**2主题模型**

2.1 隐狄利克雷分配模型Latent Dirichlet Allocation model

原始的主题模型用于挖掘出文档集合中隐含着的主题，因此方便大量文本档案的搜索、摘要与管理。Blei et.al. 提出的隐狄利克雷分配（Latent Dirichlet Allocation，LDA）模型是最经典流行的，也是最简单的主题模型，它发现隐含的主题并且基于层次式贝叶斯模型把文档集合按隐含主题进行分组。

LDA模型给离散型数据集提供了概率产生式模型。在文本数据上下文中，数据集是由D篇文档构成的集合。每篇由单词集合构成的文档被建模成在一个隐含主题集合上的有限混合模型。比如，生物信息相关的文章有很大可能是从“基因”或“数据挖掘”主题中抽取出单词组成的，但是不大可能会有来自“食物”主题的单词。每个主题被建模为单词上的多项式分布。例如，“基因主题”有较大的概率是关于基因工程的单词词汇，而“数据挖掘”主题的单词大概率是跟数据挖掘相关。

2.1.1 产生式过程

LDA模型作为一种产生式聚类模型，同时也定义了一种数据如何产生并结构化的产生式过程。在LDA模型中，语料库D中的每篇文档的产生式过程如下所示：

1. 对每篇文档，采样主题的概率.
2. 对每个单词，
3. 采样主题的分配.
4. 采样单词.

第1步是说明假设每篇文档是由不同概率的主题构成的。第2步b表明每个独立的单词是从第2步a中确定的主题中采样得到的。并且这个主题是由单词上的多项式分布构成。