Topic Modelling

-- Lu Meilin

2019.9.27

一元模型 Unigram model

定义变量：

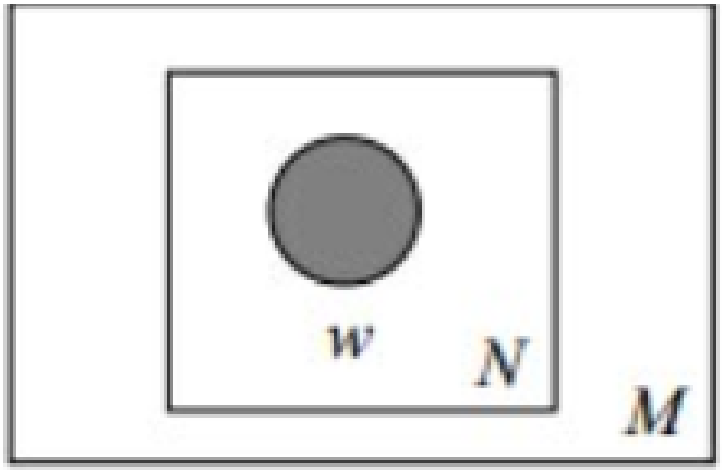
- w表示词，V表示所有单词的个数（固定值）
- z表示主题，k是主题的个数（预先给定，固定值）
- $D = (\mathbf{w}_1, \cdots, \mathbf{w}_M)$表示语料库，其中的$M$是语料库中的文档数（固定值）
- $\mathbf{w} = (w_1, w_2, \cdots, w_N)$表示文档，其中的$N$表示一个文档中的词数（随机变量）

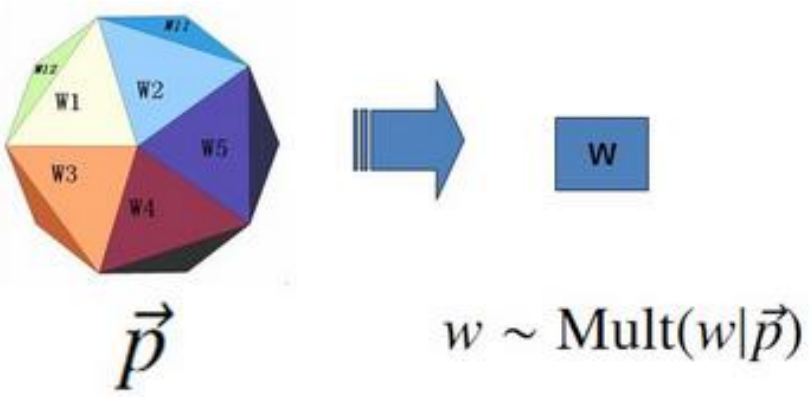
一元模型 Unigram model

给定文档，同时给定主题，不考虑词的顺序

对于文档$\mathbf{w} = (w_1, w_2, \cdots, w_N)$，用$p(w_n)$表示词$w_n$的先验概率，生成文档$\mathbf{w}$的概率为：

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n)$$

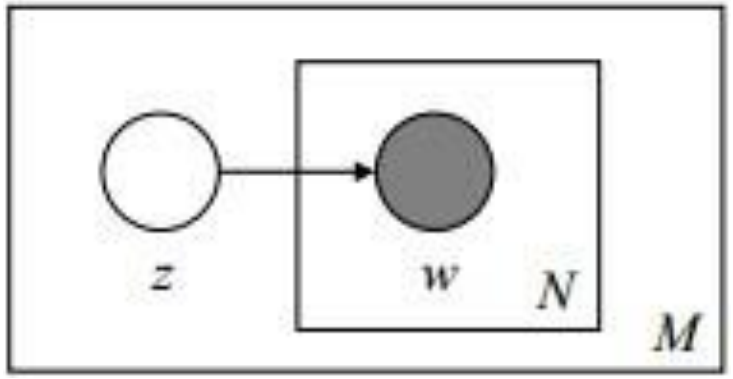




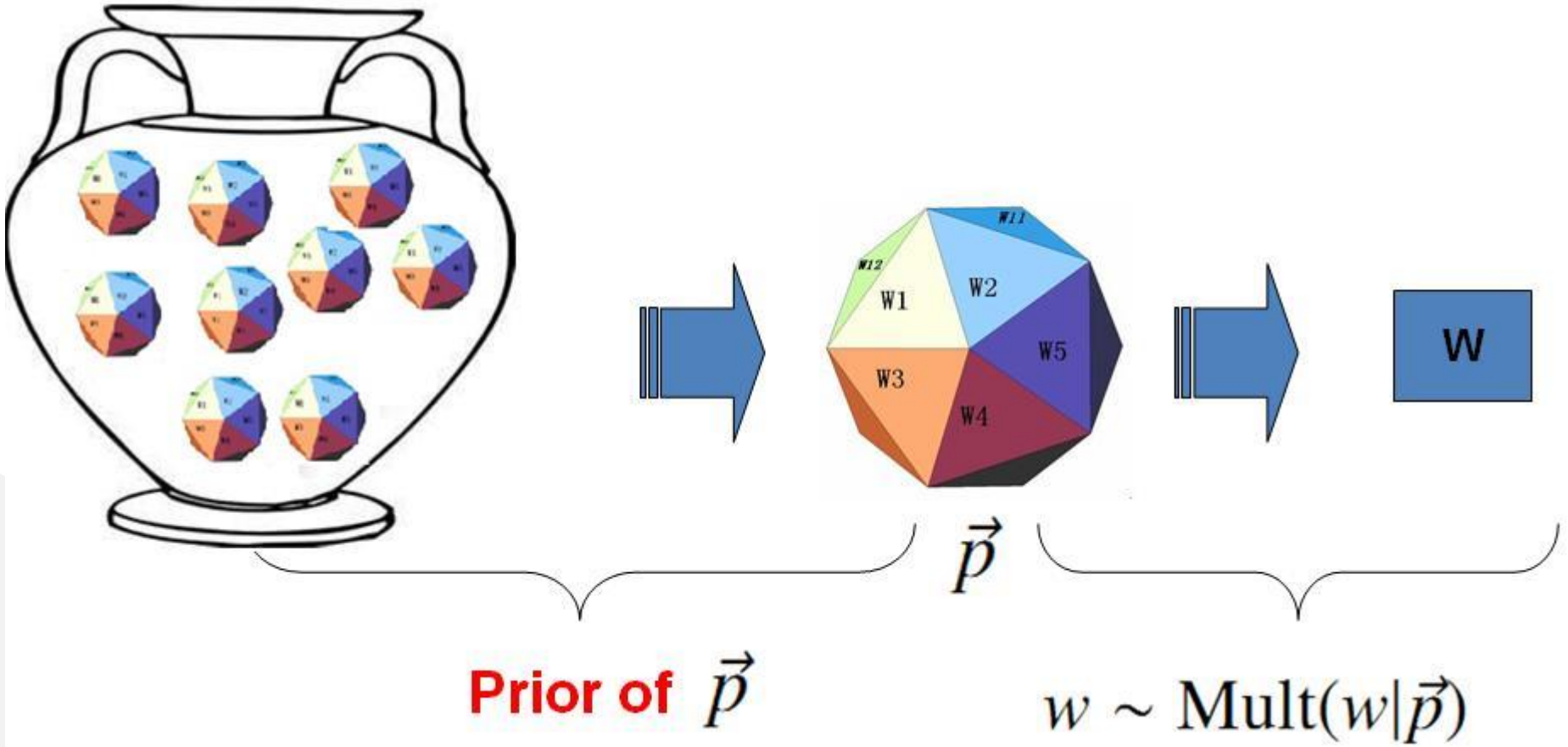
混合一元模型Mixture of unigrams models

主题未给定，一篇文档只有一个主题

$$p(\mathbf{w}) = p(z_1)\prod_{n=1}^{N} p(w_n|z_1) + \cdots + p(z_k)\prod_{n=1}^{N} p(w_n|z_k) = \sum_{z} p(z)\prod_{n=1}^{N} p(w_n|z)$$



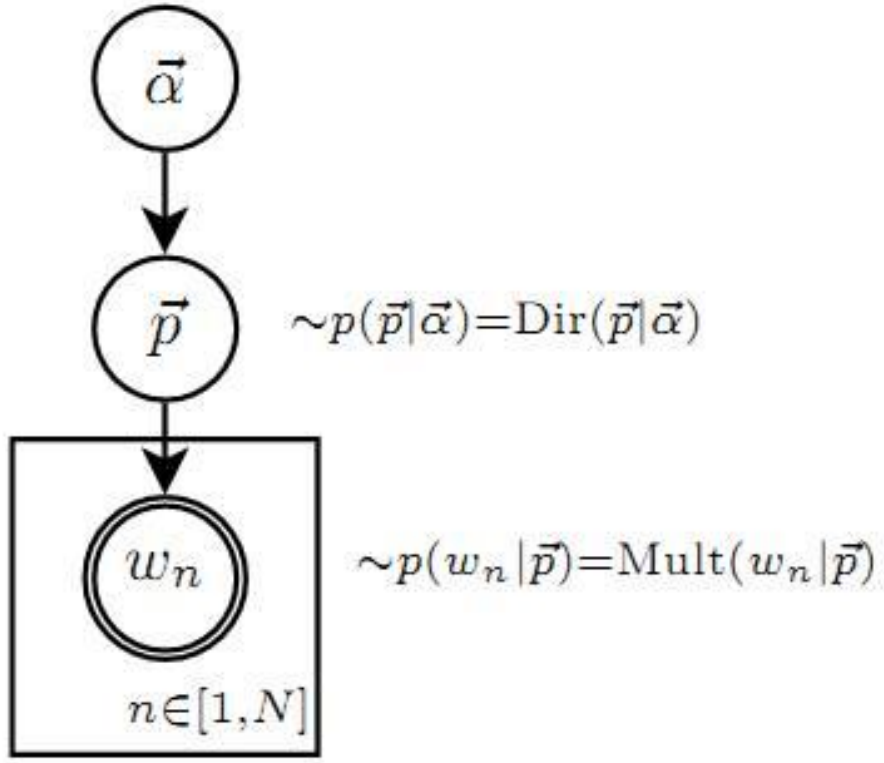
贝叶斯观点下的Unigram Model



Dirichlet 先验下的Unigram Model

p 和α是隐含未知变量

- p是词服从的Multinomial分布的参数
- α是Dirichlet 分布(Multinomial分布的先验分布) 的参数
- 一般α由经验事先给定，p由观察到的文本中出现的词学习得到，表示文本中出现每个词的概率



Dirichlet先验+多项分布的数据 ⟶ 后验分布为Dirichlet分布

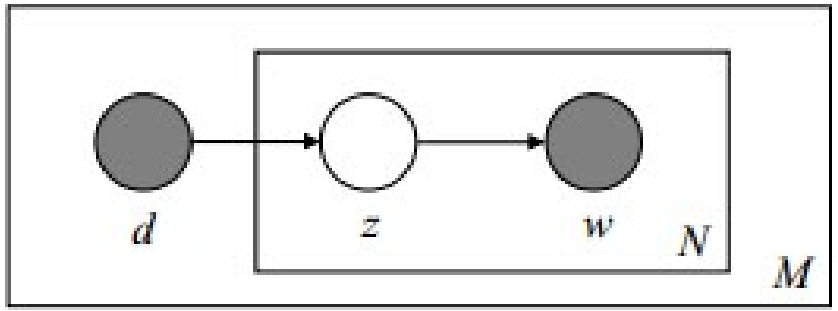
pLSA模型

定义变量：

- $P(d_i)$表示海量文档中某篇文档被选中的概率。
- $P(w_j|d_i)$表示词w_j在给定文档d_i中出现的概率。
  - 怎么计算得到呢？针对海量文档，对所有文档进行分词后，得到一个词汇列表，这样每篇文档就是一个词语的集合。对于每个词语，用它在文档中出现的次数除以文档中词语总的数目便是它在文档中出现的概率$P(w_j|d_i)$。
- $P(z_k|d_i)$表示具体某个主题z_k在给定文档d_i下出现的概率。
- $P(w_j|z_k)$表示具体某个词w_j在给定主题z_k下出现的概率，与主题关系越密切的词，其条件概率$P(w_j|z_k)$越大。

pLSA模型

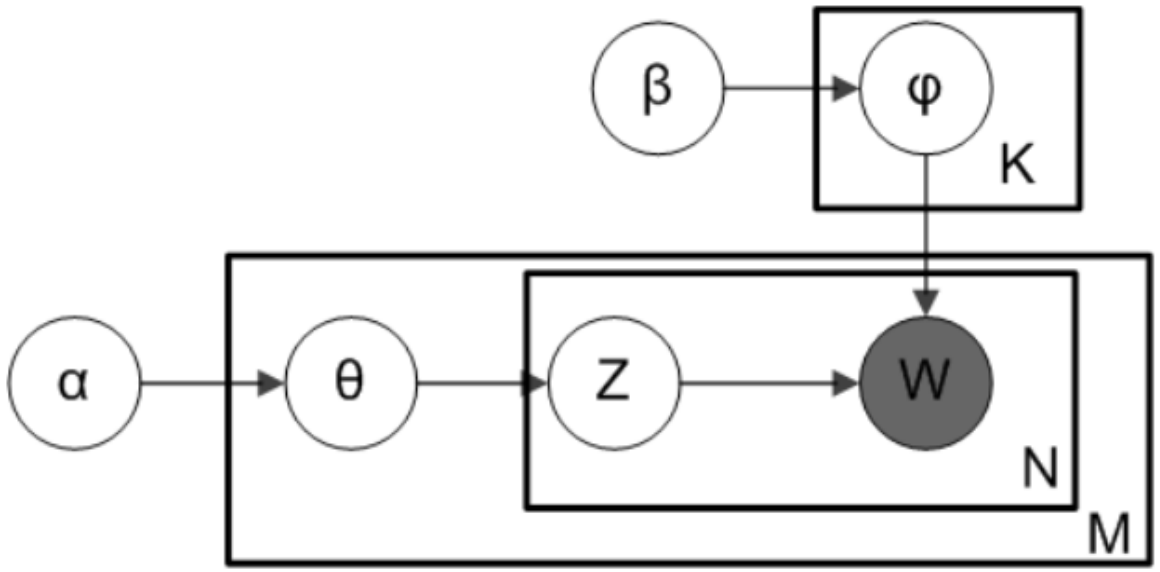
“文档-词项”生成模型



1. 按照概率$P(d_i)$选择一篇文档d_i
2. 选定文档d_i后，从主题分布中按照概率$P(z_k|d_i)$选择一个隐含的主题类别z_k
3. 选定z_k后，从词分布中按照概率$P(w_j|z_k)$选择一个词w_j

所以pLSA中生成文档的整个过程便是选定文档生成主题，确定主题生成词。

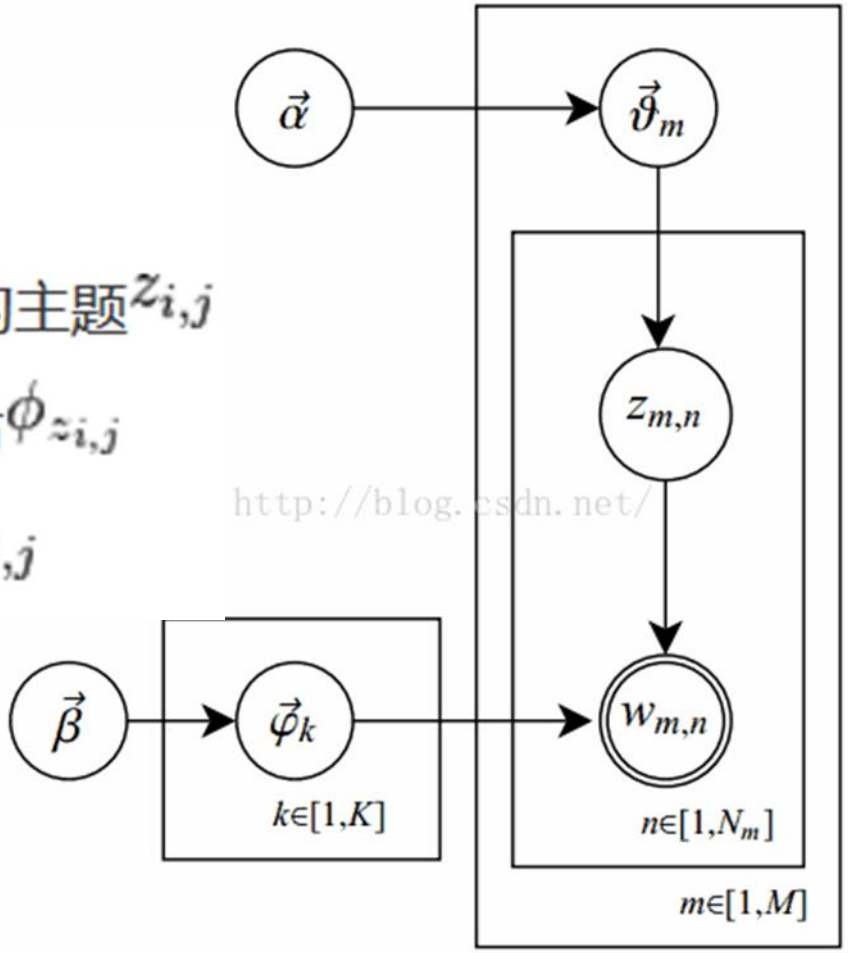
LDA （*Latent Dirichlet allocation*）



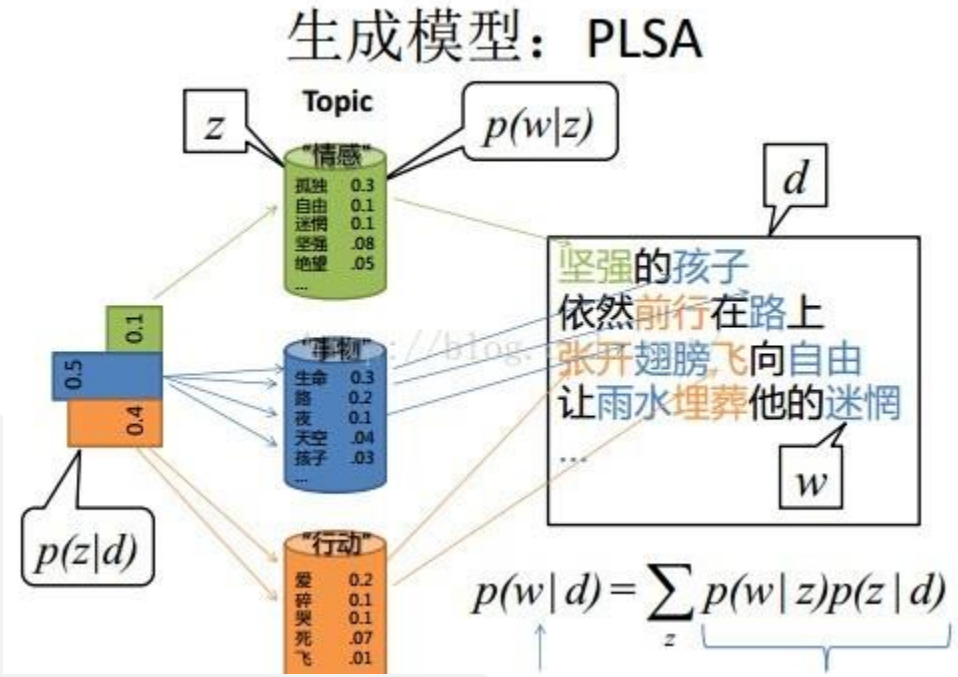
- K为主题数，N为词数，M为文档数
- α是文档主题分布的先验分布Dirichlet分布中的参数
- α是主题分布的先验分布Dirichlet分布中的参数
- β是词分布的先验分布Dirichlet分布中的参数
- $\varphi(k)$是第k个主题的词分布
- $\theta(i)$ 是第i个文档的主题分布
- $w(i,j)$ 表示第i个文档中的第j个词
- z(i,j) 表示 $w(i,j)$ 的主题

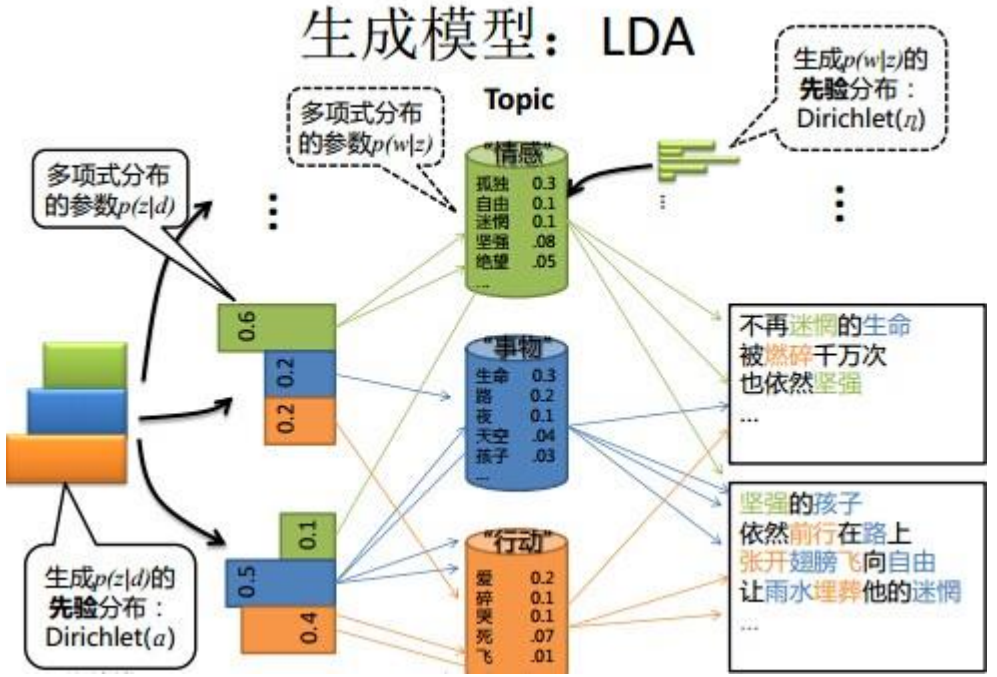
LDA （*Latent Dirichlet allocation*）

- 从狄利克雷分布α中取样生成文档i的主题分布θ_i
- 从主题的多项式分布θ_i中取样生成文档i第j个词的主题$z_{i,j}$
- 从狄利克雷分布β中取样生成主题$z_{i,j}$的词语分布$\phi_{z_{i,j}}$
- 从词语的多项式分布$\phi_{z_{i,j}}$中采样最终生成词语$w_{i,j}$



LDA （*Latent Dirichlet allocation*）





LDA （*Latent Dirichlet allocation*）

参数估计

- EM算法--pLSA
- Gibbs抽样--LDA

THANKS!