Newsgroup18828中，每篇文档只属于一个新闻组。先进行预处理，对每篇文档进行文本处理，为后续构造字典、提取特征词做准备，主要是去除停用词、非字母字符以及大小写转换。构造词典，统计每个词出现的总次数，返回<key, value>结构的词典，按key排序，value都大于4，即都是出现次数大于4的词。下面进行特征词选取，再生成20个文件夹，每个文件夹的每篇文档中存放本篇中的特征词，即本文档中在词典中出现的词。创建每次迭代所用的训练样例集合（20个文件夹）和测试样例集合（20个文件夹），并生成标注数据集合。

以上是对20个文件夹中文档数据的基本处理，下面用贝叶斯算法进行分类。

每个测试样例属于某个类别的概率 = 某个类别中出现测试样例中词的概率的乘积 （类条件概率） \* 出现某个类别的概率(先验概率)。

具体计算类条件概率和先验概率时，朴素贝叶斯不仅计算特征词出现或不出现，还要计算出现的次数。

类条件概率 p(word | category) = （类category下单词word出现在所有文档中的次数之和 + 1） / （类category下单词总数 + 训练样本中不重复的特征词总数）。

先验概率 p(category) = 类category下单词总数 / 训练样本中的特征词总数。

首先统计在训练样本中每个目录下每个单词的出现次数以及每个目录下的单词总数，再用贝叶斯算法对测试文档进行分类，计算准确率。