**LDA算法**

**解决的问题**

LDA是一种非监督学习技术，可以用来对文本进行聚类，提取文档的隐含主题

**步骤**

LDA是这样考虑问题的，它认为我们的所有文档都是由一定的主题组成，不同的文档的主题之间组成的概率是不一样的，比如先对同一个主题，在这篇文档中的概率中很可能是0.2，但在另一个文档中它可能是0.3。每个主题都是由一定的词语组成，每个词语的概率是不一样的，但是它们的和肯定是1。生成文档时就先选择一个主题分布，再按照一定的概率选取一个主题，然后按照一定的概率从主题中选择词语，经过多次选择就可以组成一篇文章。

**使用**

通过海量的数据来训练模型，可以得到一个文档-主题矩阵，主题-词语矩阵，矩阵中的每个元素就代表着一种概率。通过文档-主题矩阵，我们可以猜测该文档是属于哪一个主题这样就能进行聚类。通过训练好的model，我们还可以很快知道它的主题分布。