p}(t \mid d, w) = \frac{n_{dwt}}{n_{dw}}, \tag{1.4}$$

n_{dwt} — число троек, в которых термин w документа d связан с темой t;

$n_{dt} = \sum\limits_{w \in W} n_{dwt}$ — число троек, в которых термин документа d связан с темой t;

$n_{wt} = \sum\limits_{d \in D} n_{dwt}$ — число троек, в которых термин w связан с темой t;

$n_t = \sum\limits_{d \in D} \sum\limits_{w \in d} n_{dwt}$ — число троек, связанных с темой t.

- В вероятностном тематическом моделировании вместо метода наименьших квадратов используется метод **максимума правдоподобия**.

- **Лемматизация** – это приведение каждого слова в документе к его нормальной форме
- **Стемминг** состоит в отбрасывании изменяемых частей слов (окончаний)

- Отбрасывание «***запрещенных***» слов
- Отбрасывание ***редких*** слов (встречающихся в документе только 1 раз)

- Выделение **устойчивых оборотов** конкретной предметной области
- **Алгоритм** 1 можно использовать для генерации модельных данных по заданным распределениям

Вероятностный латентный семантический анализ

- Предложен Томасом Хофманном.

- Вероятностная модель появления пары «документ-термин» записывается тремя способами

  $p(w, d)=\sum p(t)\, p(d \mid t)\, p(w \mid t)=\sum p(d)\, p(t \mid d)\, p(w \mid t)= \sum p(w)\, p(d \mid t)\, p(t \mid w)$, где p (t) – распределение тем по всей коллекции

Алгоритм 2

Вход: коллекция документов D, число тем $|T|$, начальные приближения Θ, Φ;
Выход: распределения Θ и Φ;

повторять
  обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D,\ w \in W,\ t \in T$;
  для всех $d \in D,\ w \in d$
    $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$;
    для всех $t \in T$ таких, что $\varphi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на $\delta = n_{dw}\varphi_{wt}\theta_{td}/Z$;
  $\varphi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D,\ t \in T$;
пока Θ и Φ не сойдутся;

- **Иерархические тематические модели**

Для больших коллекций текстовых документов естественно строить иерархии вложенных друг в друга тем (называемых также категориями или рубриками), чтобы упростить поиск документов.

Иерархия – это общепринятый способ структуризации знаний.

- **Оптимизация структуры иерархии по коллекции документов** – открытая проблема

Определение тематического дерева

- Гипотеза о существовании тематического дерева. Рассмотрим дерево с множеством вершин V и корнем $t_0 \in V$. Вершины дерева соответствуют темам. Каждой теме t из V соответствует множество ее подтем – дочерних вершин в дереве $S_t \subset V$. Каждое ребро дерева соответствует паре «тема-подтема» (t, s), $s \in S_t$. Если $S_t = 0$, то тема t называется терминальной или листом тематического дерева. Для каждой вершины t в дереве V существует только одна родительская вершина, следовательно, только один путь $(t_0, \ldots, t)$ от корня дерева t_0 до темы t.
- Гипотезы: 1) если пара (d, w) связана с темой t, то она связана и со всеми темами выше вершины t на пути до корня t_0.

  2) если пара (d, w) не связана с темой t, то она не связана и со всеми подтемами в поддереве ниже вершины t.

Вероятностная интерпретация отношения «тема-подтема»

Вероятностная интерпретация отношения «тема–подтема». Каждому ребру тематического дерева (t, s) соответствует условная вероятность $p(s \mid t)$ того, что термин документа, связанный с темой t, связан также с подтемой $s \in S_t$:

$$p(s \mid t) = \frac{p(t, s)}{p(t)} = \frac{p(s)}{p(t)}. \tag{5.1}$$

Если рассматривать коллекцию документов как выборку троек (d, w, t), то частотной оценкой этой условной вероятности будет $\hat{p}(s \mid t) = n_s / n_t$ — доля троек, связанных с подтемой s, среди всех троек, связанных с темой t.

Условные вероятности подтем удовлетворяют ограничениям нормировки, которые, в силу (5.1), допускают две эквивалентные записи:

$$\sum_{s \in S_t} p(s \mid t) = 1, \quad \sum_{s \in S_t} p(s) = p(t), \quad t \in V. \tag{5.2}$$

Обозначим через T множество тем, соответствующих терминальным вершинам дерева V. Условие нормировки

$$\sum_{t \in T} p(t) = 1. \tag{5.3}$$

выполняется именно для этого множества, а не для всего множества тем в дереве V.

Вероятностная интерпретация отношения «тема-подтема»

Из (5.2) следует, что условие нормировки останется справедливым, если заменить любое из множеств S_t его родительской темой t, а также если делать такие замены многократно в произвольном порядке. В частности, для корневой темы $p(t_0) = 1$.

При разделении темы t на подтемы $s \in S_t$ условные распределения для подтем $\varphi_{ws} = p(w \mid s)$ и $\theta_{sd} = p(s \mid d)$ должны удовлетворять требованиям нормировки

$$\sum_{w \in W} \varphi_{ws} = 1, \quad s \in S_t; \qquad \sum_{s \in S_t} \theta_{sd} = \theta_{td}, \quad d \in D. \tag{5.4}$$

Распределения $p(s \mid w) = \varphi_{ws} \frac{p(s)}{p(w)}$ и $p(d \mid s) = \theta_{sd} \frac{p(d)}{p(s)}$ также должны быть нормированы, откуда следуют ещё две серии тождеств:

$$\sum_{s \in S_t} \varphi_{ws}\, p(s) = \varphi_{wt}\, p(t), \quad w \in W; \qquad \sum_{d \in D} \theta_{sd}\, p(d) = p(s), \quad s \in S_t. \tag{5.5}$$