LDA是一种聚类算法。

LDA（Latent Dirichlet Allocation，隐含狄利克雷分布）是一种文档主题生成模型，也称为一个三层[贝叶斯](https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF)概率模型，包含**词（word）**、**主题（topic）**和**文档（document）**三层结构。

狄利克雷分布 <https://en.wikipedia.org/wiki/Dirichlet_distribution>

所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“**以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语**”这样一个过程。

文档到主题服从多项式分布，主题到词服从多项式分布。

LDA是一种非监督学习，可以用来识别大规模文档集（document collection）或语料库（corpus）中潜藏的主题信息，它将每一篇文档视为一个**词频向量**，从而将文本信息转化为了易于建模的数字信息。忽略词之间的顺序。