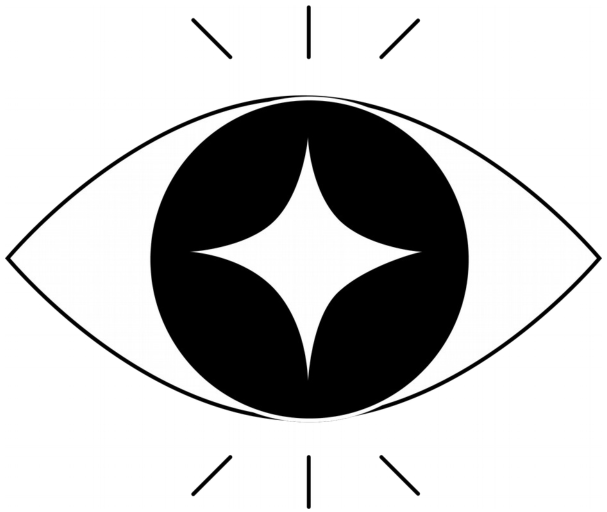
Тематическое моделирование. Введение в ЕМ-алгоритм

На этом уроке

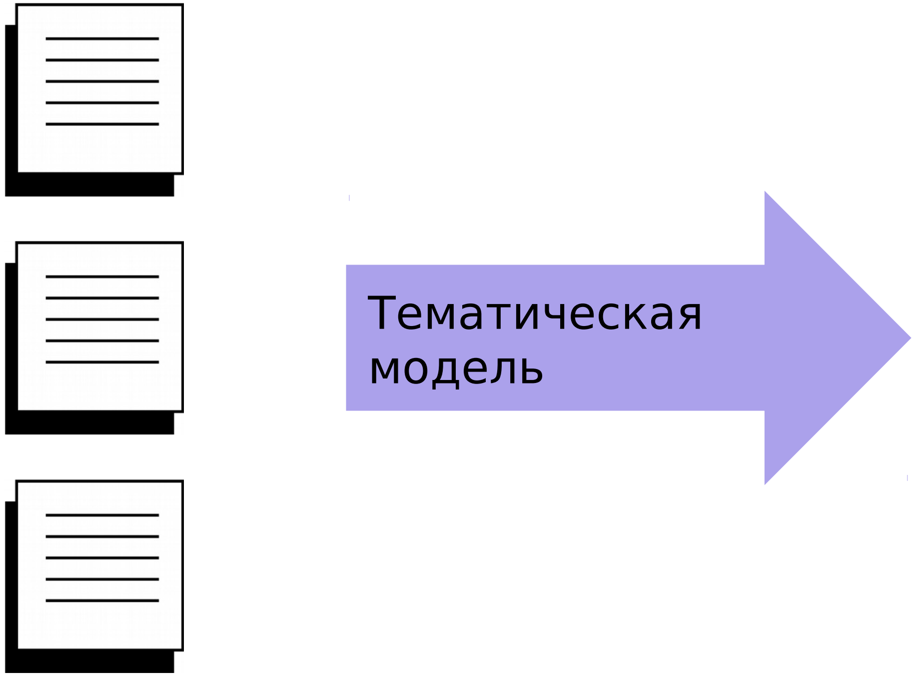
- Постановка задачи;
- LSA
- PLSA;
- EM-алгоритм.



Постановка задачи

Постановка задачи TM

Пусть дано множество (коллекция) текстов. Стоит задача выявления тем, обсуждаемых в них.



Спорт (0,3)
Экономика (0,1)
Политика (0,6)

Культура (0,75)
Туризм (0,15)
Общество (0,1)

Экономика (0,2)
Недвижимость (0,5)
Туризм (0,4)

Где используется ТМ

- Семантический поиск;
- Трендовая аналитика;
- Классификация и категоризация текстов;
- Анализ новостных потоков;
- Суммаризация текстов

Что такое тема

Тема — набор слов или их сочетаний, объединённых принадлежностью к некоторой предметной области. Если тема сформирована качественно, то эксперт в области сможет идентифицировать её смысл по словам.

«Спорт»	«Экономика»	«Политика»
Мяч	Биржа	Выборы
Чемпионат	Нефть	Министр
Хоккей	Производство	Законопроект
Судья	ВВП	Губернатор
Пловец	Министр	Оппозиция

Что такое тема

- Пусть D — коллекция тестовых документов;
- W — словарь из всех уникальных слов этой коллекции;
- Тема — это дискретное вероятностное распределение на множестве W;
- N наиболее вероятных слов в теме используется для её интерпретации.

Тема «Спорт»	Вероятность слов в теме
Мяч	0,1
Чемпионат	0.19
Хоккей	0,17
Судья	0,16
Пловец	0,14
Министр	0,05
Законопроект	0,03
ВВП	0,02
Губернатор	0,01
…	…

Тематическая модель

- Тематическая модель получает на вход коллекцию D и число тем $|T|$;
- При обучении строится два набора вероятностных распределений:
  $p(w|t)$ — распределение слов в теме ($t \in T$)
  $p(t|d)$ — распределение тем в документе ($d \in D$)
- Объединим набора распределений в матрице Φ и Θ:

$$\Phi \in \mathbb{R}^{|W|\times|T|}, \qquad \phi_{wt} = p(w|t)$$
$$\Theta \in \mathbb{R}^{|T|\times|D|}, \qquad \theta_{td} = p(t|d)$$

- Полученные матрицы содержат по столбцам вероятностные распределения и называются стохастическими.

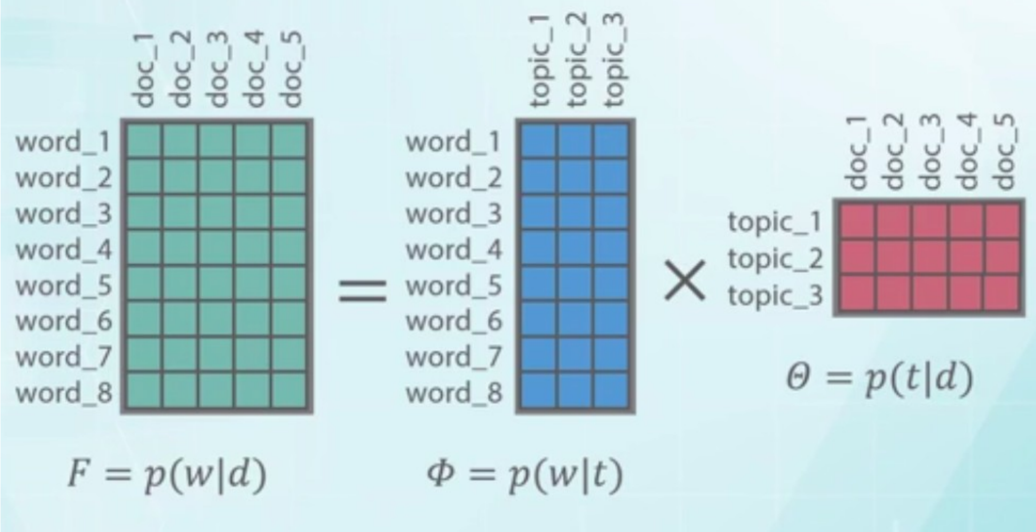
Тематическая модель

$$\Phi \in \mathbb{R}^{|W|\times|T|}, \qquad \phi_{wt} = p(w|t)$$
$$\Theta \in \mathbb{R}^{|T|\times|D|}, \qquad \theta_{td} = p(t|d)$$

Эти матрицы — искомый результат моделирования. Матрица Ф является тематической моделью, матрица Θ — результатом применения этой модели к обучающей коллекции.

ТМ как задача матричного разложения

Постановки задач построения ТМ в разных случаях могут сильно отличаться друг от друга, но по факту интересен результат такого разложения с ограничениями на значения элементов матриц:



Основные выводы

- Тематическое моделирование — методы извлечения из текстов обсуждаемых в нём тематик;
- У тематических моделей есть ряд приложений как в самостоятельном виде, так и в качестве генератора признаков для других моделей;
- Результатом моделирования являются вероятностные векторы тем и документов;
- В базовом варианте задача тематического моделирования сводится к задаче матричного разложения.

Модель PLSA

Определения и предположения

- Пусть D — коллекция тестовых документов;
- W — словарь из всех уникальных слов этой D;
- $|T|$ — множество тем.

- Порядок слов в документе не важен («мешок слов»);
- Порядок слов в коллекции не важен;
- Появление каждого слова в документе связано с некоторой темой $t \in T$:

  «Автор думал о теме t, когда писал это слово»

Определения и предположения

- Коллекция D — это выборка тр $(d_i, w_i, t_i)_{i=1}^{n} \sim p(d,w,t)$:

  $p(d,w,t)$ — распределение в дискретном вероятностном пространстве $D \times W \times T$;
  n_{dw} — число вхождений слова w в документ d;
  $n = \sum_{d \in D} \sum_{w \in d} n_{dw}$ — суммарная длина всех документов коллекции в словах;
  d_i , w_i — наблюдаемые переменные, темы t_i — скрытые.

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

  «Появление в документе слова, связанного с темой, зависит от темы и не зависит от документа».

Вероятностная модель коллекции текстов

- С учётом гипотезы условной независимости запишем вероятностную модель:

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

- Она описывает процесс появления слов в документах темами. Пусть у нас есть распределения *p(w|t)* для каждой темы и *p(t|d)* — для каждого документа
- Можно сгенерировать каждое слово каждого документа:

  Для позиции в документе *d* генерируем тему *t* из *p(t|d);*

  Затем генерируем слово *w* из *p(w|t).*

Probabilistic Latent Semantic Analysis

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

Тематическое моделирование решает обратную задачу — по документам восстановить параметры генерации.

PLSA (вероятностный латентно-семантический анализ) — исторически первая тематическая модель. Вспомним обозначения:

$$\Phi \in \mathbb{R}^{|W|\times|T|}, \qquad \phi_{wt} = p(w|t)$$
$$\Theta \in \mathbb{R}^{|T|\times|D|}, \qquad \theta_{td} = p(t|d)$$

Решаем задачу стохастического матричного разложения для фиксированного числа тем $[T]$.

Оптимизационная задача PLSA

Воспользуемся методом максимального правдоподобия. Ищем такие параметры модели, чтобы они наилучшим образом опис $(d_i, w_i)_{i=1}^n$ данные

Функционал правдоподобия:

$$\prod_{i=1}^{n} p(d_i, w_i) = \prod_{i=1}^{n} p(w_i|d_i)p(d_i) =$$

$$= \prod_{d \in D} \prod_{w \in W} p(w|d)^{n_{dw}} p(d)^{n_{dw}}$$

Прологарифмируем выражение и выбросим константы:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln p(w|d) = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi,\Theta}$$

Оптимизационная задача PLSA

Воспользуемся методом максимального правдоподобия:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

Добавим ограничения стохастичности на матрицы:

$$\sum_{w \in W} \phi_{wt} = 1, \qquad \phi_{wt} \geq 0$$

$$\sum_{t \in T} \theta_{td} = 1, \qquad \theta_{td} \geq 0$$

Получили итоговую оптимизационную задачу PLSA. Она невыпуклая, можем найти только локальный экстремум. Оптимизировать напрямую сложно.

Основные выводы

- PLSA — первая и наиболее простая модель в тематическом моделировании;
- Разрешает задачу стохастического матричного разложения методом максимального правдоподобия.

ЕМ-алгоритм

Оптимизационная задача PLSA

Воспользуемся методом максимального правдоподобия:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta} \qquad (1)$$

Добавим ограничения стохастичности на матрицы:

$$\left.\begin{array}{ll} \sum_{w \in W} \phi_{wt} = 1, & \phi_{wt} \geq 0 \\ \sum_{t \in T} \theta_{td} = 1, & \theta_{td} \geq 0 \end{array}\right\} \quad (2)$$

Получили итоговую оптимизационную задачу PLSA. Она невыпуклая, можем найти только локальный экстремум.

Идея ЕМ-алгоритма

Есть функционал (1) с ограничениями (2), который сложно напрямую оптимизировать по параметрам Φ и Θ. Введём вспомогательные переменные $p_{tdw}=p(t|d,w)$. Можем вычислять эти переменные через параметры.

Используем формулу Байеса и гипотезу условной зависимости:

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

Идея ЕМ-алгоритма

Есть функционал (1) с ограничениями (2), который сложно напрямую оптимизировать по параметрам Φ и Θ. Введём вспомогательные переменные $p_{tdw}=p(t|d,w)$. Можем вычислять эти переменные через параметры, а можем вычислять параметры через вспомогательные переменные.

Организуем итеративный процесс!

ЕМ-алгоритм для PLSA

Теорема: точка (Φ, Θ) локального экстремума задачи (1) с ограничениями (2) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}=p(t|d,w)$, если из решения исключить нулевые столбцы Φ и Θ:

Е-шаг

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}(\phi_{wt}\theta_{td})$$

М-шаг

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}(n_{wt}), \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}(n_{td}), \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}$$

где $\underset{i\in I}{\operatorname{norm}}(x_i) = \dfrac{\max\{x_i, 0\}}{\sum_{j\in I}\max\{x_j, 0\}}$

Доказательство основано на теореме ККТ.

ЕМ-алгоритм для PLSA

1. Получаем итеративный алгоритм решения задачи;
2. Сперва задаются стартовые значения Φ и Θ;
3. По ним на Е-шаге вычисляются вспомогательные переменные p_{tdw};
4. По этим переменным на М-шаге вычисляется новая версия параметров Φ и Θ;
5. Этот процесс продолжается до сходимости оптимизируемого функционала правдоподобия;

Конкретное значение локального экстремума сильно зависит от начального приближения Φ и Θ.

Метрики качества тематических моделей

Перплексия — функция правдоподобия модели, характеризует её сходимость (чем ниже, тем лучше):

$$\mathcal{P}(D;\Phi,\Theta) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d)\right)$$

Метрики качества тематических моделей

Средняя когерентность — автоматическая мера интерпретируемости тем модели (чем выше, тем лучше).

$$\mathcal{C}(\Phi) = \frac{1}{|T|}\sum_{t\in T}\mathcal{C}_t(\Phi), \qquad \mathcal{C}_t(\Phi) = \frac{2}{k(k-1)}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\mathrm{PPMI}(w_i, w_j)$$

w_i — i-е слово в списке из k наиболее вероятных слов в распределении темы t.

$$\mathrm{PPMI}(u, v) = \max\left\{0, \ln\frac{|D|N_{uv}}{N_u N_v}\right\}$$

N_{uv} — число документов, содержащих оба слова u и v в окне заданного размера.
N_u — число документов, содержащих слово u в окне заданного размера.

Основные выводы

- Для оптимизации PLSA используется итеративный EM-алгоритм;
- Сходимость гарантируется только к локальному оптимуму и сильно зависит от начального сближения параметров;
- Базовыми метриками качества тематических моделей являются перплексия и средняя когерентность тем.