作者分析（authorship analysis）是检查文档以提取文档的作者信息的过程。作者分析由几种作者身份主体构成，包括作者身份验证、剽窃检测和作者归属等。

基于作者身份的文档分解(ABDD， authorship-based document decomposition)，将多作者文档中的句子分组到不同的类中，每个类只包含一个作者写的句子。

实验思路：1.捕获文档中连续元素之间的依赖关系，以识别不同的作者组件，并构造一个用于分类的HMM（Hidden Markov Model）；2.使用无监督的学习方法来估计模型的初始参数。

实验目的：给定一个由N个共同作者编写的多作者文档C，假设文档中的每个句子都是由N个共同作者中的一个编写的。此外，每个合著者都在文件中写了长连续的句子序列。作者的数量N是事先知道的，而通常没有关于文档上下文和共同作者的信息。目标是定义由每个合著者编写的文档中的句子。

实验步骤：1.用一种新的无监督学习方法估计HMM参数的初值；2.使用Baum-Welch算法学习HMM的参数；3.应用Viterbi算法找出每个句子最有可能的作者。

隐马尔科夫模型HMM：

使用HMM模型的问题一般具有这两个特征：

1）我们的问题是基于序列的，比如时间序列，或者状态序列。

2）我们的问题中有两类数据，一类序列数据是可以观测到的，即观测序列；而另一类数据是不能观察到的，即隐藏状态序列，简称状态序列。

对于HMM模型，首先我们假设Q是所有可能的隐藏状态的集合，V是所有可能的观测状态的集合，即：

其中，N是可能的隐藏状态数，M是所有的可能的观察状态数。

对于一个长度为的序列，对应的状态序列,是对应的观察序列，即：

其中，任意一个隐藏状态，任意一个观察状态。

HMM模型有两个重要的假设：

1） 齐次马尔科夫链假设。即任意时刻的隐藏状态只依赖于它前一个隐藏状态。当然这样假设有点极端，因为很多时候我们的某一个隐藏状态不仅仅只依赖于前一个隐藏状态，可能是前两个或者是前三个。但是这样假设的好处就是模型简单，便于求解。如果在时刻的隐藏状态是，在时刻的隐藏状态是是 ，则从时刻到时刻的HMM状态转移概率可以表示为：

这样可以组成马尔科夫链的状态转移矩阵：

2） 观测独立性假设。即任意时刻的观察状态只仅仅依赖于当前时刻的隐藏状态，这也是一个为了简化模型的假设。如果在时刻的隐藏状态是，而对应的观察状态为，则该时刻观察状态在隐藏状态下生成的概率满足：

这样可以组成观测状态生成的概率矩阵：

除此之外，我们需要一组在时刻的隐藏状态概率分布：

其中。

一个HMM模型，可以由隐藏状态初始概率分布，状态转移概率矩阵和观测状态概率矩阵决定。决定状态序列，决定观测序列。因此，HMM模型可以由一个三元组表示如下：。

论文中HMM模型参数初始化的步骤：

1. 将文档分成s段，每段包含30个连续的句子，即第i段就包含了第i个由30个连续的句子组成的段。段数，其中表示文档的总句子数。按30句分割，可以保证用于训练HMM参数的句子足够长，而且分出的段数足够多。

2. 选择在文档中出现两次以上的单词，假设这样的单词共有D个，这样就能生成一个长度为D的集合。对于每一个段，创建一个D维向量，如果单词集合中第i个元素（没有）出现在段中，则向量的第i个元素为1（0）。于是文档C就能生成s个D维二进制向量，记为。

3. 使用多元高斯回合模型（GMMS）将D为向量X聚为N个分量，记为。（）会从GMMS模型中会获得一个标签，来表示这个向量所赋值的高斯分量。

4. 选择文档中至少出现一次的单词，将其存入集合，假设集合的大小为D’。与步骤2相同，生成s个D’为二进制向量，记为。

5. 运用观测序列O’和隐藏状态序列Q’构造隐马尔科夫模型。在论文模型中，O’即为。形式上，是文档C中的第i个D维二进制向量相对应的，代表的是C中的第i个片段。Q’即为观测序列O’所对应的作者。每一个代表的是最可能的作者。将从N个元素的集合中提取取一个标签，以表示它的状态。把所有的状态都赋值给后，就得到了状态转移概率A。在论文模型中只有一个状态序列，于是每个状态的可能性被定义为状态在Q’中出现的次数，即，n = 1,2,…,N

6. 给定X’和全部标签可能值的集合，则可计算出在标签下的特征的条件概率。

7. 将文档C分成句子，表示文档中所有句子的总数。用步骤4的方法给每个句子生成一个二进制集合，得到z个D为二进制向量，记为。用步骤6的方法得到的条件概率来计算B中的初始值，其中表示句子向量中的特征的值。再使用加一平滑来避免A和B为零，并使用概率的对数函数来简化计算。

HMM模型学习的步骤：

1. 用上一小节步骤7的观察序列和隐藏状态序列构建一个HMM模型。

2. 用Baum-Welch算法来学习HMM的参数。算法分两步，E-step和M-step。E-step根据观测序列O求期望作者序列Q，M-step根据状态赋值更新HMM参数。学习过程从HMM参数的初始值开始，在这两个步骤间循环，直到在π、A、B出收敛。

HMM模型在生成观测序列时存在多个状态序列，所以使用Viterbi算法来确定生成观测序列的最佳状态序列。

（维特比算法介绍：<https://www.zhihu.com/question/20136144>）