在看Classificational MedLDA之前，先看一下unsupervised LDA模型。LDA(Latent Dirichlet allocation) 模型是一个生成模型，它刻画了一个语料(有很多的普通文档组成)每个文档各个位置的单词是怎么生成的，这有点类似于解数学题里的假设变量，求解变量的过程。

在BleiNJ03中是这样描绘该生成过程的：



解释一下这个生成过程：

步骤1~3只是描述了一篇文档是怎么生成的，对于多篇文档，按照步骤1~3依次生成每篇文档。

过程中涉及到的变量和参数解释：

：主题个数

：词典中单词的个数

：泊松分布的参数，是个标量

：生成文档的长度，标量

：Dirichlet 分布的参数，维向量

：生成文档的主题分布(多项分布)，维向量

：表示生成文档第个位置的单词，维向量，如果是单词，并且为整数，则的第个位置元素为1，其它元素均为0。后面会涉及到一个相关的变量，该变量是个标量，取值0或1，若是单词，则为1，否则为0。

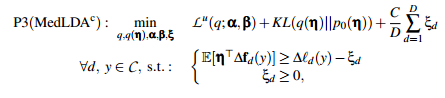
：表示生成文档第个位置的单词所属于的主题，维向量，如果是主题，并且为整数，则的第个位置元素为1，其它元素均为0。

该文档按照参数生成文档的长度N



Classificational MedLDA是 MedLDA的分类模型，它将Maximum Likelihood 和Max Margin方法统一到一个目标函数中。

目标函数：



表示 ，在unsupervised LDA 里， ,分母 是不可解的，所以用另一个分布(参数和 完全不同) 。