Тематическое моделирование

Зенкова Наталья
Калина Екатерина
Балагуров Владимир

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Санкт-Петербург
2019г.

Дано: Несколько миллионов документов.

Задача: Выявить тематики в текстовой коллекции (какие темы существуют в наших документах?).

Что такое «тема»?

- Тема — специальная терминология предметной области;
- Тема — набор часто совместно встречающихся терминов;
- Тема — семантически однородный кластер текстов.

Некоторые приложения тематического моделирования

- Разведочный поиск в электронных библиотеках;
- Поиск тематического контента в соцсетях;
- Детектирование и трекинг новостных сюжетов;
- Мультимодальный поиск текстов и изображений;
- Анализ банковских транзакционных данных.

Пусть

- D — конечное множество текстовых документов;
- W — конечное множество слов (терминов, токенов);
- T — конечное множество тем;
- $D \times W \times T$ — дискретное вероятностное простанство;
- Коллекция — это i.i.d выборка $(d_i, w_i, t_i)_{i=1}^{n} \sim p(d, w, t)$.

Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, \ldots, w_{n_d})$ из словаря W. Термин может повторяться в документе много раз.

d_i, w_i — наблюдаемые величины, t_i — скрытые

Задача тематического моделирования: Найти множество тем T, распределение $p(w|t)$ для всех тем $t \in T$, распределение $p(t|d)$ для всех документов.

Далее, найденные распределения могут использоваться для решения прикладных задач.

Что такое «тема» в коллекции текстовых документов?

Более формально

- Тема — условное распределение на множестве терминов, $p(w|t)$ — вероятность термина w в теме t;
- Тематический профиль документа — условное распределение $p(t|d)$ — вероятность темы t в документе d.

Когда автор писал термин w в документе d, он думал о теме t, и мы хотели бы выявить, о какой именно.

Тематическая модель выявляет латентные (скрытые) темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Отличие от кластеризации:

- Жесткая кластеризация: кластеризация текстов новостей (относим новостное событие к определенной теме);
- Мягкая кластеризация: кластеризация научных статей (много исследований находятся на стыке наук).

Документ d может быть связан с несколькими темами t.

Гипотезы и предположения

Гипотеза назависимости: Порядок слов в документе и порядок документов в коллекции не важны.

Гипотеза условной независимости: $p(w|d,t) = p(w|t)$.

Гипотеза разреженности: Каждый документ d и каждый термин w связан с небольшим количеством тем t.

Как получить разреженность?

- Документ относится к большому количеству тем.

  Решение: разобьем его на части, более однородные по тематике.

- Термин относится к большому числу тем.

  Решение: положим, что термин является общеупотребительным словом и несет мало полезной информации с точки зрения тематики.

Тематическая модель по формуле полной вероятности:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d).$$

Предварительная обработка документов

Мотивация: Для упрощения модели прибегают к предварительной обработке текстов.

- *Лемматизация* — приведение каждого слова в документе к его начальной форме.

  Трудоемкий процесс

- *Стемминг* — отбрасывание изменяемых частей слова.

  Большое число ошибок

- *Уменьшение словаря:*
  - $1000, 5, 23 \to$ \$number, $(5+3)$, $\frac{1}{2}\mathbf{w}\mathbf{w}^{\mathrm{T}} + C \to$ \$formula;
  - *Отбрасывание стоп-слов* — удаление слов (предлогов, союзов, вводных слов и т.д.), которые никак не характеризуют тему.
    Почти не влияет на длину словаря
  - *Отбрасывание редких слов.*
    Для коллекций коротких новостных сообщений лучше не использовать
  - *Выделение ключевых фраз.*
    Приходится привлекать экспертов

Прямая задача — порождение коллекции по $p(w|t)$ и $p(t|d)$

Тематическая модель по формуле полной вероятности:

$$p(w|d) = \sum_{t \in T} p(w|d,t)p(t|d) = \sum_{t \in T} p(w|t)p(t|d), \text{ где}$$

- $p(w|t)$ — распределение терминов в каждой теме,
- $p(t|d)$ — распределение тем в каждом документе.

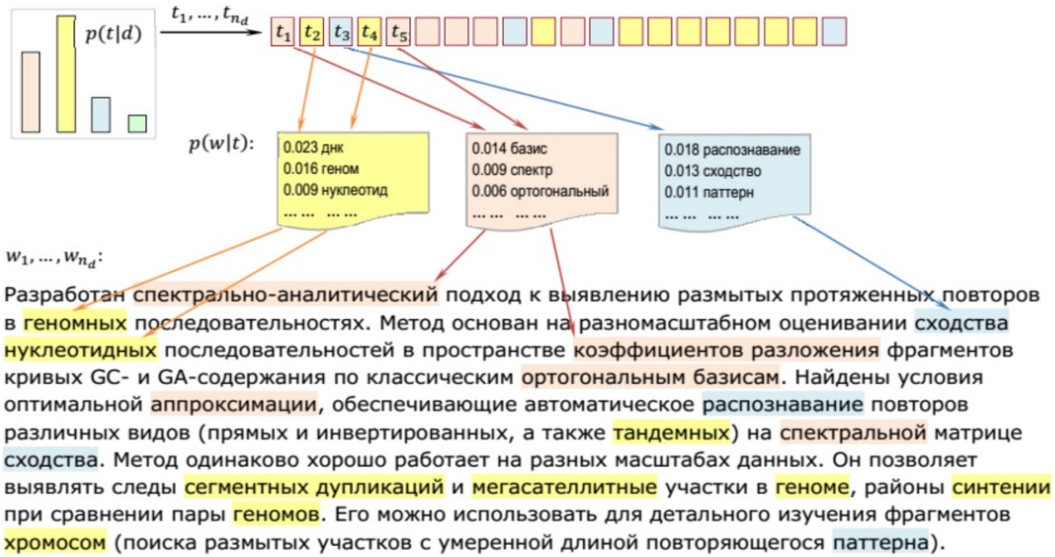


Рис. 1.1. Процесс порождения текстового документа вероятностной тематической моделью

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

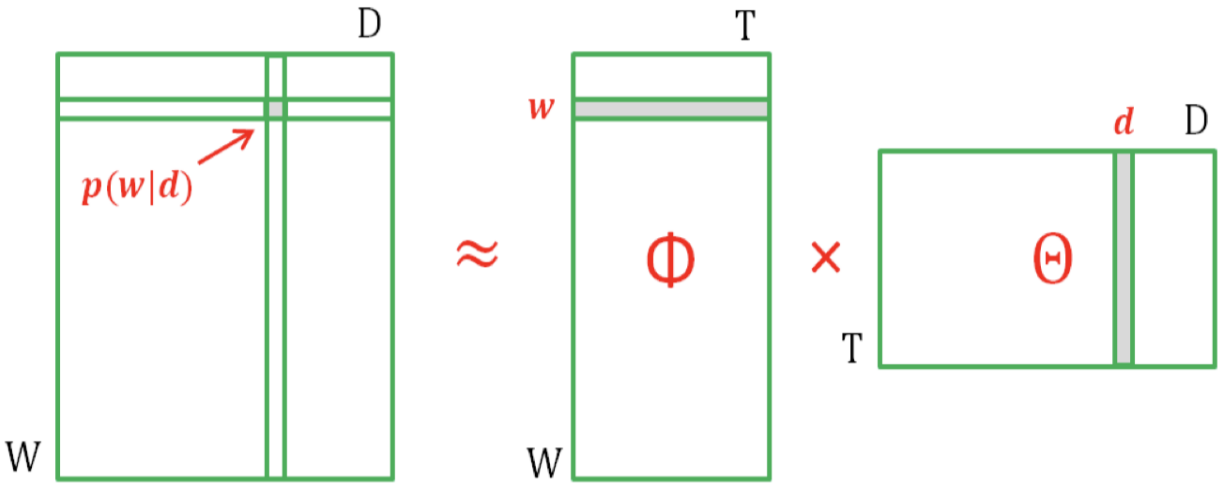
Дано: коллекция текстовых документов

- n_{dw} — частоты терминов в документах, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

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

- $\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:



- Если $\mathbf{\Phi}$ и $\mathbf{\Theta}$ — решение, то существует матрица $\mathbf{S}$ ранга $|T|$ такая, что $\mathbf{\Phi}'\mathbf{\Theta}' = (\mathbf{\Phi}\mathbf{S})(\mathbf{S}^{-1}\mathbf{\Theta})$ и $\mathbf{\Phi}'$ и $\mathbf{\Theta}'$ тоже стохастические.

Наблюдаемые частоты:

$$\hat{p}(d,w) = \frac{n_{dw}}{n}, \quad \hat{p}(d) = \frac{n_d}{n}, \quad \hat{p}(w) = \frac{n_w}{n}, \quad \hat{p}(w,d) = \frac{n_{dw}}{n_d}, \text{ где}$$

- n_{dw} — число вхождений термина w в документ d;
- $n_d = \sum_{w \in W} n_{dw}$ — длина документа d в терминах;
- $n_w = \sum_{d \in D} n_{dw}$ — число вхождений термина w во все документы коллекции;
- $n = \sum_{d \in D} \sum_{w \in W} n_{dw}$ — длина коллекции d в терминах.

Ненаблюдаемые частоты, связанные с t:

$$\hat{p}(t) = \frac{n_t}{n}, \quad \hat{p}(w|t) = \frac{n_{wt}}{n_t} \quad \hat{p}(t|d) = \frac{n_{dt}}{n_d} \quad \hat{p}(t|d,w) = \frac{n_{dwt}}{n_{dw}} \text{ где}$$

- n_{dwt} — число троек, в которых термин w в документе d связан с темой t;
- $n_{dt} = \sum_{w \in W} n_{dwt}$ — число троек, в которых термин в документе d связан с темой t;
- $n_{wt} = \sum_{d \in D} n_{dwt}$ — число троек, в которых термин w связан с темой t;
- $n_t = \sum_{d \in D} \sum_{w \in W} n_{dwt}$ — число троек, связанных с темой t.

Рассмотрим один из способов описания тематической модели представления коллекции текстовых документов.

$$\mathbf{F} \approx \mathbf{\Phi\Theta}, \text{ где } \mathbf{\Phi} = (\phi)_{W\times T}, \phi_{wt} = p(w|t); \quad \mathbf{\Theta} = (\theta)_{T\times D}, \theta_{td} = p(t|d).$$

Для оценивания параметров $\mathbf{\Phi}$ и $\mathbf{\Theta}$ тематической модели будем максимизировать функцию правдоподобия:

$$\mathcal{L}(\mathbf{\Phi},\mathbf{\Theta}) = C\prod_{d\in D}\prod_{w\in d} p(d,w)^{n_{dw}} = \prod_{d\in D}\prod_{w\in d} p(w|d)^{n_{dw}} \underbrace{Cp(d)^{n_{dw}}}_{const} \to \max_{\mathbf{\Phi\Theta}},$$

где C — нормировочный множитель.

Задача максимума правдоподобия с ограничениями:

$$\begin{cases} \mathcal{L}_{\log}(\mathbf{\Phi},\mathbf{\Theta}) = \prod\limits_{d\in D}\prod\limits_{w\in d} n_{dw} \ln \sum\limits_{t\in T} \phi_{wt}\theta_{td} \to \max\limits_{\mathbf{\Phi\Theta}} \\ \phi_{wt} \geq 0, \sum\limits_{w\in W} \phi_{wt} = 1 \\ \theta_{td} \geq 0, \sum\limits_{t\in T} \theta_{td} = 1. \end{cases}$$

Для решения задачи применяется EM-алгоритм

E-шаг. Вычисляются условные вероятности $p(t|d,w)$ всех тем $t \in T$ для каждого термина $w \in W$ в каждом документе d:

$$H_{dwt} = p(t|d,w) = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}}.$$

M-шаг. По условным вероятностям тем H_{dwt} вычисляется новое приближение параметров ϕ_{wt} и θ_{td}:

$$\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \quad \hat{n}_{wt} = \sum_{d\in D} n_{dw}H_{dwt}, \quad \hat{n}_t = \sum_{w\in W}\hat{n}_{wt};$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \quad \hat{n}_{dt} = \sum_{w\in d} n_{dw}H_{dwt}, \quad \hat{n}_d = \sum_{t\in T}\hat{n}_{dt}.$$

Начальное приближение ϕ_{wt} и θ_{td}

1. Начальное приближение можно задать нормированными случайными векторами из равномерного распределения.
2. Пройти по всей коллекции, выбрать для каждой пары (d, w) случайную тему t, вычислить частотные оценки вероятностей ϕ_{wt} и θ_{td} для всех d, w, t.

Частичное обучение (некоторые t известны заранее и имеются дополнительные данные о привязке некоторых d или w к t):

- Известно, что документ d относится к подмножеству $T_d \subset T$:

$$\theta_{td}^0 = \frac{1}{W_t} \mathbb{I}_{t \in T_d}.$$

- Известно, что подмножество терминов $W_t \subset W$ относится к теме t:

$$\phi_{td}^0 = \frac{1}{W_t} \mathbb{I}_{w \in W_t}.$$

- Известно, что некоторое множество документов $D_t \subset D$ относится к теме t:

$$\phi_{td}^0 = \frac{\sum_{d \in D} n_{dw}}{\sum_{d \in D_t} n_d}.$$

Недостатки PLSA

1. Медленно сходится на больших коллекциях, так как Φ и Θ обновляются после каждого прохода коллекции.
2. Не разреживает распределение $H_{dwt} = p(t|d, w)$.
3. Вынуждены хранить матрицу $\mathbf{H} = (H_{dwt})_{D \times W \times T}$.
4. Слишком много параметров ϕ_{wt} и θ_{td} $(|W||T| + |T||D|)$.
5. Неверно оценивает вероятность новых слов ($\hat{p}(w|t) = 0$ для слова, которого не было в обучающейся коллекции, но оно встретилось в каком-нибудь документе).
6. Не позволяет управлять разреженностью Φ и Θ:

$$(\text{в начале } \phi_{wt} = 0) \Leftrightarrow (\text{в конце } \phi_{wt} = 0),$$
$$(\text{в начале } \theta_{td} = 0) \Leftrightarrow (\text{в конце } \theta_{td} = 0).$$

Проблема: Вынуждены хранить матрицу $\mathbf{H} = (H_{dwt})_{D \times W \times T}$.

Решение: Вычислять H_{dwt} по мере необходимости.

Algorithm 1 Рациональный EM-алгоритм

Input: Коллекция D, число тем T, начальные $\mathbf{\Phi}$ и $\mathbf{\Theta}$
Output: Распределения $\mathbf{\Phi}$ и $\mathbf{\Theta}$

repeat
  обнулить $\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t$ для всех $d \in D$, $w \in W$, $t \in T$;
  for all $d \in D$, $w \in d$ **do**
    $Z := \sum_{t \in T} \phi_{wt}\theta_{td}$;
    for all $t \in T$ таких, что $\phi_{wt}\theta_{td} > 0$ **do**
      увеличить $\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t$ на $\frac{n_{dw}}{Z}\phi_{wt}\theta_{td}$;
    end for
  end for
  $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D$, $t \in T$;
until $\mathbf{\Phi}$ и $\mathbf{\Theta}$ не стабилизируются;

Проблема:PLSA медленно сходится на больших коллекциях.

Решение: Обновлять значения Φ и Θ чаще.

Algorithm 2 Обобщенный EM-алгоритм

Input: Коллекция D, число тем T, начальные Φ и Θ
Output: Распределения Φ и Θ

Обнулить $\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t$, $\hat{n}_d$, $\hat{n}_{dwt}$ для всех $d \in D$, $w \in W$, $t \in T$;
repeat
  for all $d \in D$, $w \in d$ **do**
    $Z := \sum_{t \in T} \phi_{wt}\theta_{td}$;
    for all $t \in T$ таких, что $n_{dwt} > 0$ или $\phi_{wt}\theta_{td} > 0$ **do**
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$ на $\frac{n_{dw}}{Z}\phi_{wt}\theta_{td} - n_{dwt}$;
      $n_{dwt} := \frac{n_{dw}}{Z}\phi_{wt}\theta_{td} - n_{dwt}$;
    end for
    if не первая итерация и пора обновить параметры Φ и Θ **then**
      $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$ таких, что $\hat{n}_{wt}$ изменился;
      $\theta_{td} := \hat{n_{dt}}/\hat{n}_d$ для всех $d \in D$, $t \in T$ таких, что $\hat{n}_{td}$ изменился;
    end if
  end for
until Φ и Θ не стабилизируются;

Проблема: Необходимо хранить массив $n_{dwt} = n_{dw} H_{dwt}$, который занимает $O(n|T|)$ памяти.

Решение: На M-шаге вместо распределения $H_{dwt} \equiv p(t|d,w)$ взять его несмещенную эмпирическую оценку:

$$\hat{H}_{dwt} = \hat{p}(t|d,w) = \frac{1}{s} \sum_{i=1}^{s} \mathbb{I}_{t_{dwi}=t}.$$

В ряде публикаций предложено экономное сэмплирование, когда s уменьшается до 3–5 тем,что приводит к большему разреживанию и экономии вычислительных ресурсов без существенной потери качества тематической модели.