关于LDA有两种含义，一种是线性判别分析（Linear Discriminant Analysis），一种是概率主题模型：隐含狄利克雷分布（Latent Dirichlet Allocation，简称LDA），本文讲后者。

它可以将每篇文档的主题以概率的形式表示，也是一种典型的词袋模型，对于每一个词都有不同的概率对应不同的主题，最后反推出文章所有的主题。

在讲LDA模型之前，再循序渐进理解基础模型：Unigram model、mixture of unigrams model，以及跟LDA最为接近的pLSA模型。为了方便描述，首先定义一些变量：

w

表示词，

V

示所有单词的个数（固定值）。

z

表示主题，

k

主题的个数（预先给定，固定值）。

$D = (W_1, ..., W_M)$

表示语料库，其中的M是语料库中的文档数（固定值）。

$W = (w_1, w_2, ..., w_N)$

表示文档，其中的N表示一个文档中的词数（随机变量）。

PLSA模型

Unigram model:

对于文档

$W = (w_1, w_2, ..., w_N)$

用

$p(w_n)$

表示词

w_n

的先验概率，生成文档w的概率为：

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

Mixture of unigrams model

该模型的生成过程是：给某个文档先选择一个主题z,再根据该主题生成文档，该文档中的所有词都来自一个主题。假设主题有

$z_1, ..., z_n$

，生成文档w的概率为：

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

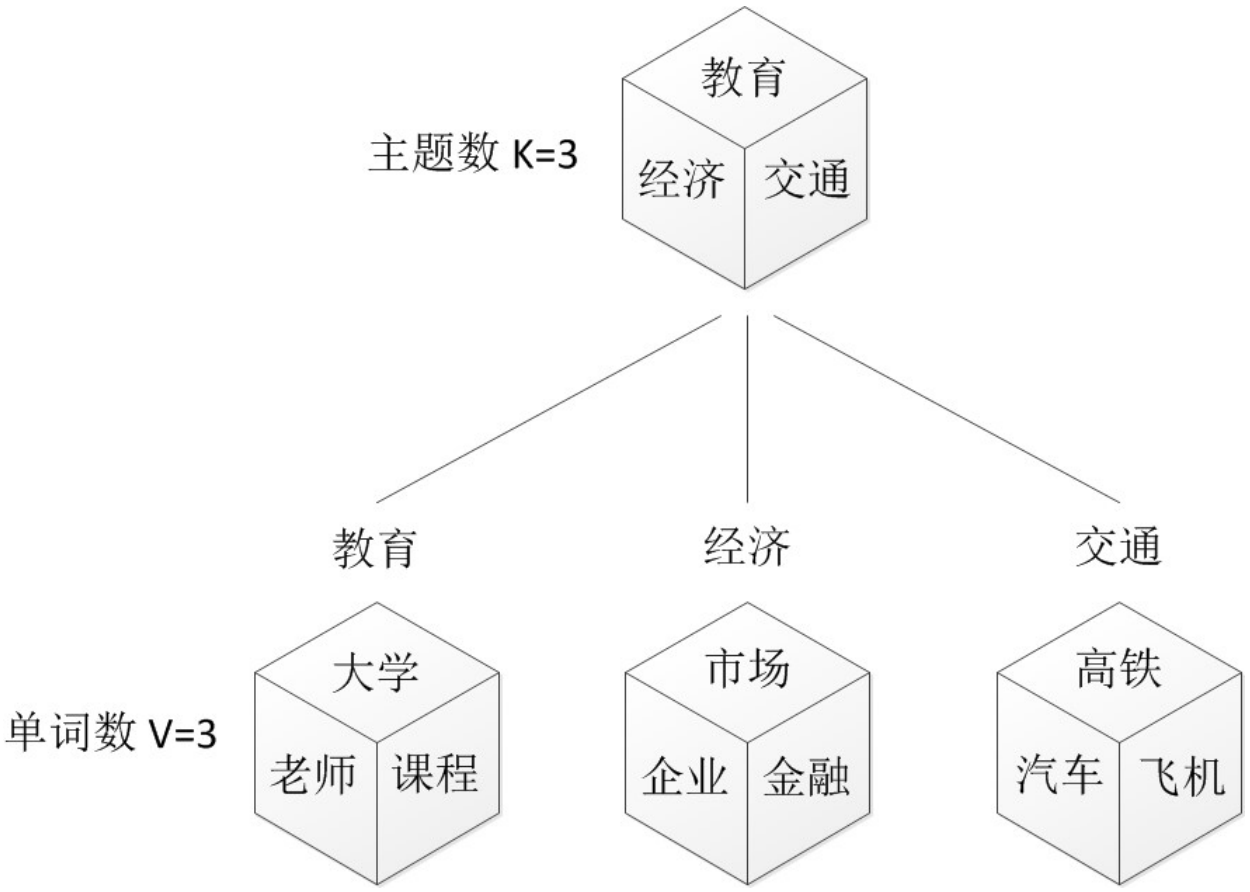
理解了pLSA模型后，到LDA模型也就一步之遥——给pLSA加上贝叶斯框架，便是LDA。

在上面的Mixture of unigrams model中，我们假定一篇文档只有一个主题生成，可实际中，一篇文章往往有多个主题，只是这多个主题各自在文档中出现的概率大小不一样。比如介绍一个国家的文档中，往往会分别从教育、经济、交通等多个主题进行介绍。那么在pLSA中，文档是怎样被生成的呢？

假定你一共有K个可选的主题，有V个可选的词，咱们来玩一个扔骰子的游戏。

一、假设你每写一篇文档会制作一颗K面的“文档-主题”骰子（扔此骰子能得到K个主题中的任意一个），和K个V面的“主题-词项” 骰子（每个骰子对应一个主题，K个骰子对应之前的K个主题，且骰子的每一面对应要选择的词项，V个面对应着V个可选的词）。

比如可令K=3，即制作1个含有3个主题的“文档-主题”骰子，这3个主题可以是：教育、经济、交通。然后令V = 3，制作3个有着3面的“主题-词项”骰子，其中，教育主题骰子的3个面上的词可以是：大学、老师、课程，经济主题骰子的3个面上的词可以是：市场、企业、金融，交通主题骰子的3个面上的词可以是：高铁、汽车、飞机。



二、每写一个词，先扔该“文档-主题”骰子选择主题，得到主题的结果后，使用和主题结果对应的那颗“主题-词项”骰子，扔该骰子选择要写的词。

先扔“文档-主题”的骰子，假设（以一定的概率）得到的主题是教育，所以下一步便是扔教育主题筛子，（以一定的概率）得到教育主题筛子对应的某个词：大学。

上面这个投骰子产生词的过程简化下便是：“先以一定的概率选取主题，再以一定的概率选取词”。

三、最后，你不停的重复扔“文档-主题”骰子和”主题-词项“骰子，重复N次（产生N个词），完成一篇文档，重复这产生一篇文档的方法M次，则完成M篇文档。

上述过程抽象出来即是PLSA的文档生成模型。在这个过程中，我们并未关注词和词之间的出现顺序，所以pLSA是一种词袋方法。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。

反过来，既然文档已经产生，那么如何根据已经产生好的文档反推其主题呢？这个利用看到的文档推断其隐藏的主题（分布）的过程（其实也就是产生文档的逆过程），便是主题建模的目的：自动地发现文档集中的主题（分布）。

文档d和词w是我们得到的样本，可观测得到，所以对于任意一篇文档，其
$P(w_j|d_i)$
是已知的。从而可以根据大量已知的文档-词项信息
$P(w_j|d_i)$
，训练出文档-主题
$P(z_k|d_i)$
和主题-词项
$P(w_j|z_k)$
，如下公式所示：

$$P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)$$

故得到文档中每个词的生成概率为：

$$P(d_i, w_j) = P(d_i)P(w_j|d_i) = P(d_i)\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)$$

由于
$P(d_i)$
可事先计算求出，而
$P(w_j|z_k)$
和
$P(z_k|d_i)$
未知，所以
$\theta = (P(w_j|z_k), P(z_k|d_i))$
就是我们要估计的参数（值），通俗点说，就是要最大化这个θ。

用什么方法进行估计呢，常用的参数估计方法有极大似然估计MLE、最大后验证估计MAP、贝叶斯估计等等。因为该待估计的参数中含有隐变量z，所以我们可以考虑EM算法。详细的EM算法可以参考之前写过的 EM算法 章节。

1.3 LDA模型

事实上，理解了pLSA模型，也就差不多快理解了LDA模型，因为LDA就是在pLSA的基础上加层贝叶斯框架，即LDA就是pLSA的贝叶斯版本（正因为LDA被贝叶斯化了，所以才需要考虑历史先验知识，才加的两个先验参数）。

下面，咱们对比下本文开头所述的LDA模型中一篇文档生成的方式是怎样的：

按照先验概率
$P(d_i)$
选择一篇文档
d_i
。

从狄利克雷分布（即Dirichlet分布）
α
中取样生成文档
d_i
的主题分布
θ_i
，换言之，主题分布
θ_i
由超参数为
α
的Dirichlet分布生成。

从主题的多项式分布
θ_i
中取样生成文档
d_i
第 j 个词的主题
$z_{i,j}$
。

从狄利克雷分布（即Dirichlet分布）
β
中取样生成主题
$z_{i,j}$
对应的词语分布
$\phi_{z_{i,j}}$
，换言之，词语分布
$\phi_{z_{i,j}}$
由参数为
β
的Dirichlet分布生成。

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中，选主题和选词依然都是两个随机的过程，依然可能是先从主题分布{教育：0.5，经济：0.3，交通：0.2}中抽取出主题：教育，然后再从该主题对应的词分布{大学：0.5，老

师：0.3，课程：0.2}中抽取出词：大学。

那PLSA跟LDA的区别在于什么地方呢？区别就在于：

PLSA中，主题分布和词分布是唯一确定的，能明确的指出主题分布可能就是{教育：0.5，经济：0.3，交通：0.2}，词分布可能就是{大学：0.5，老师：0.3，课程：0.2}。 但在LDA中，主题分布和词分布不再唯一确定不变，即无法确切给出。例如主题分布可能是{教育：0.5，经济：0.3，交通：0.2}，也可能是{教育：0.6，经济：0.2，交通：0.2}，到底是哪个我们不再确定（即不知道），因为它是随机的可变化的。但再怎么变化，也依然服从一定的分布，即主题分布跟词分布由Dirichlet先验随机确定。正因为LDA是PLSA的贝叶斯版本，所以主题分布跟词分布本身由先验知识随机给定。

换言之，LDA在pLSA的基础上给这两参数

加了两个先验分布的参数（贝叶斯化）：一个主题分布的先验分布Dirichlet分布

α

，和一个词语分布的先验分布Dirichlet分布

β

。

综上，LDA真的只是pLSA的贝叶斯版本，文档生成后，两者都要根据文档去推断其主题分布和词语分布（即两者本质都是为了估计给定文档生成主题，给定主题生成词语的概率），只是用的参数推断方法不同，在pLSA中用极大似然估计的思想去推断两未知的固定参数，而LDA则把这两参数弄成随机变量，且加入dirichlet先验。

所以，pLSA跟LDA的本质区别就在于它们去估计未知参数所采用的思想不同，前者用的是频率派思想，后者用的是贝叶斯派思想。

怎么确定LDA的topic个数？

1. 基于经验 主观判断、不断调试、操作性强、最为常用。
2. 基于困惑度（主要是比较两个模型之间的好坏）。
3. 使用Log-边际似然函数的方法，这种方法也挺常用的。
4. 非参数方法：Teh提出的基于狄利克雷过程的HDP法。
5. 基于主题之间的相似度：计算主题向量之间的余弦距离，KL距离等

> 部分参考：https://github.com/NLP-LOVE/ML-NLP/tree/master/Machine%20Learning/5.3%20Topic%20Model