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：

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
    (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
    (b) Choose a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

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：
K：主题个数
V：词典中单词的个数

ξ：泊松分布的参数，是个标量

N：生成文档的长度，标量
α：Dirichlet 分布的参数，K 维向量

θ：生成文档的主题分布(多项分布)，K 维向量，在后面θ_k表示θ的第k个元素

w_n：表示生成文档第n个位置的单词，V维向量，如果w_n是单词v，$1 \le v \le V$并且为整数,则w_n的第v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w_n^v，该变量是个标量，取值 0 或 1，若w_n是单词v，则w_n^v为 1，否则为 0，可以把它理解为w_n的第v个元素。

z_n：表示生成文档第n个位置的单词所属于的主题，K维向量，如果z_n是主题k，$1 \le k \le K$并且为整数，则z_n的第k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z_n^k，该变量是个标量，取值 0 或 1，若z_n是主题k，则z_n^k为 1，否则为 0，可以把它理解为z_n的第k个元素。

β：$K \times V$ 矩阵，行是主题，列是单词在词典中的索引。元素 $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$，即从主题$i$生成单词$j$的概率。$\beta$应该是按行正规化，

这些貌似不太重要.

即每一行的元素加和为 1。后面用 β_i 表示 β 的第 i 行。

步骤 1 ，该文档按照参数 ξ 生成文档的长度 N 。在 BleiNJ03 中，这样解释 N

that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Furthermore, note that N is independent of all the other data generating variables (θ and $\mathbf{z}$). It is thus an ancillary variable and we will generally ignore its randomness in the subsequent development.

在后面的模型中，会忽略变量 N 。生成文档的长度取值实际文档的长度。

步骤 2，从参数为 α 的 Dirichlet 分布生成文档的主题分布 θ，

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

步骤 3，对于 N 个位置，依次进行步骤(a)，(b)。

步骤(a)：从参数为 θ 的多项分布生成主题 z_n，

$$p(z_n^k = 1 \mid \theta) = \theta_k$$

步骤(b)：从参数为 β_k 的多项分布生成单词 w_n，

$$p(w_n^v = 1 \mid \beta_k) = \beta_{kv} = p(w_n^v = 1 \mid \beta, z_n^k = 1)$$

对于一篇文档，按照步骤 1~3 生成 θ，z，w 。z 是 $K \times N$ 的矩阵，即所生成文档的所有主题的表示，z_n 为其第 n 个元素。w 是 $V \times N$ 的矩阵，即所生文档所有单词的表示，w_n 为其第 n 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。

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \qquad (1)$$

在公式(1)中，参数是 α 和 β，即我们假设其是已知的，最终要求解它们。而 θ，z 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 w)之间的关系。

因为 θ，z 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 w)的模型(LDA)概率：

$$
\begin{aligned}
& p(w \mid \alpha, \beta) \\
&= \int \sum_{z} p(\theta, z, w \mid \alpha, \beta) d\theta = \int \sum_{z_1} \cdots \sum_{z_N} p(\theta, z, w \mid \alpha, \beta) d\theta \\
&= \int \sum_{z_1} \cdots \sum_{z_N} p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) d\theta \\
&= \int p(\theta \mid \alpha) \left(\sum_{z_1} p(z_1 \mid \theta) p(w_1 \mid z_1, \beta) \right) \sum_{z_2} \cdots \sum_{z_n} \prod_{n=2}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) d\theta \\
&= \int p(\theta \mid \alpha) \left(\sum_{z_1} p(z_1 \mid \theta) p(w_1 \mid z_1, \beta) \right) \cdots \left(\sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta \\
&= \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta
\end{aligned}
$$

一个语料(假设为D,共M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：

$$
p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d \quad (2)
$$

公式(2)也是 LDA 模型我们最终要优化的目标函数。

先总结一下 LDA 模型，

两个参数：α，β

两个隐含变量：θ，z（定义 I）

一个已知变量(或称观测变量)：w

模型(由参数、变量、目标函数定义)我们知道是什么样子了，下面的问题是怎么求解该模型，即求解使公式(2)最大的参数值$\hat{\alpha}$，$\hat{\beta}$。

对于这种带有隐含变量的模型，一般的求解方法是 EM 算法。下面简单看一下 EM 算法。假设一个模型已知变量是x，参数是Λ，隐含变量是h。其似然函数是：

$$
\begin{aligned}
& l(\Lambda; x) = \log p(x \mid \Lambda) \\
&= \log \sum_{h} p(x, h \mid \Lambda) \\
&= \log \sum_{h} q(h \mid x) \frac{p(x, h \mid \Lambda)}{q(h \mid x)} \\
&\geq \sum_{h} q(h \mid x) \log \frac{p(x, h \mid \Lambda)}{q(h \mid x)} \\
&\overset{\Delta}{=} L(q, \Lambda)
\end{aligned}
\quad (3)
$$

上面不等式的原理是 Jensen's inequality。Jensen's inequality 是说对于 concave function f (这里是 log 函数)，$f(E(x)) \geq E(f(x))$

我们将最大化 $l(\Lambda; x)$ 的问题转变为最大化其下界(lower bound) $L(q, \Lambda)$ 的问题。这里的下界并不是 $l(\Lambda; x)$ 的最小值，而是另一个函数($L(q, \Lambda)$)，该函数的曲线始终在原函数($l(\Lambda; x)$)之下。就像把一块布扣在一个倒置的碗上一样，我们求不出布最高的位置的坐标(二维)，通过找出碗最高点的坐标来近似。

EM 算法每次迭代分为两个子步骤：

(E step) $q^{(t+1)} = \arg\max L(q, \Lambda^{(t)})$

(M step) $\Lambda^{(t+1)} = \arg\max L(q^{(t+1)}, \Lambda)$

E-step 我们假设维持 $\Lambda^{(t)}$ 不变，改变 q 来使 L 变大。M-step 我们假设 $q^{(t+1)}$ 不变，改变 Λ 来使 L 变大。可以证明，E-step 和 M-step 都可以使 L 变大或不变，但不会变小。具体证明过程可以参考(Probabilistic graphical models, Jordan or Koller 的 EM 算法部分)。我们还可以证明，如果 $p(h \mid x, \Lambda) = q(h \mid x)$ 的话可以使(3)不等式变为等式。

如果可以计算出 $p(h \mid x, \Lambda)$ 话，我们可以省略 E-step 而直接使用 M-step。

回到我们的 LDA 模型，如果可以求解出 $p(\theta, z \mid w, \alpha, \beta)$ 的话，就可以使优化问题变得相对简单一些(EM 算法里常用方法是在 E-step 求解隐含变量的后验概率，很少有直接优化 E-step 本身的)。

但是：

$$p(\theta, z \mid w, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \phi)}{p(w \mid \alpha, \beta)}$$

其中

$$p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha)\left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta)\right) d\theta$$

是计算不出来的。

所以使用另一个简单的分布 $q(\theta, z \mid \gamma, \phi)$ 来近似 $p(\theta, z \mid w, \alpha, \beta)$。通过公式(3)可以看出，使用 $q(\theta, z \mid \gamma, \phi)$ 近似虽然不能求解出 $l(\Lambda; x)$ （这里是 $l(\alpha, \beta; w) = \log p(w \mid \alpha, \beta)$ ）真实的最大值，但能够保证求解出它的一个下界的最大值。

这种方法称为变分方法，这是因为 γ 和 ϕ 都是 $p(\theta, z \mid w, \alpha, \beta)$ 中没有的，这种方法

的效果是在目标函数中增加了新的参数。γ 和 ϕ 都是文档级的参数(相比之下，α 和 β 是语料级的参数)。γ 是一个 K 维向量，是生成该文档 Dirichlet 分布的参数。ϕ 是一个二维矩阵，行是单词的位置（或者说文档一个单词位置）的索引，列是主题的索引，元素 ϕ_{ni} 表示文档的第 n 个位置的单词属于主题 i 的概率。

$$q(\theta, z \mid \gamma, \phi) = q(\theta \mid \gamma)\prod_{n=1}^{N} q(z_n \mid \phi_n) = q(\theta \mid \gamma)q(z \mid \phi) \tag{4}$$

根据公式(3)中所示，对于一篇文档我们的优化目标由 $p(w \mid \alpha, \beta)$ 变成下面的函数：

$$\begin{aligned}
&\log p(w \mid \alpha, \beta) = \log \int \sum_z p(\theta, z, w \mid \alpha, \beta) d\theta \\
&= \log \int \sum_z \frac{p(\theta, z, w \mid \alpha, \beta) q(\theta, z)}{q(\theta, z)} d\theta \\
&\geq \int \sum_z q(\theta, z) \log p(\theta, z, w \mid \alpha, \beta) d\theta - \int \sum_z q(\theta, z) \log q(\theta, z) d\theta \\
&= E_q[\log p(\theta, z, w \mid \alpha, \beta)] - E_q[\log q(\theta, z)] \\
&= L(w \mid \{\alpha, \beta\}, \{\gamma, \phi\}) = L(\gamma, \phi; \alpha, \beta)
\end{aligned} \tag{5}$$

$L(\gamma, \phi; \alpha, \beta)$ 只是一种写法，该函数的参数有四个 α ，β (原有参数)和 γ ，ϕ (变分参数)。将公式(1)和(4)带入公式(5)中，$L(\gamma, \phi; \alpha, \beta)$ 变为下面的形式：

$$\begin{aligned}
L(\gamma, \phi; \alpha, \beta) = {} & E_q[\log p(\theta \mid \alpha)] + E_q[\log p(z \mid \theta)] + E_q[\log p(w \mid z, \beta)] \\
& - E_q[\log q(\theta)] - E_q[\log q(z)] \\
& = \log \Gamma(\textstyle\sum_{j=1}^{K} \alpha_j) - \displaystyle\sum_{i=1}^{K} \log \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\textstyle\sum_{j=1}^{k} \gamma_j)) \\
& + \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{ni} (\Psi(\gamma_i) - \Psi(\textstyle\sum_{j=1}^{k} \gamma_j)) \\
& + \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{ni} w_n^j \log \beta_{ij} \\
& - \log \Gamma(\textstyle\sum_{j=1}^{K} \gamma_j) + \displaystyle\sum_{i=1}^{K} \log \Gamma(\gamma_i) - \sum_{i=1}^{K} (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\textstyle\sum_{j=1}^{K} \gamma_j)) \\
& - \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{ni} \log \phi_{ni}
\end{aligned} \tag{6}$$

后五行分别是五个 E_q 分解开的结果，拿第一个来看看是怎么分解的。

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_k - 1}$$

$$E_q[\log p(\theta \mid \alpha)] = \int \sum_z q(\theta, z \mid \gamma, \phi) \log p(\theta \mid \alpha) d\theta$$

$$= \int \sum_z q(\theta \mid \gamma) q(z \mid \phi) \log p(\theta \mid \alpha) d\theta$$

$$= \int q(\theta \mid \gamma) \log p(\theta \mid \alpha) \sum_z q(z \mid \phi) d\theta$$

$$= \int q(\theta \mid \gamma) \log p(\theta \mid \alpha) d\theta$$

$$= \int q(\theta \mid \gamma) \log \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \mathrm{L}\, \theta_K^{\alpha_K - 1} d\theta$$

$$= \int q(\theta \mid \gamma) \left(\log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \log \theta_i \right) d\theta$$

$$= \int q(\theta \mid \gamma) \left(\log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) \right) d\theta + \int q(\theta \mid \gamma) \left(\sum_{i=1}^{K} (\alpha_i - 1) \log \theta_i \right) d\theta$$

$$= \log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \int q(\theta \mid \gamma) \log \theta_i d\theta$$

$$= \log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) E_{q(\theta|\gamma)}[\log \theta_i] d\theta$$

(7)

上式中需要求解的一个量是 $E_{q(\theta|\gamma)}[\log \theta_i]$，在 BleiNJ03 A1 中，若

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \mathrm{L}\, \theta_K^{\alpha_K - 1}$$

则 $E_{p(\theta|\alpha)}[\log \theta_i \mid \alpha] = \Psi(\alpha_i) - \Psi(\sum_{j=1}^{K} \alpha_j)$

这里比较有意思的一个问题是，从参数为 α 和参数为 γ 的 Dirichlet 分布我们都可以生成 θ 。那这里的 $E_{q(\theta|\gamma)}[\log \theta_i]$ 究竟应该是 $E_{q(\theta|\gamma)}[\log \theta_i \mid \alpha]$ 还是 $E_{q(\theta|\gamma)}[\log \theta_i \mid \gamma]$ 呢? 我们可以分析一下 $E_{q(\theta|\gamma)}[\log \theta_i]$，因为

$$E_{q(\theta|\gamma)}[\log \theta_i] = \int q(\theta \mid \gamma) \log \theta_i d\theta$$

因为我们要对 θ 求积分，那就是说 θ 是变化的，会取尽所有合法值。$q(\theta \mid \gamma)$ 说明，θ 是随着 γ 变化的，而不是随着 α 变化的。所以这里应该是 $E_{q(\theta|\gamma)}[\log \theta_i \mid \gamma]$，而不是 $E_{q(\theta|\gamma)}[\log \theta_i \mid \alpha]$。

所以可以继续公式(7)

$$= \log\Gamma\left(\sum_{i=1}^{K}\alpha_i\right) - \sum_{i=1}^{K}\log\Gamma(\alpha_i) + \sum_{i=1}^{K}(\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)\right)$$

其它四个 E_q 可以用类似的过程分解,最终得到公式(6)。

再总结一下 LDA 模型，在（定义 I）中，LDA 模型的定义如下：

目标函数：

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} p(w_d \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta)\right) d\theta_d$$

参数和变量：

两个参数：α ， β

两个隐含变量：θ ， z

一个已知变量(或称观测变量)：w

然而我们无法优化上述模型，所以采用了目标函数的一个下界来近似它，并且采用了变分的方法。这样我们的模型就变成了如下形式：

目标函数：

$$l(D \mid \alpha, \beta; \gamma, \phi) = \sum_{d=1}^{D} l(w_d \mid \alpha, \beta; \gamma_d, \phi_d)$$

$$= \sum_{d=1}^{D} \log p(w_d \mid \alpha, \beta; \gamma_d, \phi_d)$$

$$\geq \sum_{d=1}^{D} L(\alpha, \beta; \gamma_d, \phi_d)$$

即我们的目标函数是 $\sum_{d=1}^{D} L(\alpha, \beta; \gamma_d, \phi_d)$

参数和变量：

四个参数：一般参数 α ， β 和变分参数 γ ， ϕ

两个隐含变量：θ ， z （定义 II）

一个已知变量(或称观测变量)：w

虽然采用了变分方法，但是参数的求解过程依然采用 EM 算法。

E-Step ：

优化 $q(\theta, z \mid \gamma, \phi)$ ，通过优化 γ ， ϕ 实现

因为γ，ϕ是文档级的参数，所以 M-Step 是逐篇文档进行。

对于一篇文档，目标函数$L(\gamma,\phi;\alpha,\beta)$如公式(6)所示，

$L(\gamma,\phi;\alpha,\beta)$中包含$\phi_{ni}$的部分是：

$$L_{[\phi_{ni}]}=\phi_{ni}(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{K}\gamma_j))+\phi_{ni}\log\beta_{iv}-\phi_{ni}\log\phi_{ni}+\lambda_n(\sum_{j=1}^{K}\phi_{nj}-1)$$

最后一项是拉格朗日因子，ϕ_{ni}是个标量。

$L_{[\phi_{ni}]}$对ϕ_{ni}求导数，结果如下：

$$\frac{\partial L}{\partial \phi_{ni}}=\Psi(\gamma_i)-\psi(\sum_{j=1}^{K}\gamma_j)+\log\beta_{iv}-\log\phi_{ni}-1+\lambda_n$$

令导数为 0 得：

$$\phi_{ni}\propto\beta_{iv}\exp(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{K}\gamma_j))$$

$L(\gamma,\phi;\alpha,\beta)$中只包含$\gamma$的部分是

$$L_{[\gamma]}=\sum_{i=1}^{K}(\alpha_i-1)\left(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{K}\gamma_j)\right)+\sum_{n=1}^{N}\phi_{ni}(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{K}\gamma_j))\,\mathrm{i}$$

$$-\log\Gamma(\sum_{j=1}^{K}\gamma_j)+\log\Gamma(\gamma_i)-\sum_{i=1}^{K}(\gamma_i-1)(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{K}\gamma_j))$$

$$=\sum_{i=1}^{K}\left(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{K}\gamma_j)\right)\left(\alpha_i+\sum_{n=1}^{N}\phi_{ni}-\gamma_i\right)-\log\Gamma(\sum_{j=1}^{K}\gamma_j)+\log\Gamma(\gamma_i)$$

$$\frac{\partial L}{\partial \gamma_i}=\Psi'(\gamma_i)\left(\alpha_i+\sum_{n=1}^{N}\phi_{ni}-\gamma_i\right)-\Psi'(\sum_{j=1}^{K}\gamma_j)\sum_{j=1}^{K}(\alpha_j+\sum_{n=1}^{N}\phi_{nj}-\gamma_j)$$

设置导数为 0，得：

$$\gamma_i=\alpha_i+\sum_{n=1}^{N}\phi_{ni}$$

M-Step ：优化参数α ，β (它们是语料级参数)

$$L_{[\beta]}=\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{K}\sum_{j=1}^{V}\phi_{dni}w_{dn}^{j}\log\beta_{ij}+\sum_{i=1}^{K}\lambda_i(\sum_{j=1}^{V}\beta_{ij}-1)$$

求解导数并设置为 0，得：

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}$$

$$L_{[\alpha]} = \sum_{d=1}^{M} \left(\log \Gamma(\sum_{j=1}^{K} \alpha_j) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) + \sum_{i=1}^{K} \left((\alpha_i - 1) \left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{K} \gamma_{dj}\right) \right) \right) \right)$$

$$\frac{\partial L}{\partial \alpha_i} = M \left(\Psi(\sum_{j=1}^{K} \alpha_j) - \Psi(\alpha_i) \right) + \sum_{d=1}^{M} \left(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{K} \gamma_{dj}) \right)$$

$$\frac{\partial L}{\partial \alpha_i \alpha_j} = \delta(i, j) M \Psi'(\alpha_i) - \Psi'(\sum_{j=1}^{K} \alpha_j)$$

α 不能准确地解出来，所以但其导数和曲率容易求解出来，一般采用牛顿方法进行优化。

(注：α ，β 和 γ，ϕ 的计算细节仍需要整理)