第5章：文本聚类

课件制作：王新宇、吴昊一、屠可伟
讲解人：王新宇

文本聚类

- 应用
  - 新闻聚合网站

文本聚类

- 应用
  - 客户评价分类

文本聚类：定义

- 输入：
  - 一系列文档$\{d_1, d_2, \ldots, d_n\}$
- 输出：
  - 分配后的簇
    - $C_1 = \{d_1, d_3, \ldots\}$
    - $C_2 = \{d_2, d_6, \ldots\}$
    - $C_3 = \{d_4, \ldots\}$
    - ……

常用方法

- 用特征向量表示文本（第4章）
- 应用任意聚类算法
  - k均值
  - 层次凝聚聚类
  - 基于高斯混合的最大期望值法
  - ……

需要测量向量之间的距离，例如 L2

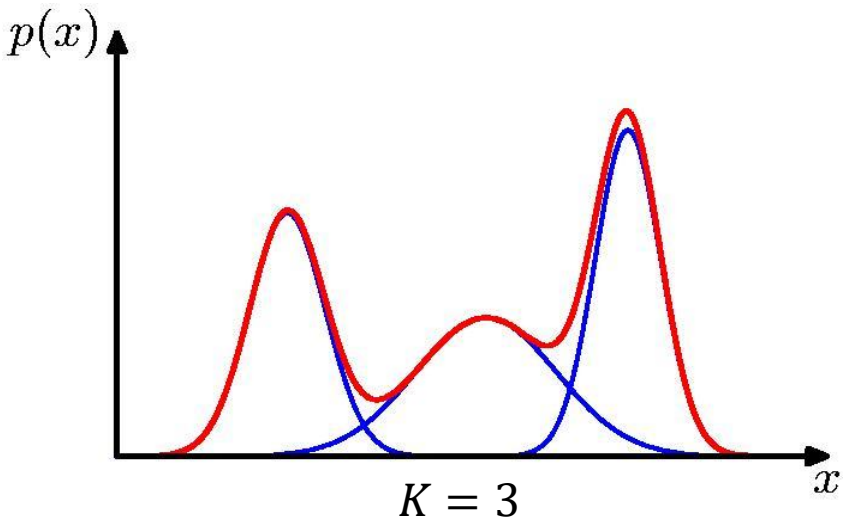
高斯混合（ MoG ）

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

高斯函数

混合系数

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

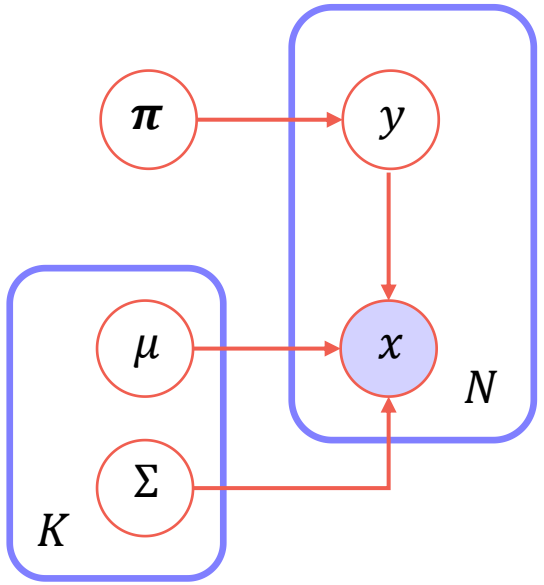


高斯混合（ MoG ）

- $P(Y)$ ： k个簇上的分布
- $P(X|Y)$：每个簇从具有均值μ_i和协方差矩阵的Σ_i的混合高斯中生成数据点的分布

每个数据点都是从***生成过程*** 中采样的：

1. 以概率π_i选择$y = i$
2. 从高斯函数$\mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i)$生成数据点

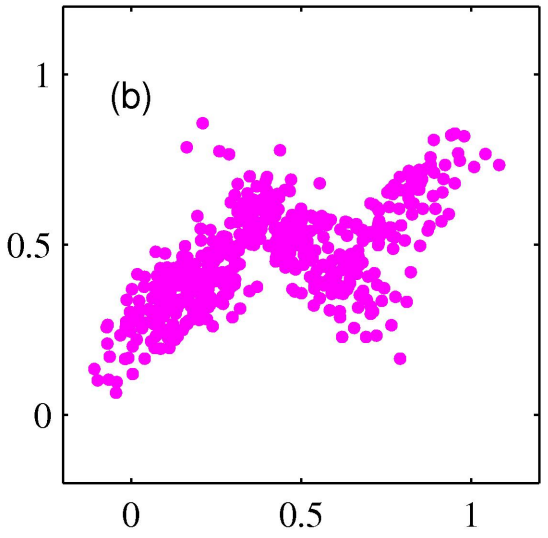


高斯混合（ MoG ）

- 在聚类中，我们不知道标签 Y的分布
- 最大化边际似然：

$$\prod_j P(\boldsymbol{x}_j) = \prod_j \sum_i P(y_j = i, \boldsymbol{x}_j) = \prod_j \sum_i \pi_i N(\boldsymbol{x}_j | \mu_i, \Sigma_i)$$

- 我们如何优化它？



期望最大值法（EM）

- 选择 K 个随机聚类模型（高斯）
- 交替执行：
  - [E步骤]将数据点按比例分配给不同的聚类模型
  - [M步骤]根据（按比例）分配的点修改每个聚类模型
- 当边际概率没有显着变化时停止

- EM = Expectation-Maximization通过坐标上升最大化边际似然

E步骤

- [E步骤]将数据点按比例分配给不同的模型
  - 计算每个数据点的标签分布

$$\mathrm{P}\left(y_j = i \middle| \boldsymbol{x}_j, \theta^{(t)}\right) \propto \pi_i{}^{(t)} N\left(\boldsymbol{x}_j \middle| \mu_i{}^{(t)}, \Sigma_i{}^{(t)}\right)$$

只需评价高斯分布 $\boldsymbol{x}_j$

M步骤

- [E步骤]将数据实例按比例分配给不同的模型
  - 计算每个数据点的标签分布

$$\mathrm{P}\left(y_j = i \middle| \boldsymbol{x}_j, \theta^{(t)}\right) \propto \pi_i{}^{(t)} N\left(\boldsymbol{x}_j \middle| \mu_i{}^{(t)}, \Sigma_i{}^{(t)}\right)$$

- [M 步骤]根据（按比例）分配的点修改每个聚类模型
  - 计算给定标签分布的参数的加权最大似然估计

$$\mu_i{}^{(t+1)} = \frac{\sum_j \mathrm{P}\left(y_j = i \middle| \boldsymbol{x}_j, \theta^{(t)}\right) \boldsymbol{x}_j}{\sum_{j'} \mathrm{P}\left(y_{j'} = i \middle| \boldsymbol{x}_{j'}, \theta^{(t)}\right)}$$

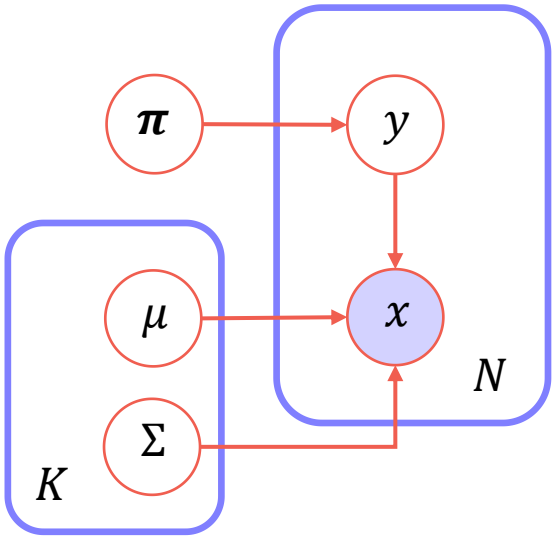
$$\pi_i{}^{(t+1)} = \frac{\sum_j \mathrm{P}\left(y_j = i \middle| \boldsymbol{x}_j, \theta^{(t)}\right)}{m}$$

m= # 训练样本

$$\Sigma_i{}^{(t+1)} = \frac{\sum_j \mathrm{P}\left(y_j = i \middle| \boldsymbol{x}_j, \theta^{(t)}\right) \left[\boldsymbol{x}_j - \mu_i{}^{(t+1)}\right]\left[\boldsymbol{x}_j - \mu_i{}^{(t+1)}\right]^T}{\sum_{j'} \mathrm{P}\left(y_{j'} = i \middle| \boldsymbol{x}_{j'}, \theta^{(t)}\right)}$$

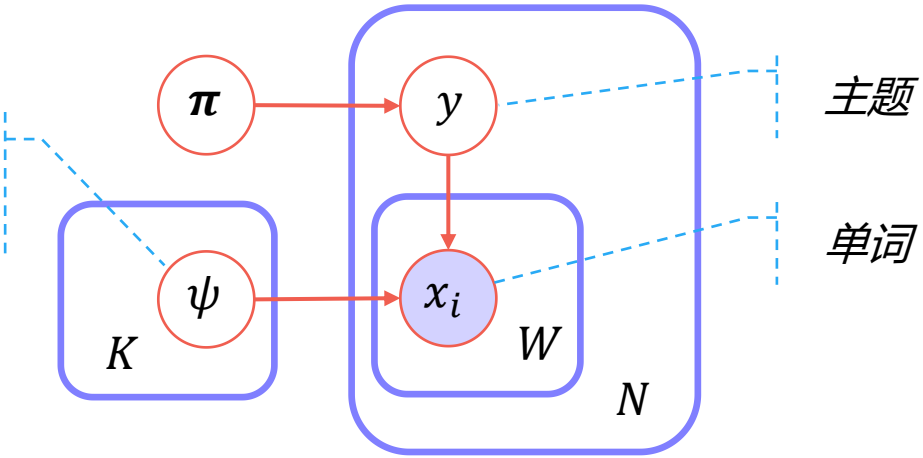
生成模型

- MoG生成具有高斯分布
- 我们可以直接生成文档（单词序列）吗？
  - 可以使用离散分布



生成模型

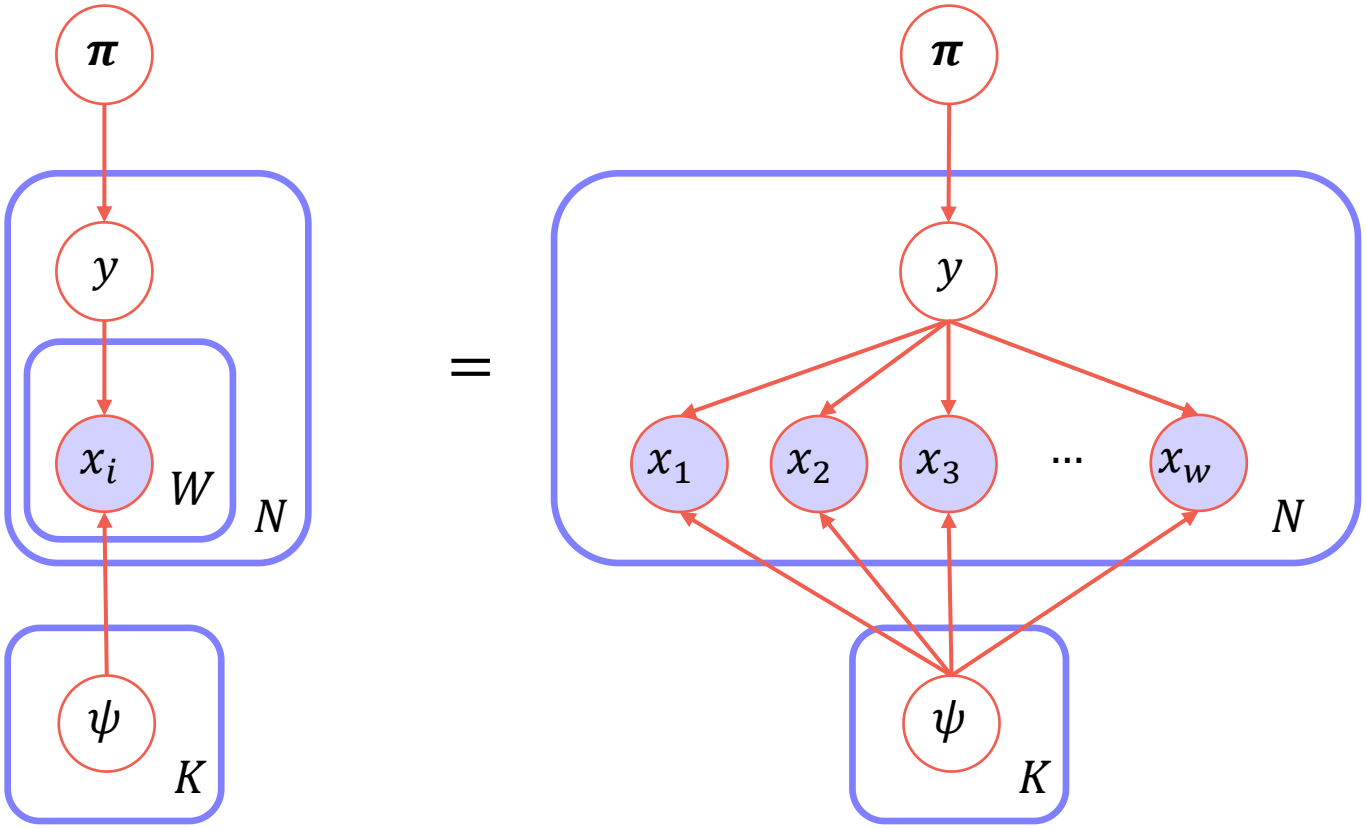
- MoG是用K个高斯分布中的一个来生成一个文档的特征向量
- 我们可以直接生成文档（单词序列）吗？
  - 可以使用离散分布



每个主题的单词分布

例如，单词-[如何、怎么、得到……]
体育：[0.1, 0.3, 0.2, …]
艺术：[0.2，0.3，0.1，…]
科学：[0.3, 0.4, 0.1, …]
……

生成模型



这正是朴素贝叶斯模型！

无监督朴素贝叶斯

- 我们可以运行 EM 来进行朴素贝叶斯的无监督学习
  - 即基于单词而不是特征的文本聚类
- [E 步骤] 根据不同的主题按比例分配文档
  - 计算每个文档的主题分布

$$\mathrm{P}\left(y_j = i \middle| x_{j,1:w}, \theta^{(t)}\right) \propto \pi_i{}^{(t)} \prod_{k=1}^{w} \mathrm{P}\left(x_{j,k} \middle| \psi_i{}^{(t)}\right)$$

无监督朴素贝叶斯

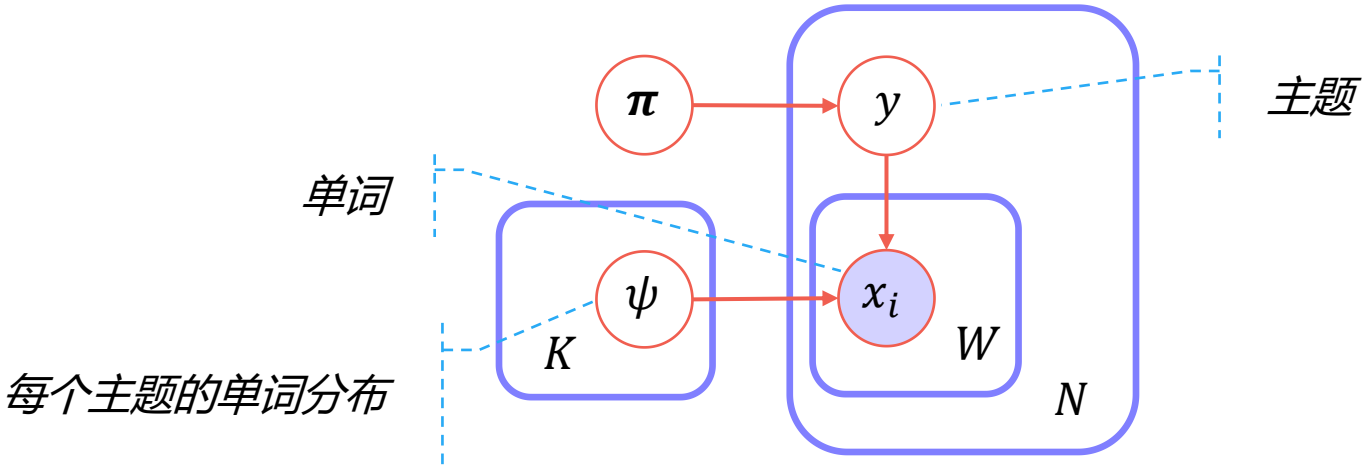
- 我们可以运行 EM 来进行朴素贝叶斯的无监督学习
  - 即基于单词而不是特征的文本聚类
- [M 步骤]根据（按比例）指定的单词修改每个主题
  - 计算给定主题分布的参数的加权最大似然估计
  - 表示$\psi_i = \{p_{i,1}, p_{i,2}, \ldots, p_{i,v}\}$

$$p_{i,l}^{(t+1)} = \frac{\sum_j \mathrm{P}(y_j = i | x_{j,1:w}, \theta^{(t)}) \sum_k \mathbf{1}(x_{j,k} = l)}{\sum_j \mathrm{P}(y_j = i | x_{j,1:w}, \theta^{(t)}) \cdot w_j} \qquad \pi_i^{(t+1)} = \frac{\sum_j \mathrm{P}(y_j = i | x_{j,1:w}, \theta^{(t)})}{m}$$

其中v是词汇量，m是训练文档的数量。

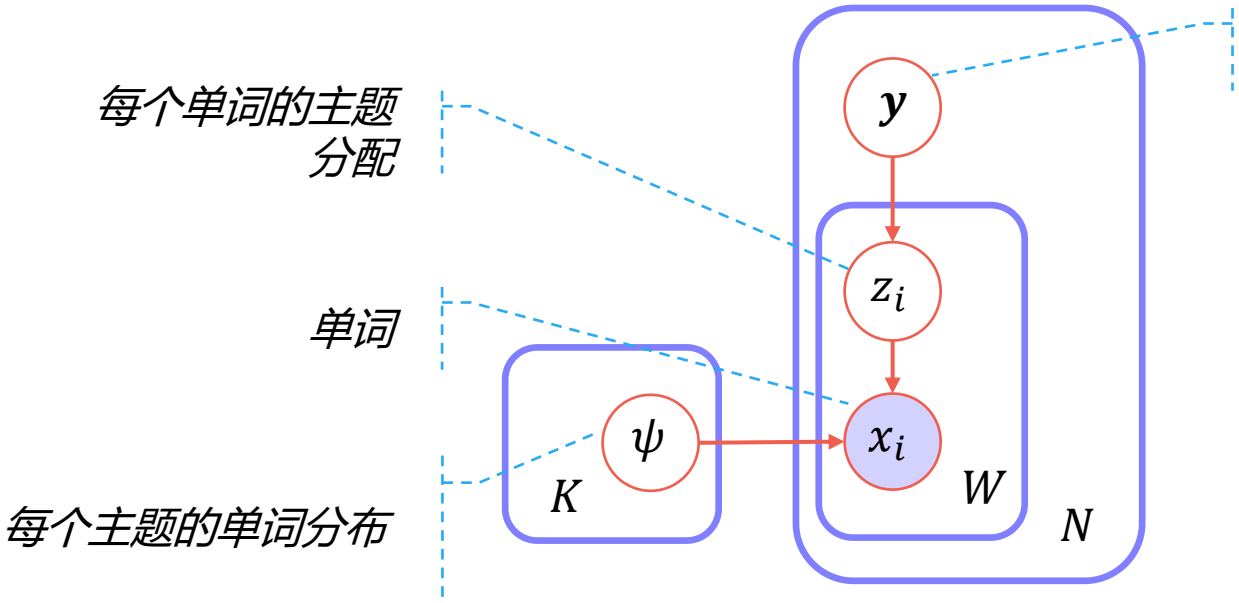
主题建模

- 文本簇可能对应不同的主题
- 到目前为止，我们假设每个文档都有一个簇标签
- 但是，一篇文档可能涵盖多个主题
  - 如何进行学习？



主题建模

- 文本簇可能对应不同的主题
- 到目前为止，我们假设每个文档都有一个簇标签
- 但是，一篇文档可能涵盖多个主题
  - 如何进行学习？

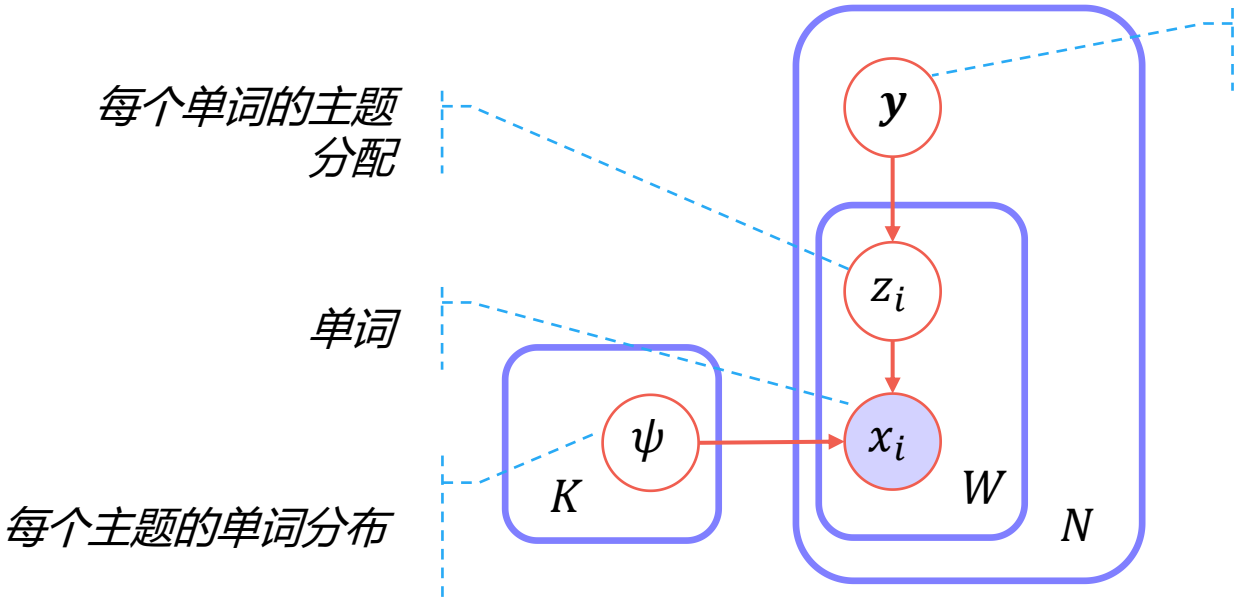


例如， **[0.1, 0.3, 0.2, …]**

运动	**0.1**
艺术	**0.3**
科学	**0.2**
……	… …

主题建模

- 这称为概率潜在语义分析（ pLSA ）
  - 同样，我们可以运行 EM 来学习它
- 我们可以进一步在主题和单词分布上添加狄利克雷先验

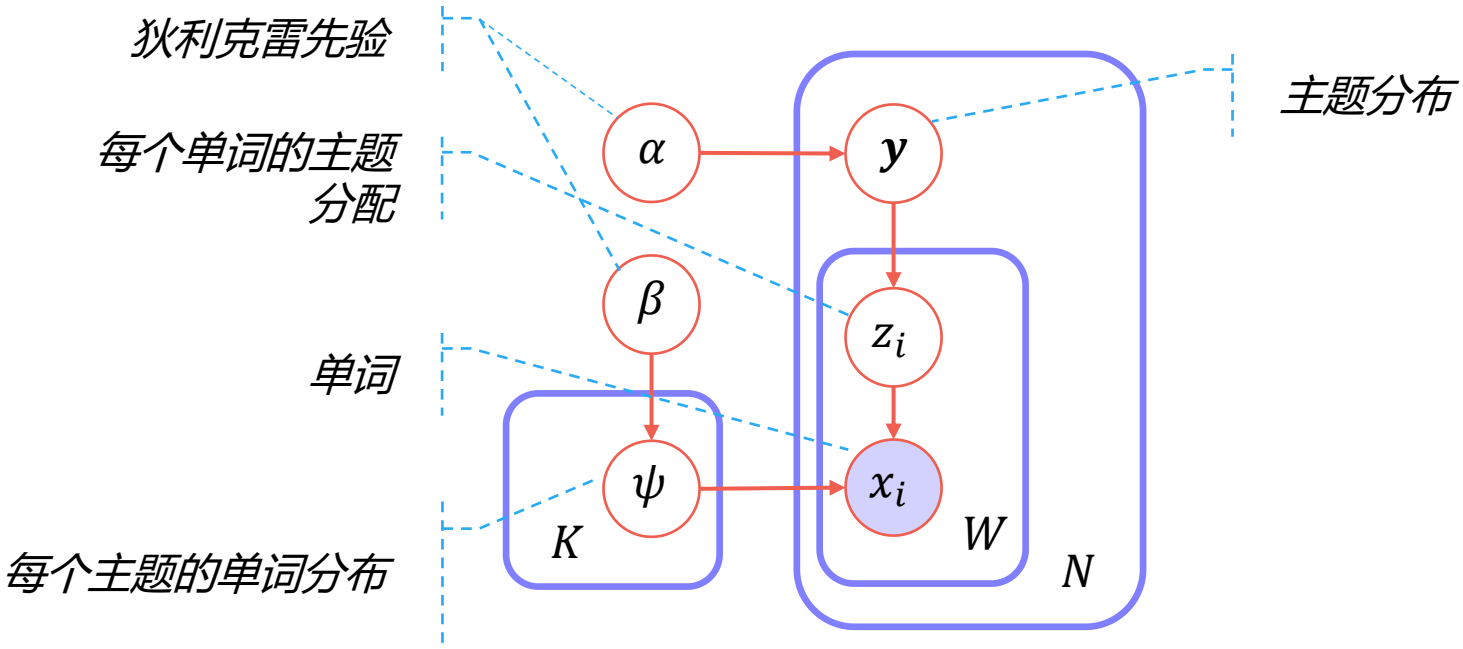


例如， **[0.1, 0.3, 0.2, …]**

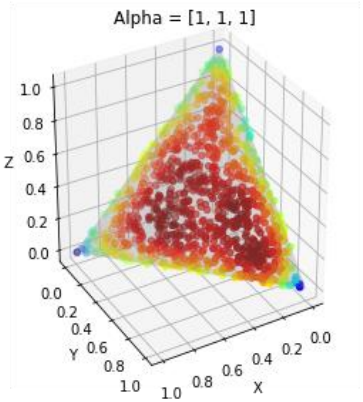
运动	**0.1**
艺术	**0.3**
科学	**0.2**
……	… …

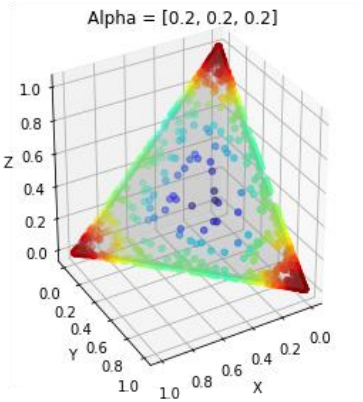
主题建模

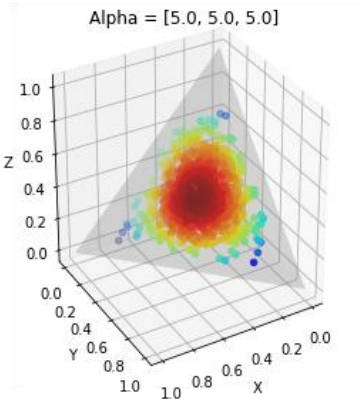
- 狄利克雷先验：鼓励主题和单词分布稀疏
  - 文档应仅涵盖几个主题
  - 在一个主题中，只有一部分单词具有高频率

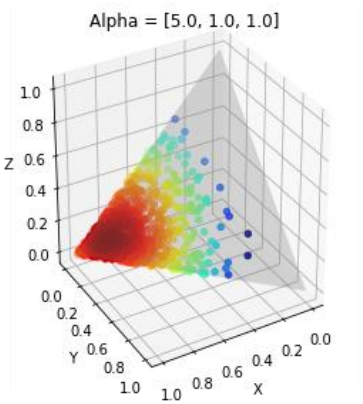


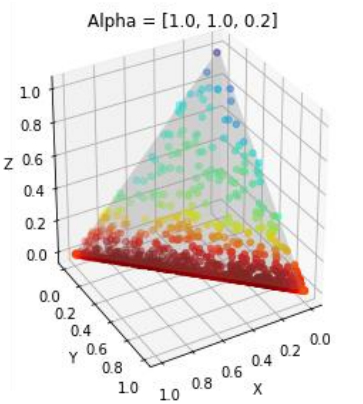
狄利克雷分布

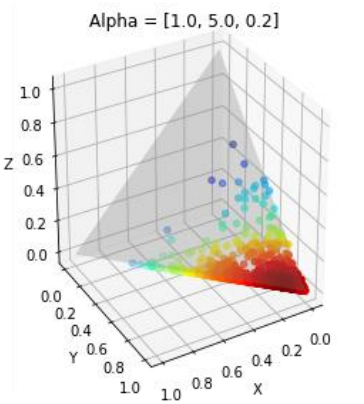










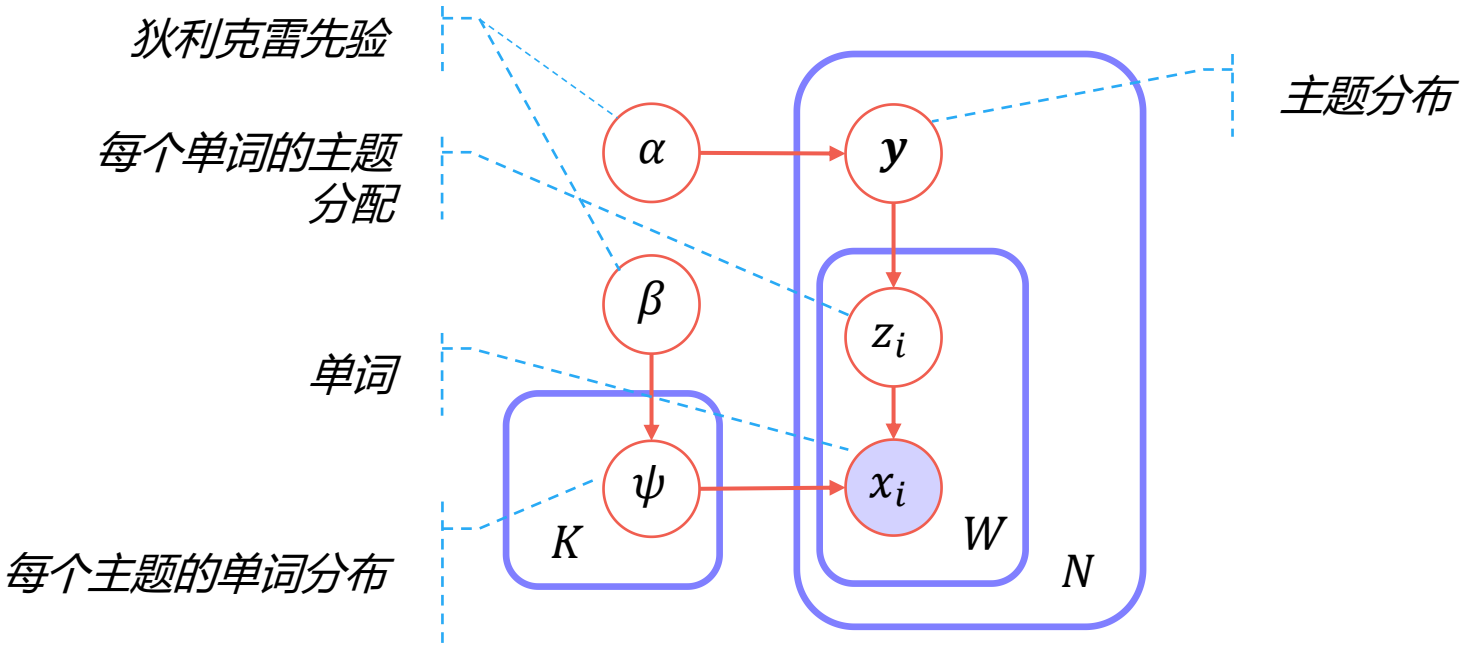


Boyu. AI

[by Thushan Ganegedara at USydney]

主题建模

- 这称为潜在狄利克雷分配（LDA）
  - 学习：变分推理或MCMC



通用EM算法

- 可用于学习任何具有隐变量（缺失数据）的模型
- 交替计算：
  - 根据当前参数值计算隐变量的分布
  - 根据隐变量的分布计算新参数值以最大化预期对数似然
- 当没有变化时停止

- 可以达到局部最优，但不一定达到全局最优

文本聚类

- 高斯混合
- 无监督朴素贝叶斯
- 主题模型
- 学习
  - 期望最大值算法