# 实验报告

201844911 杨子玉

1.实验数据来源：[20news-18828.tar.gz](http://qwone.com/~jason/20Newsgroups/20news-18828.tar.gz) - 20 Newsgroups

下载:<http://qwone.com/~jason/20Newsgroups/>

2.相关方法：

1) **朴素贝叶斯分类器**基于一个简单的假定：给定目标值时属性之间相互条件独立。换言之。该假定说明给定实例的目标值情况下。观察到联合的a1,a2...an的概率正好是对每个单独属性的概率乘积： P(a1,a2...an | Vj ) =**Π**i P( ai| Vj )2)VSM：把对文本内容的处理简化为[向量空间](https://baike.baidu.com/item/%E5%90%91%E9%87%8F%E7%A9%BA%E9%97%B4)中的向量运算。通过以上定理和“朴素”的假定，可以知道：

P( Category | Document) = P ( Document | Category ) \* P( Category) / P(Document)

2)**拉普拉斯平滑处理**：零概率问题，就是在计算实例的概率时，如果某个量x，在观察样本库（训练集）中没有出现过，会导致整个实例的概率结果是0。在文本分类的问题中，当一个词语没有在训练样本中出现，该词语调概率为0，使用连乘计算文本出现概率时也为0。这是不合理的，所以使用加1的方法。

3.处理文本数据集:

1)将实验数据分成两部分：80%的data\_train和20%的data\_test

2)对训练集和测试集创建向量[类名，所有单词的长度，出现的概率，字典]

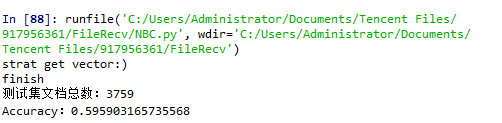
4.进行分类：

对每个待分类的文档，利用公式计算，并统计成功的文件数和失败的文件数，得到准确率

5.实验结果如下图所示

在NB1中采取Homework1中已经分好的训练集和测试集，计算步骤可能出现问题，在NB2中采用Pythonsklearn自带的贝叶斯分类器完成文本分类，使用MultinomialNB，假设特征的先验概率为多项式分布，添加新闻标签10个进行分类，可以看见越多的训练类别得到的准确度越高 ，但没有写一个添加标签的函数，直接进行导入的。

**NB1**



**NB2**

