实验报告

201844911　杨子玉

1.实验数据来源：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 $a_1,a_2...a_n$ 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

```
In [88]: runfile('C:/Users/Administrator/Documents/Tencent Files/
917956361/FileRecv/NBC.py', wdir='C:/Users/Administrator/Documents/
Tencent Files/917956361/FileRecv')
strat get vector:)
finish
测试集文档总数: 3759
Accuracy: 0.595903165735568
```

NB2

```
In [82]: runfile('C:/Users/Administrator/Desktop/Homework/Homework2/
untitled12.py', wdir='C:/Users/Administrator/Desktop/Homework/
Homework2')
训练集数量:  6113
测试集数量:  1529
Accuracy
0.8639633747547416
```

```
In [83]: runfile('C:/Users/Administrator/Desktop/Homework/Homework2/
NB2.py', wdir='C:/Users/Administrator/Desktop/Homework/Homework2')
训练集数量:  7704
测试集数量:  1926
Accuracy
0.8997923156801662
```