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，在后面表示的第个元素

：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，可以把它理解为的第个元素。

：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。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，若是主题，则为1，否则为0，可以把它理解为的第个元素。

：矩阵，行是主题，列是单词在词典中的索引。元素，即从主题生成单词的概率。应该是按行正规化，即每一行的元素加和为1。后面用表示的第行。

步骤1 ，该文档按照参数生成文档的长度。在BleiNJ03中，这样解释



在后面的模型中，会忽略变量。生成文档的长度取值实际文档的长度。

步骤2，从参数为的Dirichlet 分布生成文档的主题分布，



步骤3，对于个位置，依次进行步骤(a)，(b)。

步骤(a)：从参数为的多项分布生成主题，



步骤(b)：从参数为的多项分布生成单词，



对于一篇文档，按照步骤1~3生成，，。是的矩阵，即所生成文档的所有主题的表示，为其第个元素。是的矩阵，即所生文档所有单词的表示，为其第个元素。

按照步骤1~3，把各个子步骤的概率相乘，我们得到该篇文档的生成概率。

 (1)

在公式(1)中，参数是和，即我们假设其是已知的，最终要求解它们。而，我们假设它们就是未知的，最终也不需要求解出具体的数值，我们称这种变量为隐含变量(latent variables)，隐含变量能够有效的刻画已知变量(这里是)之间的关系。

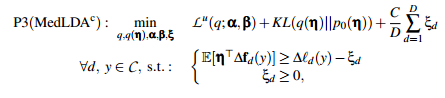
因为，从始至终我们都把它们当作未知量，所以要把它们积分出来，这样就得到一篇实际文档(假设单词都已知，就像本篇文档一样，虽然还没写完，为)的模型(LDA)概率：

一个语料(假设为,共篇文档)，每一篇文档都按照步骤1~3来生成，我们得到该实际语料的模型(LDA)概率：



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) 。