主题模型

1. 给定语料库 $\mathbb{D}=\{\mathcal{D}_1,\mathcal{D}_2,\cdots,\mathcal{D}_N\}$ ，其中包含 N 篇文档 。

   所有的单词来自于词汇表 $\mathbb{V}=\{\text{word}_1,\text{word}_2,\cdots,\text{word}_V\}$，其中 V 表示词汇表的大小。

2. `BOW:Bag of Words`：词在文档中不考虑顺序，这称作词袋模型。

一、Unigram Model

1. 假设有一个骰子，骰子有 V 个面，每个面对应于词典中的一个单词。`Unigram Model`是这样生成文档的：
   - 每次抛一次骰子，抛出的面就对应于产生一个单词
   - 如果一篇文档有 n 个单词，则独立的抛掷 n 次骰子就产生着 n 个单词。
2. 令骰子的投掷出各个面的概率为：

$$\vec{\Theta}=(\theta_1,\theta_2,\cdots,\theta_V)^T$$
$$\sum_{v=1}^{V}\theta_v=1$$

   即 $P(\text{word}_v)=\theta_v$ 。$\vec{\Theta}$ 就是待求的参数。

3. 假设文档包含 n 个单词，这些单词依次为： $\{\text{word}_{w_1},\text{word}_{w_2},\cdots,\text{word}_{w_n}\}$，其中 $w_i\in\{1,2,\cdots,V\}$，用 $(w_1,w_2,\cdots,w_n)$ 代表文档。 则生成这篇文档的概率为：

$$p(w_1,w_2,\cdots,w_n;\vec{\Theta})=p(w_1;\vec{\Theta})\prod_{i=2}^{n}p(w_i\mid w_1,w_2,\cdots,w_{i-1};\vec{\Theta})$$

   在 $p(w_i\mid w_1,w_2,\cdots,w_{i-1};\vec{\Theta})$ 中， $w_1,w_2,\cdots,w_{i-1}$ 是 w_i 的上下文。

   由于采取的是词袋模型，没有考虑上下文，所以有：

$$p(w_i\mid w_1,w_2,\cdots,w_{i-1};\vec{\Theta})=p(w_i;\vec{\Theta})$$

   于是有：

$$p(w_1,w_2,\cdots,w_n;\vec{\Theta})=\prod_{i=1}^{n}p(w_i;\vec{\Theta})$$

   - 如果考虑了上下文（即抛弃词袋模型），则各种单词的组合会导致爆炸性的复杂度增长。
   - 由于是词袋模型，因此 $p(w_1,w_2,\cdots,w_n;\vec{\Theta})$ 并不构成一个概率分布。

     $p(w_1,w_2,\cdots,w_n;\vec{\Theta})$ 仅仅是生成该文档的一种非归一化概率。

4. 假设单词 $\{\text{word}_{w_1},\text{word}_{w_2},\cdots,\text{word}_{w_n}\}$ 中，有 $\tilde{n}_1$ 个 word_1，有 $\tilde{n}_2$ 个 word_2，...有 $\tilde{n}_V$ 个 word_V ，其中 $\tilde{n}_1+\tilde{n}_2+\cdots+\tilde{n}_V=n$，则：

$$p(w_1,w_2,\cdots,w_n;\vec{\Theta})=\prod_{i=1}^{n}p(w_i;\vec{\Theta})=\prod_{v=1}^{V}P(\text{word}_v)^{\tilde{n}_v}=\prod_{v=1}^{V}\theta_v^{\tilde{n}_v}$$

5. 参数估计：就是估计骰子的投掷出各个面的概率 $\vec{\Theta}=(\theta_1,\theta_2,\cdots,\theta_V)^T$

1.1 最大似然估计

1. 假设数据集 $\mathbb{D}$ 包含 N 篇文档 $\mathbb{D}=\{\mathcal{D}_1,\mathcal{D}_2,\cdots,\mathcal{D}_N\}$。

   对文档 $\mathcal{D}_i$，假设其单词依次为 $\{\text{word}_{w_1^i},\text{word}_{w_2^i},\cdots,\text{word}_{w_{n_i}^i}\}$， 用 $(w_1^i,w_2^i,\cdots,w_{n_i}^i)$ 来表示。其中：

   - $v=w_j^i$ 表示文档 $\mathcal{D}_i$ 的第 j 个单词为单词 word_v 。
   - n_i 表示文档 $\mathcal{D}_i$ 一共有 n_i 个单词。

2. 由于每篇文档都是独立的且不考虑文档的顺序和单词的顺序，则数据集发生的概率

$$L=p(\mathbb{D})=p(w_1^1,\cdots,w_{n_1}^1,\cdots,w_1^N,\cdots,w_{n_N}^N;\vec{\Theta})=\prod_{i=1}^{N}\prod_{j=1}^{n_i}p(w_j^i;\vec{\Theta})$$

   假设单词 $\{\text{word}_{w_1^1},\text{word}_{w_2^1},\cdots,\text{word}_{w_{n_1}^1},\cdots,\text{word}_{w_1^N},\text{word}_{w_2^N},\cdots,\text{word}_{w_{n_N}^N}\}$ 中，有 $\tilde{n}_1$ 个 word_1，有 $\tilde{n}_2$ 个 word_2，...有 $\tilde{n}_V$ 个 word_V 。其中 $\tilde{n}_1+\tilde{n}_2+\cdots+\tilde{n}_V=n$， n 为所有文档的所有单词的数量。则有：

$$L=\prod_{i=1}^{N}\prod_{j=1}^{n_i}p(w_j^i;\vec{\Theta})=\prod_{v=1}^{V}P(\text{word}_v)^{\tilde{n}_v}=\prod_{v=1}^{V}\theta_v^{\tilde{n}_v}$$

3. 使用最大似然估计法，也就是最大化对数的 L：

$$LL=\log L=\log\prod_{v=1}^{V}\theta_v^{\tilde{n}_v}=\sum_{v=1}^{V}\tilde{n}_v\log\theta_v$$

   于是求解：

$$\arg\max_{\vec{\Theta}}\sum_{v=1}^{V}\tilde{n}_v\log\theta_v$$

$$s.t.\quad\sum_{v=1}^{V}\theta_v=1;\quad\theta_1,\theta_2,\cdots,\theta_V\geq 0$$

   用拉格朗日乘子法求解，其解为：

$$\hat{\theta}_v=\frac{\tilde{n}_v}{n},v=1,2,\cdots,V$$

   其物理意义为：单词 word_v 出现的概率 θ_v 等于它在数据集 $\mathcal{D}$ 中出现的频率（它出现的次数 $\tilde{n}_v$ 除以文档所有单词数 n ）。

1.2 最大后验估计

1. 根据贝叶斯学派的观点， 参数 $\vec{\Theta}$ 也是一个随机变量而不再是一个常量，它服从某个概率分布 $p(\vec{\Theta})$， 这个分布称作参数 $\vec{\Theta}$ 的先验分布。

   此时：

$$p(\mathbb{D})=p(w_1^1,\cdots,w_{n_1}^1,\cdots,w_1^N,\cdots,w_{n_N}^N)=\int p(w_1^1,\cdots,w_{n_1}^1,\cdots,w_1^N,\cdots,w_{n_N}^N,\vec{\Theta})d\vec{\Theta}$$

$$=\int p(w_1^1,\cdots,w_{n_1}^1,\cdots,w_1^N,\cdots,w_{n_N}^N\mid\vec{\Theta})p(\vec{\Theta})d\vec{\Theta}$$

   根据前面的推导有： $p(w_1^1,\cdots,w_{n_1}^1,\cdots,w_1^N,\cdots,w_{n_N}^N\mid\vec{\Theta})=\prod_{v=1}^{V}\theta^{\tilde{n}_v}$ ，则有：

$$p(\mathbb{D}) = p(w_1^1, \cdots, w_{n_1}^1, \cdots, w_1^N, \cdots, w_{n_N}^N) = \int \prod_{v=1}^{V} \theta^{\tilde{n}_v} p(\vec{\Theta}) d\vec{\Theta}$$

2. 此处先验分布 $p(\vec{\Theta})$ 有多种选择。

   注意到数据集条件概率 $p(w_1^1, \cdots, w_{n_1}^1, \cdots, w_1^N, \cdots, w_{n_N}^N \mid \vec{\Theta})$ 刚好是多项式分布的形式，于是选择先验分布为多项式分布的共轭分布，即狄利克雷分布：

$$\vec{\Theta} \sim Dir(\vec{\alpha}) : p(\vec{\Theta}; \vec{\alpha}) = \frac{1}{B(\vec{\alpha})} \prod_{v=1}^{V} \theta_v^{\alpha_v - 1}$$

   其中：

   - $\vec{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_V)^T$ 为参数向量
   - $B(\vec{\alpha})$ 为 `Beta` 函数：

$$B(\vec{\alpha}) = \frac{\prod_{v=1}^{V} \Gamma(\alpha_v)}{\Gamma(\sum_{v=1}^{V} \alpha_v)}$$

   - 显然根据定义有：

$$\int p(\vec{\Theta}; \vec{\alpha}) d\vec{\Theta} = \int \frac{1}{B(\vec{\alpha})} \prod_{v=1}^{V} \theta_v^{\alpha_v - 1} d\vec{\Theta} = 1 \longrightarrow \int \prod_{v=1}^{V} \theta_v^{\alpha_v - 1} d\vec{\Theta} = B(\vec{\alpha})$$

3. 令 $\vec{\mathbf{n}} = (\tilde{n}_1, \tilde{n}_2, \cdots, \tilde{n}_V)^T$ 为词频向量，其每个元素代表了对应的单词在数据集 $\mathbb{D}$ 中出现的次数。

   此时有：

$$\begin{aligned} p(\mathbb{D}) = p(w_1^1, \cdots, w_{n_1}^1, \cdots, w_1^N, \cdots, w_{n_N}^N) &= \int \prod_{v=1}^{V} \theta_v^{\tilde{n}_v} p(\vec{\Theta}) d\vec{\Theta} \\ &= \int \prod_{v=1}^{V} \theta_v^{\tilde{n}_v} \frac{1}{B(\vec{\alpha})} \prod_{v=1}^{V} \theta_v^{\alpha_v - 1} d\vec{\Theta} \\ &= \frac{1}{B(\vec{\alpha})} \int \prod_{v=1}^{V} \theta_v^{\tilde{n}_v + \alpha_v - 1} d\vec{\Theta} \\ &= \frac{B(\vec{\alpha} + \vec{\mathbf{n}})}{B(\vec{\alpha})} \end{aligned}$$

   因此 $p(\mathbb{D})$ 仅由 $\vec{\alpha}$ 决定，记作：$p(\mathbb{D}) = \frac{B(\vec{\alpha} + \vec{\mathbf{n}})}{B(\vec{\alpha})}$

4. 后验概率：

$$\begin{aligned} p(\vec{\Theta} \mid w_1^1, \cdots, w_{n_1}^1, \cdots, w_1^N, \cdots, w_{n_N}^N; \vec{\alpha}) &= \frac{p(w_1^1, \cdots, w_{n_1}^1, \cdots, w_1^N, \cdots, w_{n_N}^N \mid \vec{\Theta}) p(\vec{\Theta})}{p(w_1^1, \cdots, w_{n_1}^1, \cdots, w_1^N, \cdots, w_{n_N}^N; \vec{\alpha})} \\ &= \prod_{v=1}^{V} \theta_v^{\tilde{n}_v} \frac{1}{B(\vec{\alpha})} \prod_{v=1}^{V} \theta_v^{\alpha_v - 1} \frac{B(\vec{\alpha})}{B(\vec{\alpha} + \vec{\mathbf{n}})} \\ &= \frac{1}{B(\vec{\alpha} + \vec{\mathbf{n}})} \prod_{v=1}^{V} \theta_v^{\tilde{n}_v + \alpha_v - 1} \end{aligned}$$

   可见后验概率服从狄利克雷分布 $Dir(\vec{\alpha} + \vec{\mathbf{n}})$ 。

5. 因为这时候的参数 $\vec{\Theta}$ 是一个随机变量，而不再是一个固定的数值，因此需要通过对后验概率 $p(\vec{\Theta} \mid \mathbb{D};\vec{\alpha})$ 最大化或者期望来求得。

   这里使用期望值 $\mathbb{E}(\vec{\Theta} \mid \mathbb{D};\vec{\alpha})$ 来做参数估计。

   由于后验分布 $p(\vec{\Theta} \mid \mathbb{D};\vec{\alpha})$ 服从狄利克雷分布 $Dir(\vec{\alpha}+\vec{\tilde{\mathbf{n}}})$， 则有期望：

$$\mathbb{E}(\vec{\Theta} \mid \mathbb{D};\vec{\alpha})=\left(\frac{\tilde{n}_1+\alpha_1}{\sum_{v=1}^{V}(\tilde{n}_v+\alpha_v)},\frac{\tilde{n}_2+\alpha_2}{\sum_{v=1}^{V}(\tilde{n}_v+\alpha_v)},\cdots,\frac{\tilde{n}_V+\alpha_V}{\sum_{v=1}^{V}(\tilde{n}_v+\alpha_v)}\right)$$

   即参数 θ_v 的估计值为：

$$\hat{\theta}_v=\frac{\tilde{n}_v+\alpha_v}{\sum_{v=1}^{V}(\tilde{n}_v+\alpha_v)}$$

   考虑到 α_v 在狄利克雷分布中的物理意义为：事件的先验的伪计数。因此该估计式物理意义为：估计值是对应事件计数（伪计数+真实计数）在整体计数中的比例。

1.3 文档生成

1. `Unigram Model` 生成文档的步骤为：
   - 根据参数为 $\vec{\alpha}$ 的狄利克雷分布 $p(\vec{\Theta};\vec{\alpha})\sim Dir(\vec{\alpha})$ 随机采样一个词汇分布 $\vec{\tilde{\Theta}}=(\tilde{\theta}_1,\tilde{\theta}_2,\cdots,\tilde{\theta}_V)^T$。

     所谓随机采样一个词汇分布，即：根据狄里克雷分布生成一个随机向量。选择时要求： $$

$$\sum_{v=1}^{V}\tilde{\theta}_v=1$$

$$\tilde{\theta}_v\geq 0,\quad v=1,2,\cdots,V$$

   - 根据词汇分布 $\vec{\tilde{\Theta}}$：
     - 根据该分布从词汇表中独立重复采样 n 次，获取 n 个单词。则这些单词就组成一篇文档。
     - 重复 N 次，即得到 N 篇文档组成的文档集合 。

   > 所有文档共享同一个词汇分布 $\vec{\tilde{\Theta}}$，而不是每个文档各自采样一个词汇分布。

二、pLSA Model

1. `Unigram Model` 模型过于简单。事实上人们写一篇文章往往需要先确定要写哪几个主题。

   如：写一篇计算机方面的文章，最容易想到的词汇是：内存、CPU、编程、算法等等。之所以能马上想到这些词，是因为这些词在对应的主题下出现的概率相对较高。

   因此可以很自然的想到：一篇文章通常由多个主题构成，而每个主题大概可以用与该主题相关的频率最高的一些词来描述。

   上述直观的想法由 `Hoffman` 在 1999 年的 `probabilistic Latent Semantic Analysis:pLSA` 模型中首先进行了明确的数学化。

2. 主题 `topic`：表示一个概念。具体表示为一系列相关的词，以及它们在该概念下出现的概率。
   - 与某个主题相关性比较强的词，在该主题下出现概率较高
   - 与某个主题相关性比较弱的词，在该主题下出现概率较低
3. 主题示例：给定一组词： `证明,推导,对象,酒庄,内存`，下列三个主题可以表示为：
   - 数学主题：$(0.45,0.35,0.2,0,0)$

- 计算机主题：$(0.2, 0.15, 0.45, 0, 0.2)$
- 红酒主题：$(0, 0, 0.2, 0.8, 0)$

	证明	推导	对象	酒庄	内存
数学	0.45	0.35	0.2	0	0
计算机	0.2	0.15	0.45	0	0.2
红酒	0	0	0.2	0.8	0

2.1 文档生成

1. 假设话题集合 $\mathbb{T}$ 有 T 个话题，分别为 $\mathbb{T} = \{\text{topic}_1, \text{topic}_2, \cdots, \text{topic}_T\}$ 。

   `pLSA` 模型的文档生成规则：

   - 首先以概率 $p(\mathcal{D}_i)$ 选中第 i 篇文档。
   - 然后在文档 $\mathcal{D}_i$ 中，以概率 $p(\text{topic}_t \mid \mathcal{D}_i)$ 选中第 t 个话题 topic_t 。
   - 然后在话题 topic_t 中，以概率 $p(\text{word}_v \mid \text{topic}_t)$ 选中第 v 个单词 word_v 。
   - 重复执行`挑选话题--> 挑选单词` n 次，则得到一篇包含 n 个单词 $\{\text{word}_{w_1}, \text{word}_{w_2}, \cdots, \text{word}_{w_n}\}$ 的文档。

     其中：$1 \le w_1, \cdots, w_n \le V$ ；$v = w_j$ 表示文档的第 j 个单词为 word_v 。

2. 重复执行上述文档生成规则 N 次，即得到 N 篇文档组成的文档集合 $\mathbb{D}$ 。

2.2 模型原理

1. 令

$$\varphi_{i,t} = p(\text{topic}_t \mid \mathcal{D}_i), i = 1, 2, \cdots, N; t = 1, 2, \cdots, T$$
$$\theta_{t,v} = p(\text{word}_v \mid \text{topic}_t), v = 1, 2, \cdots, V; t = 1, 2, \cdots, T$$

   - $\varphi_{i,t}$ 表示：选中第 i 篇文档 $\mathcal{D}_i$ 的条件下，选中第 t 个话题 topic_t 的概率
   - $\theta_{t,v}$ 表示：选中第 t 个话题 topic_t 的条件下，选中第 v 个单词 word_v 的概率

   待求的是参数 Φ 和 Θ：

$$\Phi = \{\varphi_{i,t}\} = \begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} & \cdots & \varphi_{1,T} \\ \varphi_{2,1} & \varphi_{2,2} & \cdots & \varphi_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{N,1} & \varphi_{N,2} & \cdots & \varphi_{N,T} \end{bmatrix} \quad \Theta = \{\theta_{t,v}\} = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,V} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,V} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{T,1} & \theta_{T,2} & \cdots & \theta_{T,V} \end{bmatrix}$$

2. 根据概率的定义，有约束条件：

$$\sum_{t=1}^{T} \varphi_{i,t} = 1, \quad i = 1, 2, \cdots, N;$$
$$\sum_{v=1}^{V} \theta_{t,v} = 1, \quad t = 1, 2, \cdots, T;$$
$$\varphi_{i,t} \ge 0, \quad \theta_{t,v} \ge 0, \quad i = 1, 2, \cdots, N; \quad t = 1, 2, \cdots, T; \quad v = 1, 2, \cdots, V$$

3. 根据 pLSA 概率图模型（由盘式记法得到），结合成对马尔可夫性有：

$$p(\text{word}_v, \mathcal{D}_i \mid \text{topic}_t) = p(\text{word}_v \mid \text{topic}_t)p(\mathcal{D}_i \mid \text{topic}_t)$$

即：文档和单词关于主题条件独立。

4. 对于给定文档 $\mathcal{D}_i$ 中的单词 word_v，有：

$$\begin{aligned}
p(\mathcal{D}_i, \text{word}_v) &= \sum_{t=1}^{T} p(\mathcal{D}_i, \text{word}_v, \text{topic}_t) \\
&= \sum_{t=1}^{T} p(\mathcal{D}_i, \text{word}_v \mid \text{topic}_t)p(\text{topic}_t) \\
&= \sum_{t=1}^{T} p(\mathcal{D}_i \mid \text{topic}_t)p(\text{word}_v \mid \text{topic}_t)p(\text{topic}_t) \\
&= \sum_{t=1}^{T} p(\mathcal{D}_i, \text{topic}_t)p(\text{word}_v \mid \text{topic}_t) \\
&= \sum_{t=1}^{T} p(\text{topic}_t \mid \mathcal{D}_i)p(\mathcal{D}_i)p(\text{word}_v \mid \text{topic}_t) \\
&= p(\mathcal{D}_i) \sum_{t=1}^{T} p(\text{topic}_t \mid \mathcal{D}_i)p(\text{word}_v \mid \text{topic}_t)
\end{aligned}$$

根据该等式，可以得到：

$$p(\text{word}_v \mid \mathcal{D}_i) = \sum_{t=1}^{T} p(\text{topic}_t \mid \mathcal{D}_i)p(\text{word}_v \mid \text{topic}_t)$$

即：给定文档 $\mathcal{D}_i$ 的条件下，某个的单词 word_v 出现的概率可以分成三步：

- 首先得到给定的文档 $\mathcal{D}_i$ 的条件下，获取某个话题 topic_t 的概率
- 再得到该话题 topic_t 生成该单词 word_v 的概率
- 对所有的话题累加 $\sum_{t=1}^{T}$ 即得到给定的单词 word_v 在给定文档 $\mathcal{D}_i$ 中出现的概率

5. 对于给定文档 $\mathcal{D}_i$ 中主题 topic_t 生成的单词 word_v，有：

$$\begin{aligned}
p(\mathcal{D}_i, \text{topic}_t, \text{word}_v) &= p(\mathcal{D}_i)p(\text{word}_v, \text{topic}_t \mid \mathcal{D}_i) \\
&= p(\mathcal{D}_i)p(\text{word}_v \mid \text{topic}_t, \mathcal{D}_i)p(\text{topic}_t \mid \mathcal{D}_i) \\
&= p(\mathcal{D}_i)p(\text{topic}_t \mid \mathcal{D}_i)\frac{p(\text{word}_v, \mathcal{D}_i \mid \text{topic}_t)}{p(\mathcal{D}_i \mid \text{topic}_t)} \\
&= p(\mathcal{D}_i)p(\text{topic}_t \mid \mathcal{D}_i)p(\text{word}_v \mid \text{topic}_t)
\end{aligned}$$

则已知文档 $\mathcal{D}_i$ 中出现了单词 word_v 的条件下，该单词由主题 topic_t 生成的概率为：

$$\begin{aligned}
p(\text{topic}_t \mid \mathcal{D}_i, \text{word}_v) &= \frac{p(\mathcal{D}_i, \text{word}_v, \text{topic}_t)}{p(\mathcal{D}_i, \text{word}_v)} \\
&= \frac{p(\mathcal{D}_i)p(\text{topic}_t \mid \mathcal{D}_i)p(\text{word}_v \mid \text{topic}_t)}{p(\mathcal{D}_i)\sum_{t'=1}^{T} p(\text{topic}_{t'} \mid \mathcal{D}_i)p(\text{word}_v \mid \text{topic}_{t'})} \\
&= \frac{p(\text{topic}_t \mid \mathcal{D}_i)p(\text{word}_v \mid \text{topic}_t)}{\sum_{t'=1}^{T} p(\text{topic}_{t'} \mid \mathcal{D}_i)p(\text{word}_v \mid \text{topic}_{t'})}
\end{aligned}$$

- 其物理意义为：给定文档 $\mathcal{D}_i$ 的条件下，单词 word_v 由主题 topic_t 生成的概率占单词 word_v 出现的概率的比例。
- 若话题 topic_{10} 仅仅与单词 word_1 相关，而且文档 $\mathcal{D}_3$ 中刚好出现了单词 word_1 ，则说明文档 $\mathcal{D}_3$ 由话题 topic_{10} 组成的概率很高。

2.3 参数求解

1. `pLSA` 模型由两种参数求解方法：
   - 矩阵分解
   - `EM` 算法

2.3.1 矩阵分解

1. 根据前面的推导，有：$p(\text{word}_v \mid \mathcal{D}_i) = \sum_{t=1}^{T} p(\text{topic}_t \mid \mathcal{D}_i) p(\text{word}_v \mid \text{topic}_t)$ 。

   其中文档 $\mathcal{D}_i$ 和单词 word_v 是观测到的，主题 topic_t 是未观测到的、未知的。

   令 $p_{i,v}^{\mathcal{D}} = p(\text{word}_v \mid \mathcal{D}_i)$ ，根据：

   $$\varphi_{i,t} = p(\text{topic}_t \mid \mathcal{D}_i), \quad \theta_{t,v} = p(\text{word}_v \mid \text{topic}_t)$$

   则有：

   $$p_{i,v}^{\mathcal{D}} = \sum_{t=1}^{T} \varphi_{i,t} \theta_{t,v}$$

2. 令：

   $$\mathbf{P}^{\mathcal{D}} = \begin{bmatrix} p_{1,1}^{\mathcal{D}} & p_{1,2}^{\mathcal{D}} & \cdots & p_{1,V}^{\mathcal{D}} \\ p_{2,1}^{\mathcal{D}} & p_{2,2}^{\mathcal{D}} & \cdots & p_{2,V}^{\mathcal{D}} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N,1}^{\mathcal{D}} & p_{N,2}^{\mathcal{D}} & \cdots & p_{N,V}^{\mathcal{D}} \end{bmatrix}$$

   则有：

   $$\mathbf{P}^{\mathcal{D}} = \Phi\Theta$$

   由于 $\mathbf{P}^{\mathcal{D}}$ 是观测的、已知的，所以`pLSA`对应着矩阵分解。其中要求：

   $$\sum_{v=1}^{V} p_{i,v}^{\mathcal{D}} = 1, \quad i = 1, 2, \cdots, N$$

   $$\sum_{t=1}^{T} \varphi_{i,t} = 1, \quad i = 1, 2, \cdots, N;$$

   $$\sum_{v=1}^{V} \theta_{t,v} = 1, \quad t = 1, 2, \cdots, T;$$

   $$p_{i,v}^{\mathcal{D}} \geq 0, \quad \varphi_{i,t} \geq 0, \quad \theta_{t,v} \geq 0, \quad i = 1, 2, \cdots, N; \quad t = 1, 2, \cdots, T; \quad v = 1, 2, \cdots, V$$

2.3.2 EM 算法

1. 在文档 $\mathcal{D}_i$ 中，因为采用词袋模型，所以单词的生成是独立的。假设文档 $\mathcal{D}_i$ 中包含单词 $\{\text{word}_{w_1^i},\cdots,\text{word}_{w_{n_i}^i}\}$，其中：
   - n_i 表示文档 $\mathcal{D}_i$ 的单词总数。
   - $v=w_j^i$ 表示文档 $\mathcal{D}_i$ 的第 j 个单词为 word_v 。

   则有：
   $$p(\text{word}_{w_1^i},\cdots,\text{word}_{w_{n_i}^i}\mid\mathcal{D}_i)=\prod_{j=1}^{n_i}p(\text{word}_{w_j^i}\mid\mathcal{D}_i)$$

2. 根据前面的推导，有：$p(\text{word}_v\mid\mathcal{D}_i)=\sum_{t=1}^{T}p(\text{topic}_t\mid\mathcal{D}_i)p(\text{word}_v\mid\text{topic}_t)$。则：
   $$p(\text{word}_{w_1^i},\cdots,\text{word}_{w_{n_i}^i}\mid\mathcal{D}_i)=\prod_{j=1}^{n_i}\sum_{t=1}^{T}p(\text{topic}_t\mid\mathcal{D}_i)p(\text{word}_{w_j^i}\mid\text{topic}_t)$$

   则有：
   $$p(\text{word}_{w_1^i},\cdots,\text{word}_{w_{n_i}^i},\mathcal{D}_i)=p(\mathcal{D}_i)\prod_{j=1}^{n_i}\sum_{t=1}^{T}p(\text{topic}_t\mid\mathcal{D}_i)p(\text{word}_{w_j^i}\mid\text{topic}_t)$$

3. 假设文档 $\mathcal{D}_i$ 的单词 $\{\text{word}_{w_1^i},\cdots,\text{word}_{w_{n_i}^i}\}$ 中，单词 word_v 有 $c(i,v)$ 个，$v=1,2,\cdots,V$。则有：
   $$p(\text{word}_{w_1^i},\cdots,\text{word}_{w_{n_i}^i},\mathcal{D}_i)=p(\mathcal{D}_i)\prod_{v=1}^{V}\left[\sum_{t=1}^{T}p(\text{topic}_t\mid\mathcal{D}_i)p(\text{word}_v\mid\text{topic}_t)\right]^{c(i,v)}$$

   $c(i,v)$ 的物理意义为：即文档 $\mathcal{D}_i$ 中单词 word_v 的数量。

4. 考虑观测变量 $X_i=(\text{word}_{w_1^i},\cdots,\text{word}_{w_{n_i}^i},\mathcal{D}_i)$，它表示第 i 篇文档 $\mathcal{D}_i$ 以及该文档中的 n_i 个单词。

   则有：
   $$p(X_i)=p(\mathcal{D}_i)\prod_{v=1}^{V}\left[\sum_{t=1}^{T}p(\text{topic}_t\mid\mathcal{D}_i)p(\text{word}_v\mid\text{topic}_t)\right]^{c(i,v)}$$

   由于文档之间是相互独立的，因此有：
   $$p(X_1,X_2,\cdots,X_N)=\prod_{i=1}^{N}p(X_i)$$
   $$=\prod_{i=1}^{N}\prod_{v=1}^{V}p(\mathcal{D}_i)\prod_{v=1}^{V}\left[\sum_{t=1}^{T}p(\text{topic}_t\mid\mathcal{D}_i)p(\text{word}_v\mid\text{topic}_t)\right]^{c(i,v)}$$
   $$=\prod_{i=1}^{N}\prod_{v=1}^{V}p(\mathcal{D}_i)\left[\sum_{t=1}^{T}\varphi_{i,t}\theta_{t,v}\right]^{c(i,v)}$$

要使得观测结果发生，则应该最大化 $p(X_1,X_2,\cdots,X_N)$。但是这里面包含了待求参数的乘积，其最大化难于求解，因此使用 EM 算法求解。

5. 考虑完全变量 $Y_i=(\text{word}_{w_1^i},\cdots,\text{word}_{w_{n_i}^i},\text{topic}_{z_1^i},\cdots,\text{topic}_{z_{n_i}^i},\mathcal{D}_i)$，其中 $\{\text{topic}_{z_1^i},\cdots,\text{topic}_{z_{n_i}^i}\}$ 为文档中 $\mathcal{D}_i$ 中每个单词对应的话题。

   - 由于采用词袋模型，所以生成单词是相互独立的，因此有：

   $$\begin{aligned}p(Y_i)&=p(\mathcal{D}_i)p(\text{word}_{w_1^i},\cdots,\text{word}_{w_{n_i}^i},\text{topic}_{z_1^i},\cdots,\text{topic}_{z_{n_i}^i}\mid\mathcal{D}_i)\\&=p(\mathcal{D}_i)\prod_{j=1}^{n_i}p(\text{word}_{w_j^i},\text{topic}_{z_j^i}\mid\mathcal{D}_i)\end{aligned}$$

   - 根据 $p(\mathcal{D}_i,\text{word}_v,\text{topic}_t)=p(\mathcal{D}_i)p(\text{topic}_t\mid\mathcal{D}_i)p(\text{word}_v\mid\text{topic}_t)$ 有：

   $$p(\text{word}_v,\text{topic}_t\mid\mathcal{D}_i)=p(\text{topic}_t\mid\mathcal{D}_i)p(\text{word}_v\mid\text{topic}_t)$$

   于是：

   $$p(Y_i)=p(\mathcal{D}_i)\prod_{j=1}^{n_i}p(\text{topic}_{z_j^i}\mid\mathcal{D}_i)p(\text{word}_{w_j^i}\mid\text{topic}_{z_j^i})$$

   - 由于文档之间是相互独立的，因此有：

   $$p(Y_1,Y_2,\cdots,Y_N)=\prod_{i=1}^{N}p(Y_i)=\prod_{i=1}^{N}p(\mathcal{D}_i)\prod_{j=1}^{n_i}p(\text{topic}_{z_j^i}\mid\mathcal{D}_i)p(\text{word}_{w_j^i}\mid\text{topic}_{z_j^i})$$

6. 假设在所有文档 $\{\mathcal{D}_1,\mathcal{D}_2,\cdots,\mathcal{D}_N\}$ 中，单词 word_v 不论出现在哪个文档的哪个位置，都是由同一个话题 topic_t 产生的。

   则有：

   $$\begin{aligned}&p(\mathcal{D}_i)\prod_{j=1}^{n_i}p(\text{topic}_{z_j^i}\mid\mathcal{D}_i)p(\text{word}_{w_j^i}\mid\text{topic}_{z_j^i})\\=&p(\mathcal{D}_i)\prod_{v=1}^{V}[p(\text{topic}_t\mid\mathcal{D}_i)p(\text{word}_v\mid\text{topic}_t)]^{c(i,v)}\end{aligned}$$

   则有：

   $$p(Y_1,Y_2,\cdots,Y_N)=\prod_{i=1}^{N}p(\mathcal{D}_i)\prod_{v=1}^{V}[p(\text{topic}_t\mid\mathcal{D}_i)p(\text{word}_v\mid\text{topic}_t)]^{c(i,v)}$$

   则完全数据的对数似然函数为：

   $$LL=\log p(Y_1,Y_2,\cdots,Y_N)=\sum_{i=1}^{N}p(\mathcal{D}_i)+\sum_{i=1}^{N}\sum_{v=1}^{V}[c(i,v)(\log p(\text{topic}_t\mid\mathcal{D}_i)+\log p(\text{word}_v\mid\text{topic}_t)]$$

7. E 步：求取 Q 函数，为 LL 关于后验概率 $p(\text{topic}_t\mid\mathcal{D}_i,\text{word}_v)$ 的期望。

   根据前面的推导，有：

   $$p(\text{topic}_t\mid\mathcal{D}_i,\text{word}_v)=\frac{p(\text{topic}_t\mid\mathcal{D}_i)p(\text{word}_v\mid\text{topic}_t)}{\sum_{t'=1}^{T}p(\text{topic}_{t'}\mid\mathcal{D}_i)p(\text{word}_v\mid\text{topic}_{t'})}=\frac{\tilde{\varphi}_{i,t}\tilde{\theta}_{t,v}}{\sum_{t'=1}^{T}\tilde{\varphi}_{i,t'}\tilde{\theta}_{t',v}}$$

其中 $\tilde{\varphi}_{i,t},\tilde{\theta}_{t,v}$ 均为上一轮迭代的结果，为已知量。

则有：

$$
\begin{aligned}
&Q=\mathbb{E}[LL]_{p(\text{topic}_t|\mathcal{D}_i,\text{word}_v)}\\
&=\sum_{i=1}^{N}p(\mathcal{D}_i)+\mathbb{E}[\sum_{i=1}^{N}\sum_{v=1}^{V}[c(i,v)(\log p(\text{topic}_t\mid\mathcal{D}_i)+\log p(\text{word}_v\mid\text{topic}_t)]]_{p(\text{topic}_t|\mathcal{D}_i,\text{word}_v)}\\
&=\sum_{i=1}^{N}p(\mathcal{D}_i)+\sum_{i=1}^{N}\sum_{v=1}^{V}c(i,v)\sum_{t=1}^{T}p(\text{topic}_t\mid\mathcal{D}_i,\text{word}_v))(\log p(\text{topic}_t\mid\mathcal{D}_i)+\log p(\text{word}_v\mid\text{topic}_t)\\
&=\sum_{i=1}^{N}p(\mathcal{D}_i)+\sum_{i=1}^{N}\sum_{v=1}^{V}c(i,v)\sum_{t=1}^{T}\frac{\tilde{\varphi}_{i,t}\tilde{\theta}_{t,v}}{\sum_{t'=1}^{T}\tilde{\varphi}_{i,t'}\tilde{\theta}_{t',v}}(\log\varphi_{i,t}+\log\theta_{t,v})
\end{aligned}
$$

8. `M` 步：最大化 `Q` 函数，同时考虑约束条件：

$$
\begin{gathered}
\sum_{t=1}^{T}\varphi_{i,t}=1,i=1,2,\cdots,N;\\
\sum_{v=1}^{V}\theta_{t,v}=1,t=1,2,\cdots,T;\\
\varphi_{i,t}\geq 0,\quad \theta_{t,v}\geq 0,\quad i=1,2,\cdots,N;\quad t=1,2,\cdots,T;\quad v=1,2,\cdots,V
\end{gathered}
$$

对每个参数进行求导并使之等于0，联立方程求解得到：

$$
\begin{aligned}
\varphi_{i,t}&=\frac{\sum_{v=1}^{V}c(i,v)p(\text{topic}_t\mid\mathcal{D}_i,\text{word}_v)}{n_i},\quad t=1,2\cdots,T;i=1,2,\cdots,N\\
\theta_{t,v}&=\frac{\sum_{i=1}^{N}c(i,v)p(\text{topic}_t\mid\mathcal{D}_i,\text{word}_v)}{\sum_{v'=1}^{V}\sum_{i=1}^{N}c(i,v')p(\text{topic}_t\mid\mathcal{D}_i,\text{word}_{v'})},\quad t=1,2,\cdots,T;v=1,2,\cdots,V
\end{aligned}
$$

其物理意义为：

- 文档-主题概率 $\varphi_{i,t}$：它等于文档 $\mathcal{D}_i$ 中，所有单词对应于主题 topic_t 的后验概率的加权和。权重为每个单词出现的频率。
- 主题-单词概率 $\theta_{t,v}$：它等于单词 word_v 在所有文档的主题 topic_t 上的后验概率加权和（权重为它出现的词频），占所有单词在所有文档的主题 topic_t 上的后验概率加权和（权重为每个单词出现的词频）的比例。

9. `pLSA` 的 `EM` 算法：
   - 输入：文档集合 $\mathbb{D}$，话题集合 $\mathbb{T}$，字典集合 $\mathbb{V}$
   - 输出：参数 $\Phi=\{\varphi_{i,t}\}$ 和 $\Theta=\{\theta_{t,v}\}$，其中：

$$
\begin{gathered}
\sum_{t=1}^{T}\varphi_{i,t}=1,i=1,2,\cdots,N;\\
\sum_{v=1}^{V}\theta_{t,v}=1,t=1,2,\cdots,T;\\
\varphi_{i,t}\geq 0,\quad \theta_{t,v}\geq 0,\quad i=1,2,\cdots,N;\quad t=1,2,\cdots,T;\quad v=1,2,\cdots,V
\end{gathered}
$$

   - 算法步骤：
     - 初始化：令 $m=0$，为 $\varphi_{i,t}^{<m>}$ 和 $\theta_{t,v}^{<m>}$ 赋初值，$i=1,2,\cdots,N;v=1,2,\cdots,V;t=1,2,\cdots,T$。
     - 迭代，迭代收敛条件为参数变化很小或者 `Q` 函数的变化很小。迭代步骤如下：

- `E` 步：计算 Q 函数。
  - 先计算后验概率：

$$p(\text{topic}_t \mid \mathcal{D}_i, \text{word}_v)^{<m>} = \frac{\varphi_{i,t}^{<m>}\theta_{t,v}^{<m>}}{\sum_{t'=1}^{T}\varphi_{i,t'}^{<m>}\theta_{t',v}^{<m>}}$$

  - 再计算 Q 函数：

$$Q = \sum_{i=1}^{N} p(\mathcal{D}_i) + \sum_{i=1}^{N}\sum_{v=1}^{V} c(i,v)\sum_{t=1}^{T} p(\text{topic}_t \mid \mathcal{D}_i, \text{word}_v)^{<m>}(\log\varphi_{i,t} + \log\theta_{t,v})$$

- `M` 步：计算 Q 函数的极大值，得到参数的下一轮迭代结果：

$$\varphi_{i,t}^{<m+1>} = \frac{\sum_{v=1}^{V} c(i,v)p(\text{topic}_t \mid \mathcal{D}_i, \text{word}_v)^{<m>}}{n_i},\quad t = 1,2\cdots,T; i = 1,2,\cdots,N$$

$$\theta_{t,v}^{<m+1>} = \frac{\sum_{i=1}^{N} c(i,v)p(\text{topic}_t \mid \mathcal{D}_i, \text{word}_v)^{<m>}}{\sum_{i=1}^{N}\sum_{v=1}^{V} c(i,v)p(\text{topic}_t \mid \mathcal{D}_i, \text{word}_v)^{<m>}},\quad t = 1,2,\cdots,T; v = 1,2,\cdots,V$$

- 重复上面两步直到收敛

三、LDA Model

1. 在 `pLSA` 模型中，参数 Φ, Θ 是常数。而在 `LDA` 模型中，假设 Φ, Θ 也是随机变量：
   - 参数 $\vec{\varphi}^{\,i} = (\varphi_{i,1}, \varphi_{i,2}, \cdots, \varphi_{i,T})$ 为文档 $\mathcal{D}_i$ 的主题分布（离散型的），其中 $i = 1,2,\cdots,N$。该分布也是一个随机变量，服从分布 $p(\vec{\varphi}^{\,i})$（连续型的）。
   - 参数 $\vec{\theta}^{\,t} = (\theta_{t,1}, \theta_{t,2}, \cdots, \theta_{t,V})$ 为主题 topic_t 的单词分布（离散型的），其中 $t = 1,2,\cdots,T$。该分布也是一个随机变量，服从分布 $p(\vec{\theta}^{\,t})$（连续型的）。

   事实上，`LDA` 模型是 `pLSA` 模型的贝叶斯版本。
2. 例：

   在 `pLSA` 模型中，给定一篇文档，假设：
   - 主题分布为 `{教育：0.5，经济：0.3，交通：0.2}`，它就是 $p(\text{topic}_t \mid \mathcal{D}_i)$。
   - 主题 `教育` 下的主题词分布为 `{大学:0.5,老师:0.2,课程:0.3}`，它就是 $p(\text{word}_v \mid \text{topic}_t)$。

   在 `LDA` 中：
   - 给定一篇文档，主题分布 $p(\text{topic}_t \mid \mathcal{D}_i)$ 不再固定。
     - 可能为 `{教育：0.5，经济：0.3，交通：0.2}`，也可能为 `{教育：0.3，经济：0.5，交通：0.2}`，也可能为 `{教育：0.1，经济：0.8，交通：0.1}`。
     - 但是它并不是没有规律的，而是服从一个分布 $p(\vec{\varphi})$。

       即：主题分布取某种分布的概率可能较大，取另一些分布的概率可能较小。
   - 主题 `教育` 下的主题词分布也不再固定。
     - 可能为 `{大学:0.5,老师:0.2,课程:0.3}`，也可能为 `{大学:0.8,老师:0.1,课程:0.1}`。
     - 但是它并不是没有规律，而是服从一个分布 $p(\vec{\theta})$。

       即：主题词分布取某种分布的概率可能较大，取另一些分布的概率可能较小。

3.1 文档生成

1. `LDA`模型的文档生成规则：
   - 以概率 $p(\mathcal{D}_i)$ 选中第 i 篇文档。
   - 根据参数为 $\vec{\alpha}$ 的狄利克雷分布随机采样，生成文档 $\mathcal{D}_i$ 的一个话题分布 $\vec{\varphi}_i=(\varphi_{i,1},\varphi_{i,2},\cdots,\varphi_{i,T})$
   - 根据参数为 $\vec{\eta}$ 的狄利克雷分布随机采样，对每个话题 topic_t 生成一个单词分布 $\vec{\theta}_t=(\theta_{t,1},\theta_{t,2},\cdots,\theta_{t,V})$
   - 在文档 $\mathcal{D}_i$ 中，根据话题分布 $p(\text{topic}_t\mid\mathcal{D}_i)=\varphi_{i,t}$ 来随机挑选一个话题。
   - 在话题 topic_t 中，根据单词分布 $p(\text{word}_v\mid\text{topic}_t)=\theta_{t,v}$ 来随机挑选一个单词。
   - 重复执行`挑选话题--> 挑选单词` n 次，则得到一篇包含 n 个单词 $\{\text{word}_{w_1},\text{word}_{w_2},\cdots,\text{word}_{w_n}\}$ 的文档。

     其中：$1\le w_1,\cdots,w_n\le V$；$v=w_j$ 表示文档的第 j 个单词为 word_v 。
2. 重复执行上述文档生成规则 N 次，即得到 N 篇文档组成的文档集合 $\mathbb{D}$ 。
3. 由于两次随机采样，导致 `LDA` 模型的解会呈现一定程度的随机性。
   - 所谓随机性，就是：当多次运行`LDA`算法，获得解可能会各不相同
   - 当采样的样本越稀疏，则采样的方差越大，则`LDA`的解的方差越大。
     - 文档数量越少，则文档的话题分布的采样越稀疏
     - 文档中的单词越少，则话题的单词分布的采样越稀疏

3.2 模型原理

1. 由于使用词袋模型，`LDA` 生成文档的过程可以分解为两个过程：
   - $\vec{\alpha}\rightarrow\vec{\varphi}_i\rightarrow\{\text{topic}_{z_1^i},\text{topic}_{z_2^i},\cdots,\text{topic}_{z_{n_i^i}}\}$：该过程表示，在生成第 i 篇文档 $\mathcal{D}_i$ 的时候，先从`文档-主题`分布 $\vec{\varphi}_i$ 中生成 n_i 个主题。

     其中：
     - $t=z_j^i$ 表示文档 $\mathcal{D}_i$ 的第 j 个单词由主题 topic_t 生成。
     - n_i 表示文档 $\mathcal{D}_i$ 一共有 n_i 个单词。
   - $\vec{\eta}\rightarrow\vec{\theta}_t\rightarrow\{\text{word}_{w_1^t},\text{word}_{w_2^t},\cdots,\text{word}_{w_{n_t}^t}\}$：该过程表示，在已知主题为 topic_t 的条件下，从`主题-单词`分布 $\vec{\theta}_t$ 生成 n_t 个单词。

     其中：
     - $v=w_j^t$ 表示由主题 topic_t 生成的的第 j 个单词为 word_v 。
     - n_t 为由 topic_t 生成的单词的数量。

3.2.1 主题生成过程

1. 主题生成过程用于生成第 i 篇文档 $\mathcal{D}_i$ 中每个位置的单词对应的主题。
   - $\vec{\alpha}\rightarrow\vec{\varphi}_i$ ：对应一个狄里克雷分布
   - $\vec{\varphi}_i\rightarrow\{\text{topic}_{z_1^i},\text{topic}_{z_2^i},\cdots,\text{topic}_{z_{n_i}^i}\}$：对应一个多项式分布
   - 该过程整体对应一个`狄里克雷-多项式`共轭结构：

$$p(\text{topic}_{z_1^i},\text{topic}_{z_2^i},\cdots,\text{topic}_{z_{n_i}^i};\vec{\alpha})=\int p(\text{topic}_{z_1^i},\text{topic}_{z_2^i},\cdots,\text{topic}_{z_{n_i}^i}\mid\vec{\varphi}_i)p(\vec{\varphi}_i;\vec{\alpha})d\vec{\varphi}_i$$
$$=\int\left[\prod_{j=1}^{n_i}p(\text{topic}_{z_j^i}\mid\vec{\varphi}_i)\right][Dir(\vec{\varphi}_i;\vec{\alpha})]\,d\vec{\varphi}_i$$

2. 合并文档 $\mathcal{D}_i$ 中的同一个主题。设 $n_z(i,t)$ 表示文档 $\mathcal{D}_i$ 中，主题 topic_t 出现的次数。则有：

$$\prod_{j=1}^{n_i}p(\text{topic}_{z_j^i}\mid\vec{\varphi}_i)=\prod_{t=1}^{T}\varphi_{i,t}^{n_z(i,t)}$$

则有：

$$p(\text{topic}_{z_1^i},\text{topic}_{z_2^i},\cdots,\text{topic}_{z_{n_i}^i};\vec{\alpha})=$$
$$=\int\left[\prod_{j=1}^{n_i}p(\text{topic}_{z_j^i}\mid\vec{\varphi}_i)\right][Dir(\vec{\varphi}_i;\vec{\alpha})]\,d\vec{\varphi}_i$$
$$=\int\left[\prod_{t=1}^{T}\varphi_{i,t}^{n_z(i,t)}\right]\left[\frac{1}{B(\vec{\alpha})}\prod_{t=1}^{T}\varphi_{i,t}^{\alpha_t-1}\right]d\vec{\varphi}_i$$
$$=\frac{1}{B(\vec{\alpha})}\int\prod_{t=1}^{T}\varphi_{i,t}^{n_z(i,t)+\alpha_t-1}d\vec{\varphi}_i$$
$$=\frac{B(\vec{\mathbf{n}}_z(i)+\vec{\alpha})}{B(\vec{\alpha})}$$

其中 $\vec{\mathbf{n}}_z(i)=(n_z(i,1),n_z(i,2),\cdots,n_z(i,T))$ 表示文档 $\mathcal{D}_i$ 中，各主题出现的次数。

3. 由于语料库中 N 篇文档的主题生成相互独立，则得到整个语料库的主题生成概率：

$$p(\text{topic}_{z_1^1},\text{topic}_{z_2^1},\cdots,\text{topic}_{z_{n_1}^1},\cdots,\text{topic}_{z_1^N},\cdots,\text{topic}_{z_{n_N}^N};\vec{\alpha})$$
$$=\prod_{i=1}^{N}p(\text{topic}_{z_1^i},\text{topic}_{z_2^i},\cdots,\text{topic}_{z_{n_i}^i};\vec{\alpha})=\prod_{i=1}^{N}\frac{B(\vec{\mathbf{n}}_z(i)+\vec{\alpha})}{B(\vec{\alpha})}$$

3.2.2 单词生成过程

1. 单词生成过程用于生成所有文档 $\mathbb{D}$ 的所有主题的单词。
   - $\vec{\eta}\to\vec{\theta}_t$ ： 对应一个狄里克雷分布
   - $\vec{\theta}_t\to\{\text{word}_{w_1^t},\text{word}_{w_2^t},\cdots,\text{word}_{w_{n_t}^t}\}$ ： 对应一个多项式分布
   - 数据集 $\mathbb{D}$ 中，由主题为 topic_t 生成的所有单词的分布对应一个 狄里克雷-多项式 共轭结构：

$$p(\text{word}_{w_1^t},\text{word}_{w_2^t},\cdots,\text{word}_{w_{n_t}^t}\mid\text{topic}_t;\vec{\eta})$$
$$=\int p(\text{word}_{w_1^t},\text{word}_{w_2^t},\cdots,\text{word}_{w_{n_t}^t}\mid\vec{\theta}_t,\text{topic}_t)p(\vec{\theta}_t\mid\text{topic}_t;\vec{\eta})d\vec{\theta}_t$$
$$=\int\prod_{i=1}^{n_t}p(\text{word}_{w_i^t}\mid\vec{\theta}_t,\text{topic}_t)Dir(\vec{\theta}_t;\vec{\eta})d\vec{\theta}_t$$

2. 合并主题 topic_t 生成的同一个单词。设 $n_v(t,v)$ 表示中主题 topic_t 生成的单词中，word_v 出现的次数。则有：

$$\prod_{i=1}^{n_t} p(\text{word}_{w_i^t} \mid \vec{\theta}_t, \text{topic}_t) = \prod_{v=1}^{V} \theta_{t,v}^{n_t(t,v)}$$

则有：

$$\begin{aligned}&p(\text{word}_{w_1^t}, \text{word}_{w_2^t}, \cdots, \text{word}_{w_{n_t}^t} \mid \text{topic}_t; \vec{\eta})\\ =& \int \prod_{v=1}^{V} \theta_{t,v}^{n_t(t,v)} \left[\frac{1}{B(\vec{\eta})} \prod_{v=1}^{V} \theta_{t,v}^{\eta_v - 1}\right] d\vec{\theta}_t = \frac{B(\vec{\mathbf{n}}_v(t) + \vec{\eta})}{B(\vec{\eta})}\end{aligned}$$

其中 $\vec{\mathbf{n}}_v(t) = (n_v(t,1), n_v(t,2), \cdots, n_v(t,V))$ 表示由主题 topic_t 生成的单词的词频。

3. 考虑数据集 $\mathbb{D}$ 中的所有主题，则有：

$$\begin{aligned}&p(\text{word}_{w_1^1}, \cdots, \text{word}_{w_{n_1}^1}, \cdots, \text{word}_{w_1^T}, \cdots, \text{word}_{w_{n_T}^T} \mid \text{topic}_1, \cdots, \text{topic}_T; \vec{\eta})\\ =& \prod_{t=1}^{T} p(\text{word}_{w_1^t}, \text{word}_{w_2^t}, \cdots, \text{word}_{w_{n_t}^t} \mid \text{topic}_t; \vec{\eta})\\ =& \prod_{t=1}^{T} \frac{B(\vec{\mathbf{n}}_v(t) + \vec{\eta})}{B(\vec{\eta})}\end{aligned}$$

3.2.3 联合概率

1. 根据 $p(\text{topic}_{z_1^1}, \text{topic}_{z_2^1}, \cdots, \text{topic}_{z_{n_1}^1}, \cdots, \text{topic}_{z_1^N}, \cdots, \text{topic}_{z_{n_N}^N}; \vec{\alpha})$ 以及 $p(\text{word}_{w_1^1}, \cdots, \text{word}_{w_{n_1}^1}, \cdots, \text{word}_{w_1^T}, \cdots, \text{word}_{w_{n_T}^T} \mid \text{topic}_1, \cdots, \text{topic}_T; \vec{\eta})$，可以得到数据集 $\mathbb{D}$ 的联合概率分布为：

$$\begin{aligned}&p(\text{word}_{w_1^1}, \cdots, \text{word}_{w_{n_1}^1}, \cdots, \text{word}_{w_1^T}, \cdots, \text{word}_{w_{n_T}^T}, \text{topic}_1, \cdots, \text{topic}_T; \vec{\alpha}, \vec{\eta})\\ =& p(\text{word}_{w_1^1}, \cdots, \text{word}_{w_{n_1}^1}, \cdots, \text{word}_{w_1^T}, \cdots, \text{word}_{w_{n_T}^T} \mid \text{topic}_1, \cdots, \text{topic}_T; \vec{\eta}) \times p(\text{topic}_1, \cdots, \text{topic}_T)\\ =& \prod_{t=1}^{T} \frac{B(\vec{\mathbf{n}}_v(t) + \vec{\eta})}{B(\vec{\eta})} \times \prod_{i=1}^{N} \frac{B(\vec{\mathbf{n}}_z(i) + \vec{\alpha})}{B(\vec{\alpha})}\\ =& \prod_{i=1}^{N} \prod_{t=1}^{T} \frac{B(\vec{\mathbf{n}}_z(i) + \vec{\alpha})}{B(\vec{\alpha})} \frac{B(\vec{\mathbf{n}}_v(t) + \vec{\eta})}{B(\vec{\eta})}\end{aligned}$$

其中：

- $\vec{\mathbf{n}}_z(i) = (n_z(i,1), n_z(i,2), \cdots, n_z(i,T))$ 表示文档 $\mathcal{D}_i$ 中，各主题出现的次数。
- $\vec{\mathbf{n}}_v(t) = (n_v(t,1), n_v(t,2), \cdots, n_v(t,V))$ 表示主题 topic_t 生成的单词中，各单词出现的次数。

3.2.4 后验概率

1. 若已知文档 $\mathcal{D}_i$ 中的主题 $\{\text{topic}_{z_1^i}, \text{topic}_{z_2^i}, \cdots, \text{topic}_{z_{n_i}^i}\}$，则有：

$$
\begin{aligned}
p(\vec{\varphi}_i \mid \text{topic}_{z_1^i}, \text{topic}_{z_2^i}, \cdots, \text{topic}_{z_{n_i}^i}) &= \frac{p(\vec{\varphi}_i, \text{topic}_{z_1^i}, \text{topic}_{z_2^i}, \cdots, \text{topic}_{z_{n_i}^i})}{p(\text{topic}_{z_1^i}, \text{topic}_{z_2^i}, \cdots, \text{topic}_{z_{n_i}^i})} \\
&= \frac{p(\text{topic}_{z_1^i}, \text{topic}_{z_2^i}, \cdots, \text{topic}_{z_{n_i}^i} \mid \vec{\varphi}_i) p(\vec{\varphi}_i)}{p(\text{topic}_{z_1^i}, \text{topic}_{z_2^i}, \cdots, \text{topic}_{z_{n_i}^i})} \\
&= \frac{\left[\prod_{t=1}^{T} \varphi_{i,t}^{n_z(i,t)}\right]\left[\frac{1}{B(\vec{\alpha})}\prod_{t=1}^{T} \varphi_{i,t}^{\alpha_t - 1}\right]}{\frac{B(\vec{\mathbf{n}}_z(i)+\vec{\alpha})}{B(\vec{\alpha})}} \\
&= \frac{\prod_{t=1}^{T} \varphi_{i,t}^{n_z(i,t)+\alpha_t-1}}{B(\vec{\mathbf{n}}_z(i)+\vec{\alpha})}
\end{aligned}
$$

则有：$p(\vec{\varphi}_i \mid \text{topic}_{z_1^i}, \text{topic}_{z_2^i}, \cdots, \text{topic}_{z_{n_i}^i}) = Dir(\vec{\varphi}_i; \vec{\mathbf{n}}_z(i) + \vec{\alpha})$ 。

这说明参数 $\vec{\varphi}_i$ 的后验分布也是狄里克雷分布。

2. 若已知主题 topic_t 及其生成的单词 $\{\text{word}_{w_1^t}, \cdots, \text{word}_{w_{n_t}^t}\}$ 则有：

$$
\begin{aligned}
p(\vec{\theta}_t \mid \text{topic}_t, \text{word}_{w_1^t}, \cdots, \text{word}_{w_{n_t}^t}) &= \frac{p(\vec{\theta}_t, \text{word}_{w_1^t}, \cdots, \text{word}_{w_{n_t}^t} \mid \text{topic}_t)}{p(\text{word}_{w_1^t}, \cdots, \text{word}_{w_{n_t}^t} \mid \text{topic}_t)} \\
&= \frac{p(\text{word}_{w_1^t}, \cdots, \text{word}_{w_{n_t}^t} \mid \text{topic}_t, \vec{\theta}_t) p(\vec{\theta}_t \mid \text{topic}_t)}{p(\text{word}_{w_1^t}, \cdots, \text{word}_{w_{n_t}^t} \mid \text{topic}_t)} \\
&= \frac{\left[\prod_{v=1}^{V} \theta_{t,v}^{n_v(t,v)}\right]\left[\frac{1}{B(\vec{\eta})}\prod_{v=1}^{V} \theta_{t,v}^{\eta_v - 1}\right]}{\frac{B(\vec{\eta}+\vec{\mathbf{n}}_v(t))}{B(\vec{\eta})}} \\
&= \frac{\prod_{v=1}^{V} \theta_{t,v}^{n_v(t,v)+\eta_v-1}}{B(\vec{\eta}+\vec{\mathbf{n}}_v(t))}
\end{aligned}
$$

则有：$p(\vec{\theta}_t \mid \text{topic}_t, \text{word}_{w_1^t}, \cdots, \text{word}_{w_{n_t}^t}) = Dir(\vec{\theta}_t; \vec{\mathbf{n}}_v(t) + \vec{\eta})$ 。

这说明参数 $\vec{\theta}_t$ 的后验分布也是狄里克雷分布。

3.3 模型求解

1. `LDA` 的求解有两种办法：变分推断法、吉布斯采样法。

3.3.1 吉布斯采样

1. 对于数据集 $\mathbb{D}$，其所有的单词 $\{\text{word}_{w_1^1}, \cdots, \text{word}_{w_{n_1}^1}, \cdots, \text{word}_{w_1^N}, \cdots, \text{word}_{w_{n_N}^N}\}$ 是观测的已知数据，记作 $\mathbf{WORD}$ 。

   这些单词对应的主题 $\{\text{topic}_{z_1^1}, \cdots, \text{topic}_{z_{n_1}^1}, \cdots, \text{topic}_{z_1^N}, \cdots, \text{topic}_{z_{n_N}^N}\}$ 是未观测数据，记作 $\mathbf{TOPIC}$。

   需要求解的分布是：

   $$p(\mathbf{WORD} \mid \mathbf{TOPIC})$$

   其中：

   - $v = w_j^i$ 表示文档 $\mathcal{D}_i$ 的第 j 个单词为 word_v 。
   - $t = z_j^i$ 表示文档 $\mathcal{D}_i$ 的第 j 个单词由主题 topic_t 生成。

2. 定义 $\mathbf{TOPIC}_{\neg(i,j)}$ 为：去掉 $\mathcal{D}_i$ 的第 j 个单词背后的那个生成主题（注：只是对其词频减一）：

$$\mathbf{TOPIC}_{\neg(i,j)}=\{\text{topic}_{z_1^1},\cdots,\text{topic}_{z_{n_1}^1}\cdots,\text{topic}_{z_1^i},\cdots,\text{topic}_{z_{j-1}^i},\text{topic}_{z_{j+1}^i},\cdots,\text{topic}_{z_{n_i}^i}\\,\cdots,\text{topic}_{z_1^N},\cdots,\text{topic}_{z_{n_N}^N}\}$$

定义 $\mathbf{WORD}_{\neg(i,j)}$ 为：去掉 $\mathcal{D}_i$ 的第 j 个单词：

$$\mathbf{WORD}_{\neg(i,j)}=\{\text{word}_{w_1^1},\cdots,\text{word}_{w_{n_1}^1}\cdots,\text{word}_{w_1^i}\cdots,\text{word}_{w_{j-1}^i},\text{word}_{w_{j+1}^i},\cdots,\text{word}_{w_{n_i}^i}\\,\cdots,\text{word}_{w_1^N},\cdots,\text{word}_{w_{n_N}^N}\}$$

3. 根据吉布斯采样的要求，需要得到条件分布：

$$p(\text{topic}_{z_j^i}\mid\mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD})$$

根据条件概率有：

$$p(\text{topic}_{z_j^i}\mid\mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD})=\frac{p(\text{topic}_{z_j^i},\text{word}_{w_j^i}\mid\mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD}_{\neg(i,j)})}{p(\text{word}_{w_j^i})}$$

则有：

$$p(\text{topic}_{z_j^i}\mid\mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD})\propto p(\text{topic}_{z_j^i},\text{word}_{w_j^i}\mid\mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD}_{\neg(i,j)})$$

4. 对于文档 $\mathcal{D}_i$ 的第 j 个位置，单词 $\text{word}_{w_j^i}$ 和对应的主题 $\text{topic}_{z_j^i}$ 仅仅涉及到如下的两个“狄里克雷-多项式”共轭结构

   - 文档 $\mathcal{D}_i$ 的主题分布：$\vec{\alpha}\rightarrow\vec{\varphi}_i$
   - 已知主题为 topic_t 的情况下，单词的分布：$\vec{\eta}\rightarrow\vec{\theta}_t$

   对于这两个共轭结构，去掉文档 $\mathcal{D}_i$ 的第 j 个位置的主题和单词时：

   - 先验分布（狄里克雷分布）：保持不变。
   - 文档 $\mathcal{D}_i$ 的主题分布：主题 $\text{topic}_{z_j^i}$ 频数减少一次，但是该分布仍然是多项式分布。

     其它 $N-1$ 个文档的主题分布完全不受影响。因此有：

$$p(\mathbf{TOPIC}_{\neg(i,j)};\vec{\alpha})=\prod_{i=1}^{N}\frac{B(\vec{\mathbf{n}}_z'(i)+\vec{\alpha})}{B(\vec{\alpha})}$$

   - 对于所有文档集合 $\mathbb{D}$，主题 $\text{topic}_{z_j^i}$ 的单词分布：单词 $\text{word}_{w_j^i}$ 频数减少一次，但是该分布仍然是多项式分布。

     其它 $T-1$ 个主题的单词分布完全不受影响。因此有：

$$p(\mathbf{WORD}_{\neg(i,j)}\mid\mathbf{TOPIC}_{\neg(i,j)};\vec{\eta})=\prod_{t=1}^{T}\frac{B(\vec{\mathbf{n}}_v'(t)+\vec{\eta})}{B(\vec{\eta})}$$

   - 根据主题分布和单词分布有：

$$p(\mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD}_{\neg(i,j)};\vec{\alpha},\vec{\eta})=\prod_{i=1}^{N}\prod_{t=1}^{T}\frac{B(\vec{\mathbf{n}}_z'(i)+\vec{\alpha})}{B(\vec{\alpha})}\frac{B(\vec{\mathbf{n}}_v'(t)+\vec{\eta})}{B(\vec{\eta})}$$

     其中：

     - $\vec{\mathbf{n}}_z'(i)=(n_z'(i,1),n_z'(i,2),\cdots,n_z'(i,T))$ 表示去掉文档 $\mathcal{D}_i$ 的第 j 个位置的单词和主题之后，第 i 篇文档中，各主题出现的次数。

- $\vec{\mathbf{n}}'_v(t)=(n'_v(t,1),n'_v(t,2),\cdots,n'_v(t,V))$ 表示去掉文档 $\mathcal{D}_i$ 的第 j 个位置的单词和主题之后，数据集 $\mathbb{D}$ 中，由主题 topic_t 生成的单词中，各单词出现的次数。

5. 考虑 $p(\text{topic}_{z_j^i},\text{word}_{w_j^i}\mid \mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD}_{\neg(i,j)})$。记 $t=z_j^i,v=w_j^i$，则有：

$$\begin{aligned}
&p(\text{topic}_t,\text{word}_v\mid \mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD}_{\neg(i,j)})\\
=&\int p(\text{topic}_t,\text{word}_v,\vec{\varphi}_i,\vec{\theta}_t\mid \mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD}_{\neg(i,j)})d\vec{\varphi}_i d\vec{\theta}_t\\
=&\int p(\text{topic}_t,\vec{\varphi}_i\mid \mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD}_{\neg(i,j)})\\
&\times p(\text{word}_v,\vec{\theta}_t\mid \mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD}_{\neg(i,j)})d\vec{\varphi}_i d\vec{\theta}_t\\
=&\int p(\text{topic}_t\mid\vec{\varphi}_i)\times p(\vec{\varphi}_i\mid \mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD}_{\neg(i,j)})\\
&\times p(\text{word}_v\mid\vec{\theta}_t)\times p(\vec{\theta}_t\mid \mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD}_{\neg(i,j)})d\vec{\varphi}_i d\vec{\theta}_t\\
=&\int \varphi_{i,t}Dir(\vec{\varphi}_i;\vec{\mathbf{n}}'_z(i)+\vec{\alpha})\times\theta_{t,v}Dir(\vec{\theta}_t;\vec{\mathbf{n}}'_v(t)+\vec{\eta})d\vec{\varphi}_i d\vec{\theta}_t\\
=&\int \varphi_{i,t}Dir(\vec{\varphi}_i;\vec{\mathbf{n}}'_z(i)+\vec{\alpha})d\vec{\varphi}_i\times\int\theta_{t,v}Dir(\vec{\theta}_t;\vec{\mathbf{n}}'_v(t)+\vec{\eta})d\vec{\theta}_t\\
=&\mathbb{E}[\varphi_{i,t}]_{Dir}\times\mathbb{E}[\theta_{t,v}]_{Dir}
\end{aligned}$$

根据狄里克雷分布的性质有：

$$\begin{aligned}
\mathbb{E}[\varphi_{i,t}]_{Dir}&=\frac{n'_z(i,t)+\alpha_t}{\sum_{t'=1}^{T}\left[n'_z(i,t')+\alpha_{t'}\right]}\\
\mathbb{E}[\theta_{t,v}]_{Dir}&=\frac{n'_v(t,v)+\eta_v}{\sum_{v'=1}^{V}\left[n'_v(t,v')+\eta_{v'}\right]}
\end{aligned}$$

则有：

$$\begin{aligned}
&p(\text{topic}_t,\text{word}_v\mid \mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD}_{\neg(i,j)})\\
=&\frac{n'_z(i,t)+\alpha_t}{\sum_{t'=1}^{T}\left[n'_z(i,t')+\alpha_{t'}\right]}\times\frac{n'_v(t,v)+\eta_v}{\sum_{v'=1}^{V}\left[n'_v(t,v')+\eta_{v'}\right]}
\end{aligned}$$

其中：$t=z_j^i$ 为文档 $\mathcal{D}_i$ 的第 j 个位置的单词背后的主题在主题表的编号；$v=w_j^i$ 为文档 $\mathcal{D}_i$ 的第 j 个位置的单词在词汇表中的编号。

6. 根据上面的推导，得到吉布斯采样的公式：

$$p(\text{topic}_{z_j^i}\mid \mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD})\propto\frac{n'_z(i,z_j^i)+\alpha_{z_j^i}}{\sum_{t'=1}^{T}\left[n'_z(i,t')+\alpha_{t'}\right]}\times\frac{n'_v(z_j^i,w_j^i)+\eta_{w_j^i}}{\sum_{v'=1}^{V}\left[n'_v(z_j^i,v')+\eta_{v'}\right]}$$

- 第一项刻画了：文档 $\mathcal{D}_i$ 中，第 j 个位置的单词背后的主题占该文档所有主题的比例（经过 $\vec{\alpha}$ 先验频数调整）。
- 第二项刻画了：在数据集 $\mathbb{D}$ 中，主题 $\text{topic}_{z_j^i}$ 中，单词 $\text{word}_{w_j^i}$ 出现的比例（经过 $\vec{\eta}$ 先验频数调整）。
- 它整体刻画了：文档 $\mathcal{D}_i$ 中第 j 个位置的单词为 $\text{word}_{w_j^i}$，且由主题 $\text{topic}_{z_j^i}$ 生成的可能性。

3.3.2 模型训练

1. 定义文档-主题计数矩阵 $\mathbf{T}$ 为：

$$\mathbf{T}=\begin{bmatrix} n_z(1,1) & n_z(1,2) & \cdots & n_z(1,T) \\ n_z(2,1) & n_z(2,2) & \cdots & n_z(2,T) \\ \vdots & \vdots & \ddots & \vdots \\ n_z(N,1) & n_z(N,2) & \cdots & n_z(N,T) \end{bmatrix}$$

其中第 i 行代表文档 $\mathcal{D}_i$ 的主题计数。

定义主题-单词计数矩阵 $\mathbf{W}$ 为：

$$\mathbf{W}=\begin{bmatrix} n_v(1,1) & n_v(1,2) & \cdots & n_v(1,V) \\ n_v(2,1) & n_v(2,2) & \cdots & n_v(2,V) \\ \vdots & \vdots & \ddots & \vdots \\ n_v(T,1) & n_v(T,2) & \cdots & n_v(T,V) \end{bmatrix}$$

其中第 t 行代表 主题 topic_t 的单词计数

2. `LDA` 训练的吉布斯采样算法
   - 输入：
     - 单词词典 $\mathbb{V}$
     - 超参数 $\vec{\alpha},\vec{\eta}$
     - 主题数量 T
     - 语料库 $\mathbb{D}$
   - 输出：
     - 文档-主题分布 $\vec{\varphi}_i$ 的估计量
     - 主题-单词分布 $\vec{\theta}_t$ 的估计量

       > 因为这两个参数都是随机变量，因此使用它们的期望来作为一个合适的估计
   - 算法步骤：
     - 设置全局变量：
       - $n_{i,t}^z$ 表示文档 $\mathcal{D}_i$ 中，主题 topic_t 的计数。它就是$n_z(i,t)$，也就是 $\mathbf{T}$ 的第 i 行第 t 列。
       - $n_{t,v}^v$ 表示主题 topic_t 中，单词 word_v 的计数。它就是$n_v(t,v)$，也就是 $\mathbf{W}$ 的第 t 行第 v 列。
       - n_i^z 表示各文档 $\mathcal{D}_i$ 中，主题的总计数。它也等于该文档的单词总数，也就是文档长度，也就是 $\mathbf{T}$的第 i 行求和。
       - n_t^v 表示单主题 topic_t 中，单词的总计数。它也就是 $\mathbf{W}$的第 t 行求和。
     - 随机初始化：
       - 对全局变量初始化为 0 。
       - 遍历文档：$i\in\{1,2,\cdots,N\}$
         - 对文档 $\mathcal{D}_i$ 中的每一个位置 $j=1,2,\cdots,n_i$，其中 n_i 为文档 $\mathcal{D}_i$ 的长度：
           - 随机初始化每个位置的单词对应的主题：$\text{topic}_{z_j^i}\rightarrow z_j^i=t\sim Mult(\frac{1}{T})$
           - 增加“文档-主题”计数：$n_{i,t}^z+=1$
           - 增加“文档-主题”总数：$n_i^z+=1$
           - 增加“主题-单词”计数：$n_{t,v}^v+=1$，其中 $v=w_j^i$
           - 增加“主题-单词”总数：$n_t^v+=1$
     - 迭代下面的步骤，直到马尔科夫链收敛：

- 遍历文档：$i \in \{1, 2, \cdots, N\}$
  - 对文档 i 中的每一个位置 $j = 1, 2, \cdots, n_i$，其中 n_i 为文档 $\mathcal{D}_i$ 的长度：
    - 删除该位置的主题计数，设 $t = z_j^i$：

$$
\begin{aligned}
n_{i,t}^{z} -= 1 \\
n_{i}^{z} -= 1 \\
n_{t,v}^{v} -= 1 \\
n_{t}^{v} -= 1
\end{aligned}
$$

    - 根据下面的公式，重新采样得到该单词的新主题 $\text{topic}_{z_j^i}$：

$$
\begin{aligned}
p(\text{topic}_{z_j^i} \mid \mathbf{TOPIC}_{\neg(i,j)}, \mathbf{WORD}) \propto & \frac{n_z'(i, z_j^i) + \alpha_{z_j^i}}{\sum_{t'=1}^{T} \left[n_z'(i, t') + \alpha_{t'} \right]} \\
& \times \frac{n_v'(z_j^i, w_j^i) + \eta_{w_j^i}}{\sum_{v'=1}^{V} \left[n_v'(z_j^i, v') + \eta_{v'} \right]}
\end{aligned}
$$

    - 记新的主题在主题表中的编号为 $\tilde{t}$，则增加该单词的新的主题计数：

$$
\begin{aligned}
n_{i,\tilde{t}}^{z} += 1 \\
n_{i}^{z} += 1 \\
n_{\tilde{t},v}^{v} += 1 \\
n_{\tilde{t}}^{v} += 1
\end{aligned}
$$

- 如果马尔科夫链收敛，则根据下列公式生成`文档-主题分布` Φ 的估计，以及`主题-单词`分布 Θ 的估计：

$$
\begin{aligned}
\hat{\varphi}_{i,t} &= \mathbb{E}[\varphi_{i,t}]_{Dir} = \frac{n_z'(i, t) + \alpha_t}{\sum_{t'=1}^{T} \left[n_z'(i, t') + \alpha_{t'} \right]} \\
\hat{\theta}_{t,v} &= \mathbb{E}[\theta_{t,v}]_{Dir} = \frac{n_v'(t, v) + \eta_v}{\sum_{v'=1}^{V} \left[n_v'(t, v') + \eta_{v'} \right]}
\end{aligned}
$$

3.3.3 模型推断

1. 一旦得到 `LDA` 模型，则对于新的文档 $\mathcal{D}_{new}$，其推断过程与训练过程完全类似。

   推断过程中，需要让吉布斯采样公式中的主题-单词计数矩阵 $\mathbf{W}$ 稳定不变。所以采样过程只需要更新该文档的`文档-主题分布` $\vec{\varphi}_{new}$。

2. `LDA` 推断算法
   - 输入：
     - 单词词典 $\mathbb{V}$
     - 超参数 $\vec{\alpha}$
     - 已有 `LDA` 模型的主题-单词计数矩阵 $\mathbf{W}$
     - 新的文档 $\mathcal{D}_{new}$
   - 输出：文档-主题分布 $\vec{\varphi}_{new}$ 的估计量
   - 算法步骤：

- 设置全局变量：
  - $n_{i,t}^{z}$ 表示文档 $\mathcal{D}_i$ 中，主题 topic_t 的计数。它就是$n_z(i,t)$，也就是 $\mathbf{T}$ 的第 i 行第 t 列。
  - n_i^z 表示文档 $\mathcal{D}_i$ 中各主题的总数。它也等于该文档的单词总数，也就是文档长度，也就是 $\mathbf{T}$ 的第 i 行求和。
- 随机初始化：
  - 对全局变量初始化为 0
  - 遍历文档：$i \in \{1,2,\cdots,N\}$
    - 对文档 $\mathcal{D}_i$ 中的每一个位置 $j=1,2,\cdots,n_i$，其中 n_i 为文档 $\mathcal{D}_i$ 的长度：
      - 随机初始化位置 j 处单词对应的主题：$z_j^i=t\sim Mult(\frac{1}{K})$
      - 增加`文档-主题`计数：$n_{i,t}^{z}+=1$
      - 增加`文档-主题`总数：$n_i^z+=1$
- 迭代下面的步骤，直到马尔科夫链收敛：
  - 遍历文档：$i \in \{1,2,\cdots,N\}$
    - 对文档 $\mathcal{D}_i$ 中的每一个位置 $j=1,2,\cdots,n_i$，其中 n_i 为文档 $\mathcal{D}_i$ 的长度：
      - 删除位置 j 处单词对应的主题 topic_t 的计数：

$$n_{i,t}^{z}-=1$$
$$n_i^z-=1$$

      - 根据下面的公式，重新采样得到位置 j 处单词的新主题 $\text{topic}_{z_j^i}$：

$$p(\text{topic}_{z_j^i}\mid \mathbf{TOPIC}_{\neg(i,j)},\mathbf{WORD})\propto \frac{n_z'(i,z_j^i)+\alpha_{z_j^i}}{\sum_{t'=1}^{T}\left[n_z'(i,t')+\alpha_{t'}\right]} \times \frac{n_v'(z_j^i,w_j^i)+\eta_{w_j^i}}{\sum_{v'=1}^{V}\left[n_v'(z_j^i,v')+\eta_{v'}\right]}$$

      - 记新的主题在主题表中的编号为 $\tilde{t}$，则增加该单词的新的主题计数：

$$n_{i,\tilde{t}}^{z}+=1$$
$$n_{\tilde{t}}^{z}+=1$$

  - 如果马尔科夫链收敛，则根据下列公式生成`文档-主题分布` Φ 的估计

$$\hat{\varphi}_{i,t}=\mathbb{E}[\varphi_{i,t}]_{Dir}=\frac{n_z'(i,t)+\alpha_t}{\sum_{t'=1}^{T}\left[n_z'(i,t')+\alpha_{t'}\right]}$$

四、模型讨论

1. `LSA` 的主要缺点：
   - 缺乏可解释性。主题的成分可能是随机的。
   - 需要大量的文档和单词才能获取较好的结果。

4.1 过拟合

1. `pLSA` 容易陷入过拟合。

   在`pLSA` 中，认为：

   - 文档-主题分布 $\vec{\varphi}_i, i=1,2,\cdots,N$ 不是随机变量，而是我们不知道的常量
   - 主题-单词分布 $\vec{\theta}_t, t=1,2,\cdots,T$ 也不是随机变量，也是我们不知道的常量

   `pLSA` 通过拟合训练数据集来求解这些参数，这意味着这些参数只能表征当前的训练集的文档的特征。

   对于测试集的文档，`pLSA` 认为它也符合训练集的文档特征。事实上这就是一种过拟合，尤其是当训练集的文档数量太少时，非常容易陷入过拟合。

2. 以人口抽样问题为类比。`pLSA` 的思想认为：人口的男女比例是一个常数。

   给出一个人口集合，`pLSA` 先统计男女比例（假如训练集是从医院获取的）。假如结果为 `2 :1`， 则 `pLSA` 会断言：所有的人口比例都是 `2 : 1`。

   于是训练集越小，`pLSA` 越容易陷入过拟合，离真实结果也就越远。

3. `LDA` 会给 $\vec{\varphi}_i, \vec{\theta}_t$ 加入一些先验性的知识。

   - 当数据量较小，先验性的知识会占据主导地位
   - 当数据量较大，真实数据占据主导地位

4. 以人口抽样为类比。`LDA` 会首先假设男女比例为 `1000:1000`。 然后再统计人口集合中男女的人数，最终得到的结果。

   假设人口集合中男女的人数分别为 `200:100`，则最终`LDA` 得到男女比例为 `1200:1100` 。

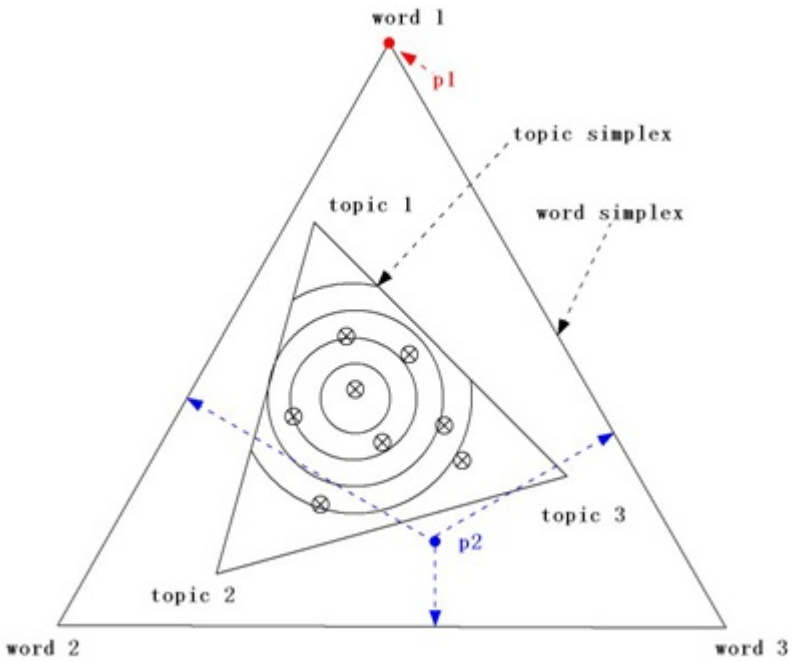
   虽然该结果离真实的结果可能有偏差，但是它比 `pLSA` 的结果要更好。

5. 当数据量足够大的时候， `pLSA` 跟 `LDA` 的结果相差无几。

   这是因为当数据量足够大时，真实数据的信息会淹没掉先验知识。

4.2 模型比较

1. 下面一张图形比较了 `Mixture of Unigram`， `pLSA` 与 `LDA` 的区别。



   假设有词汇表共有 3 个单词，主题表共有 3个主题。

   - 最外的三角形为单词三角形。

内部每个点（如 p_2）表示一个概率分布，表示产生 `word1,word2,word3` 这三个词的概率的大小。

- 靠内的三角形为主题三角形。

  内部每个点表示一个概率分布，表示产生 `topic1,topic2,topic3` 这三个主题的概率的大小。

2. `Mixture of Unigram` 模型：该模型中，只有一个主题，所以随机在三个主题中选择一个。

   假设选择 `topic1`，于是根据 `topic1` 到外边三角形的各边的距离来随机生成单词。

   于是主题三角形的三个顶点的任意一点即代表了 `Mixture of Unigram` 模型。

3. `pLSA` 模型：主题三角形内任意一些点（如带叉的点所示）就是一个 `pLSA` 模型。

   产生文档的过程：

   - 先根据主题三角形内带叉的点到主题三角形的三条边的距离来选择一个主题。
   - 然后根据该主题到单词三角形的各边的距离来选择一个单词。
   - 重复执行`选择主题-选择单词`的过程，即可得到一篇文档。

4. `LDA` 模型： 就是一个 `LDA` 模型。

   - 主题三角形内，每一条曲线表示了`topic`分布的分布，即 `topic` 分布取某些值的概率较大，取另外一些值的概率较小。它刻画了 `LDA` 模型选择主题的过程。
   - 单词三角形内，每一条曲线表示了`word`分布的分布。它刻画了 `LDA` 模型选择单词的过程。