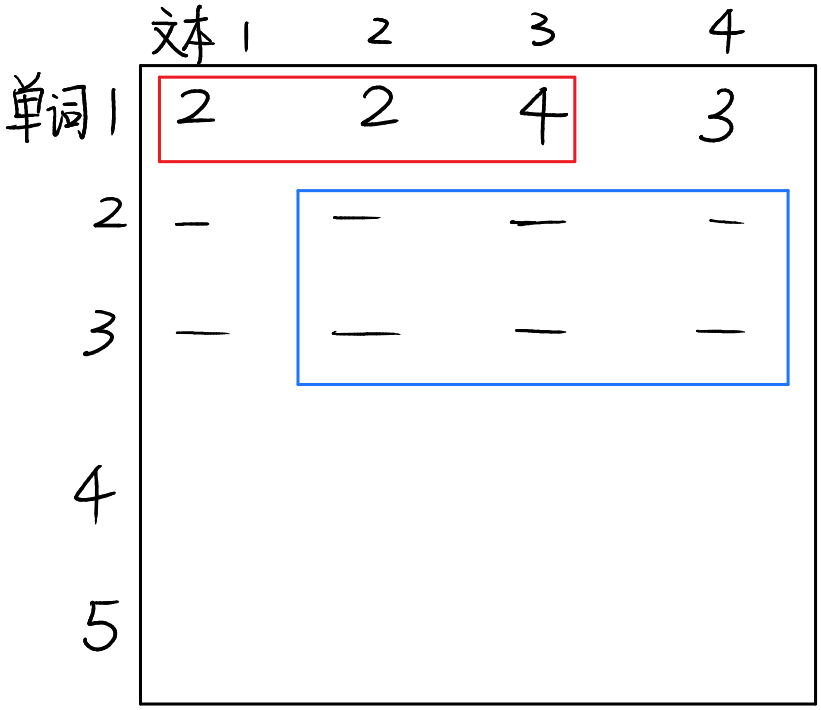
概率潜在语义分析

1. 基本想法

一个文本(document)集合包含多个话题，一个话题由若干单词组成，话题无法直接观测需要数据估计。

目标：发现隐变量表示的话题，即潜在语义



语义相近的单词，语义相近的文本会被聚向相同的"软的类别"中

↓

话题

2. 生成模型

数据

(1) 单词集合 $W=\{w_1, w_2, \cdots, w_M\}$

(2) 文本集合 $D=\{d_1, d_2, \cdots, d_N\}$

(3) 话题集合 $Z=\{z_1, z_2, \cdots, z_K\}$

观测数据(word-doc)矩阵

	$\Lambda(w,d)$	

d → z → w

① $\xrightarrow{P(d)}$ 文本d

② $\xrightarrow{P(w|d)}$ 话题z

③ $\xrightarrow{P(w|z)}$ 单词w

概率表示

单词文本矩阵T的生成概率

$$P(T)=\prod_{(w,d)} P(w,d)^{\Lambda(w,d)}$$

$$P(w,d)=P(d)P(w|d)=P(d)\sum_z P(w,z|d)$$

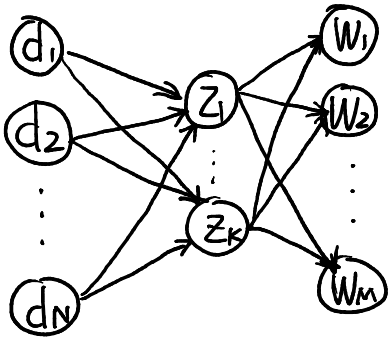
$$=P(d)\sum_z P(z|d)P(w|z)$$

3. 模型参数

(1) 如果认为未知参数是P(w,d)　M×N

(2) 参数个数 → 引入话题　O(M×K+K×N)

其中K<<M,N, 大大减少待估参数



4. 概率潜在语义模型求解

对数似然函数

$$L(\theta)=\sum_{i=1}^{M}\sum_{j=1}^{N}\Lambda(w_i,d_j)\log P(w_i,d_j)$$

$$=\sum_{i=1}^{M}\sum_{j=1}^{N}\Lambda(w_i,d_i)\log\{P(d_j)\sum_{K}P(w_i|z_k)P(z_k|d_j)\}$$

$r_{ijk}=1$，代表d_j产生了话题z_k，z_k产生了w_i

$$P(w_i|z_k)P(z_k|d_j)$$

· E步：（计算Q函数）

$$Q(\theta,\theta^{(i)})=\sum_{z}\log P(Y,z|\theta)\cdot P(z|Y,\theta^{(i)})$$

完全样本似然函数

$$P(w,d,r|\theta)=\prod_{i=1}^{N}\prod_{j=1}^{M}P(w_i,d_j,r_{ij1},r_{ij2},\cdots,r_{ijK})$$

$$=\prod_{i=1}^{N}\prod_{j=1}^{M}\left[\prod_{k=1}^{K}\{P(w_i|z_k)P(z_k|d_j)\}^{r_{ijk}}\cdot P(d_j)\right]^{\Lambda(w_i,d_j)}$$

$$\log P(w,d,z|\theta)=\sum_{i=1}^{N}\sum_{j=1}^{M}\Big[\Lambda(w_i,d_j)\Big\{\log P(d_j)$$

$$+\sum_{K=1}^{K}r_{ijk}\log P(w_i|z_k)\cdot P(z_k|d_j)\Big\}\Big]$$

$$r_{ijk}^{(t)} = E[r_{ijk} | W, d, \theta^{(t)}]$$

$$= P(z_k | W, d, \theta) = \frac{P(z_k, w_i | d_j)}{P(w_i | d_j)}$$

$$= \frac{P(w_i | z_k) P(z_k | d_j)}{\sum_k P(w_i | z_k) P(z_k | d_j)}$$，$P(w_i | z_k) P(z_k | d_j)$用上一步迭代代入

M步：

$$r_{ijk}^{(t)} = E\{r_{ijk} | w, d, \theta^{(t)}\}$$

$$Q(\theta, \theta^{(i)}) = \sum_{k=1}^{K} \sum_{i=1}^{M} \sum_{j=1}^{N} r_{ijk}^{(t)} \Delta(w_i, d_j) \log\{P(w_i | z_k) P(z_k | d_j)\}$$

约束：$\sum_{i=1}^{M} P(w_i | z_k) = 1$, for $k = 1, 2, \cdots, K$

$\sum_{k=1}^{K} P(z_k | d_j) = 1$, for $j = 1, 2, \cdots, M$

定义Lagrange函数：

$$\mathcal{L}(\theta) = Q(\theta, \theta^{(i)}) + \sum_{k=1}^{K} z_k \{1 - \sum_{i=1}^{M} P(w_i | z_k)\} + \sum_{j=1}^{N} d_j \{1 - \sum_{k=1}^{K} P(z_k | d_j)\}$$

$$P(w_i | z_k) = \frac{\sum_{j=1}^{N} \Delta(w_i, d_j) r_{ijk}}{\sum_{i=1}^{M} \sum_{j=1}^{N} \Delta(w_i, d_j) r_{ijk}} \qquad P(z_k | d_j) = \frac{\sum_{i=1}^{M} \Delta(w_i, d_j) r_{ijk}}{\sum_{k=1}^{K} \sum_{i=1}^{M} \Delta(w_i, d_j) r_{ijk}}$$