NBC

姓名：祝瑶佳　　学号：201834893

项目：使用朴素贝叶斯分类器分类文档_

一、实验方法

1. 将文档分成训练集和测试集
2. 将训练集进行分词处理(去除重复词，去掉符号)得到词典
3. 将每个类的所有文档分词得到词典
4. 将训练集得到的词典与每个类的词典进行比较，如果训练词典中的词在类词典中出现，则记录出现次数，没出现则记为 0，得到训练集对应每个类词频向量
5. 统计每个词在每个类出现的次数和此类中所有词出现的次数，计算得到 p(每个词|每个类)
6. 将测试集文档进行分词处理并得到词典。
7. 比较测试集词典和训练集词典，如果有相同的词，则对应词的 p(每个词|每个类)可以从训练集中延用，可以放入测试集概率向量列表中。
8. 将测试集中词典的所有 p(每个词|每个类)相乘再乘以类概率 pclass 得到该文档属于此类的概率，依照此方法计算出每个测试文档属于各个类的概率
9. 比较每个训练文档属于各个类的概率，找到最大的概率，并将文档归于相应的类。

二、实验任务

1，实现朴素贝叶斯分类器，测试其在 20 Newsgroups 数据集上的效果

三、实验数据

20 Newsgroups

四、实验步骤

1，文本处理

（1）分词并创建词典

读取文档按空格分词，并且去掉符号和重复。将文档划分成单词，

（2）划分训练集和测试集

将数据集划分成训练集和测试集，从每个类抽取 80%为训练集，20%为测试集，组成最终的训练集和测试集。

（3）得到训练集对应每个类词频向量

2，训练贝叶斯分类器

（1） 得到训练集的每个词属于每个类的概率向量

3，测试贝叶斯分类器

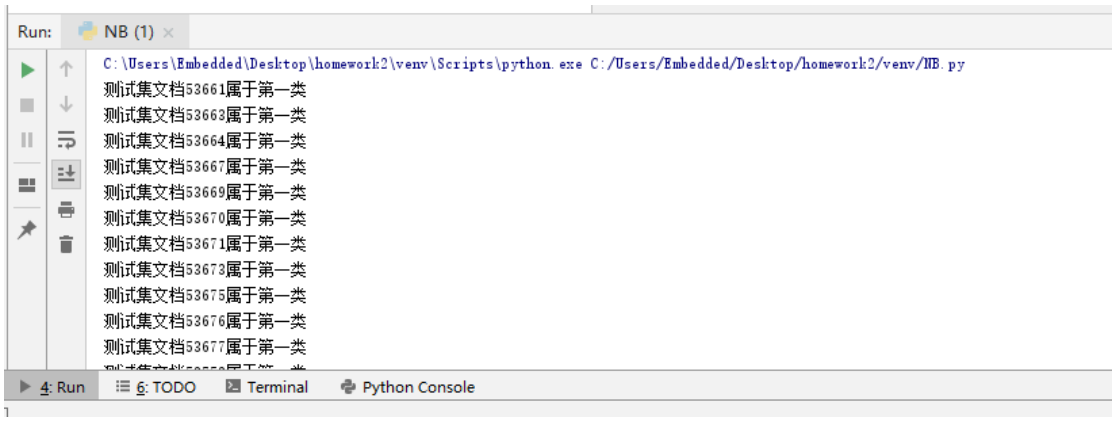
（1） 根据训练效果得到测试集每个词属于每个类的概率向量
（2） 得到每个文档属于每个类的概率并选择概率最大的，将文档归于此类。

五、实验结果

训练集的概率向量：

```
C:\Users\Embedded\Desktop\homework2\venv\Scripts\python.exe C:/Users/Embedded/Desktop/homework2/venv/NB.py
1.04551186327e-07

2.09102372653e-07

8.36409490613e-07

1.25461423592e-06

1.35916542225e-06
```

分类结果：



六、实验结论和感想

词频向量是统计每个词在每个类的所有文档中出现的次数。
概率向量是统计每个词在每个类的所有文档中出现的概率。
贝叶斯分类器的特点就是简单实用，分类效果也不错。