**1、主题模型与Bag-Of-Words（或N-gram）的文本表示模型有什么优点？**

Bag-Of-Words（或N-gram）的文本表示模型就是无法识别出不同的词（或词组）具有相同主题的情况。我们需要一种技术能够将具有相同主题的词（或词组）映射到同一维度上去，于是产生了主题模型(Topic Model)。

**2、依据什么原理判定两个不同的词具有相同的主题？**

两个词可能有更高的概率出现在同一主题的文档中；换句话说，给定某一主题，这两个词的产生概率都是比较高的，而另一些不太相关的词产生的概率则是较低的。

**3、主题模型如何表示主题向量？主题模型要解决什么问题？**

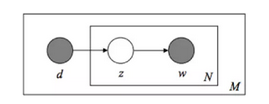
假设有K个主题，我们可以把任意文章表示成一个K维的主题向量，其中向量的每一维代表一个主题，权重代表这篇文章属于该主题的概率。主题模型所解决的事情，就是从语料库中发现有代表性的主题（得到每个主题上面词的分布），并且计算出每篇文章对应着哪些主题。这样具有相似主题的文章拥有相似的主题向量表示，从而能够更好地表示文章的语义，提高文本分类、信息检索等应用的效果。

**4、 常见的主题模型有哪些？试介绍其原理。**

常用的主题模型当属pLSA和LDA，下面分别介绍其原理：

(1) pLSA

pLSA(probabilistic Latent Semantic Analysis)用一个生成模型来建模文章的生成过程。假设有K个主题，M篇文章；对语料库中的任意文章d, 假设该文章有N个词，则对于其中的每一个词, 我们首先选择一个主题z, 然后在当前主题的基础上生成一个词w。这一过程表示成图模型(Graphical Model)如下图所示：



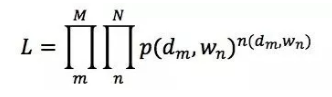
生成主题z和词w的过程遵照一个确定的概率分布。设在文章d中生成主题z的概率为p(z|d), 在选定主题的条件下生成词w的概率为p(w|z)，则给定文章d，生成词w的概率可以写成：

clipboard.png

假设给定主题z的条件下，生成词w的概率是与特定的文章无关的，则公式可以简化为：

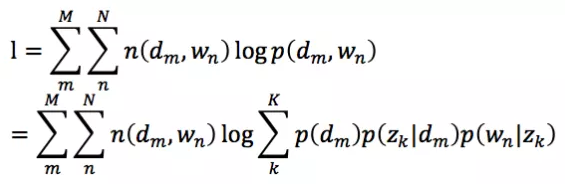
clipboard.png

整个语料库中的文本生成概率可以用以下公式表示，我们称之为似然函数（Likelihood Function）：



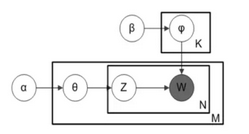
其中*p*(*d*m, *w*n)是在第m篇文章中，第n个单词为*w*n的概率，与上文中*p*(*w*|*d*)的含义是相同的，只是换了一种符号表达。*n*(*d*m, *w*n)表示单词*wn*在文章*d*m中出现的次数。

于是，对数似然函数可以写成：



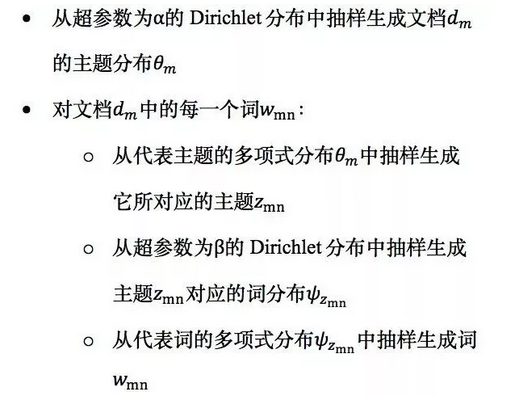
在上面的公式中，定义在文章上的主题分布*p*(*zk*|*dm*)和定义在主题上的词分布*p*(*wn*|*zk*)是待估计的参数 。我们需要找到最优的参数，使得整个语料库的对数似然函数最大化。由于参数中包含的*zk*是隐含变量（即无法直接观测到的变量），因此无法用最大似然估计直接求解，可以利用EM（Expectation-Maximization）算法来解决。

（2）LDA(Latent Dirichlet Allocation)[2]可以看作是pLSA的贝叶斯版本，其文本生成过程与pLSA基本相同，不同的是为主题分布和词分布分别加了狄利克雷(Direchlet)先验。为什么要加入狄利克雷先验呢？这就要从频率学派和贝叶斯学派的区别说起。pLSA采用的是频率派思想，将每篇文章对应的主题分布*p*(*zk*|*dm*)和每个主题对应的词分布*p*(*wn*|*zk*)看成确定的未知常数，并可以求解出来；而LDA采用的是贝叶斯学派的思想，认为待估计的参数（主题分布和词分布）不再是一个固定的常数，而是服从一定分布的随机变量。这个分布符合一定的先验概率分布（即Dirichlet分布），并且在观察到样本信息之后，可以对先验分布进行修正，从而得到后验分布。LDA之所以选择Dirichlet分布做为先验分布，是因为它是多项式分布的共轭先验概率分布，后验概率依然服从Dirichlet分布，这样做可以为计算带来便利。LDA的图模型表示如下：



其中α，β分别为两个Dirichlet分布的超参数，为人工设定。语料库的生成过程如下。

对文本库中的每一篇文档*dm*：



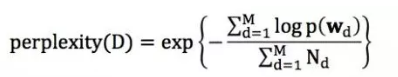
这里主题分布*θm*以及词分布

是待估计的参数，可以用吉布斯采样（Gibbs Sampling）求解其期望。具体做法为，首先随机给定每个词的主题，然后在其它变量固定的情况下，根据转移概率抽样生成每个词的新主题。对于每个词来说，转移概率可以理解为：给定文章中的所有词以及除自身以外其它所有词的主题，在此条件下该词对应各个新主题的概率。最后，经过反复迭代，我们可以根据收敛后的采样结果计算主题分布和词分布的期望。

**5、如何确定LDA模型中的主题个数?**

在LDA中，主题的个数K是一个预先指定的超参数。对于模型超参数的选择，实践中的做法一般是将全部数据集分成训练集、验证集、和测试集3部分，然后利用验证集对超参数进行选择。例如，在确定LDA的主题个数时，我们可以随机选取60%的文档组成训练集，另外20%的文档组成验证集，剩下20%的文档组成测试集。我们在训练时尝试多组超参数的取值，并在验证集上检验哪一组超参数所对应的模型取得了最好的效果。最终，在验证集上效果最好的一组超参数和其对应的模型将被选定，并在测试集上进行测试。

为了衡量LDA模型在验证集和测试集上的效果，我们需要寻找一个合适的评估指标。一个常用的评估指标是困惑度（perplexity）。在文档集合D上，模型的困惑度被定义为：



其中M为文档的总数，**w**d为文档d中单词所组成的词袋向量，p(**w**d)为模型所预测的文档d的生成概率，Nd为文档d中单词的总数。

一开始，随着主题个数的增多，模型在训练集和验证集的困惑度呈下降趋势，但是当主题数目足够大的时候，会出现过拟合，导致困惑度指标在训练集上继续下降但在验证集上反而增长。这时，我们可以取困惑度极小值点所对应的主题个数作为超参数。实践中，困惑度的极小值点可能出现在主题数目非常大的时候，然而实际应用并不能承受如此大的主题数目，这时就需要在实际应用中合理的主题数目范围内进行选择，比如选择合理范围内困惑度的下降明显变慢（拐点）的时候。

另外一种方法是在LDA基础之上融入分层狄利克雷过程（Hierarchical Dirichlet Process，HDP）[3]，构成一种非参数主题模型HDP-LDA。非参数主题模型的好处是不需要预先指定主题的个数，模型可以随着文档数目的变化而自动对主题个数进行调整；它的缺点是在LDA基础上融入HDP之后使得整个概率图模型更加复杂， 训练速度也更加缓慢，因此在实际应用中还是经常采用第一种方法确定合适的主题数目。