LDA算法处理思路

通过海量的文章建立词料库，词料库是一个一维数组，每个元素代表一个词语，这些元素都不相同

先为每一篇文章中每一个词随机指定一个主题（序号），以文章为单位遍历文章中每一个词，根据一个数学公式估算这个词在每一个主题下的概率，通过Gibbs Sampling算法得出每个词是属于哪个主题，记录这个主题下这个词语的次数，这个文章中出现这个主题的次数（知道词语属于哪个主题，这个词语出现就代表着这个主题出现）。经过足够多的训练后，就可以知道在这个数据下，主题下不同词语出现得次数，就可以统计一个主题下不同词语出现的概率，同理也可以得到文章-主题概率

Gibbs Sampling算法：

1的概率是0.3,2的概率是0.3,3的概率是0.4.，那么我们可以随机取一个在0~1直接的数字，如果这个数字在0~0.3表明取的是1，如果是0.3到0.6，那么可以代表数字是2,。如果是0.6到1，那么可以代表这个数字是3。