Naive Bayes Classifier

一、 实验目的

通过实验，加深对 Naive Bayes 的理解，锻炼编程能力。

二、 实验要求

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

三、 实验环境

windows 10

python 3.6.5

四、 实验过程

1、计算每个类别中每个单词出现的次数以及每个类别中所有单词的总数。

2、计算条件概率和先验概率，使用多项式模型，使用平滑技术(避免零的情况)和取对数(加速计算)。

3、对文本进行分类，选择概率最大的类别，并将结果保存到文档中。

4、读取保存的结果文档(编号、类别、预测类别)，以空格划分，比较类别和预测类别是否相同，然后计算准确率。

五、 出现的问题及解决

1、git 添加文件夹的问题

（1）在 git add Homework2 时出错：

```
warning: LF will be replaced by CRLF in Homework2/.idea/workspace.xml.
The file will have its original line endings in your working directory
```

原因:路径中存在 / 的符号转义问题,false 就是不转换符号默认是 true，相当于把路径的 / 符号进行转义，这样添加的时候就有问题。CRLF 和 LF 是两种不同的换行格式，git 工作区默认为 CRLF 来作为换行符，所以当我们项目文件里有用的地方使用 LF 作为换行符，这个时候我们再继续 git add 或则 git commit 的时候就会弹出警告，当最终 push 到远程仓库的时候 git 会统一格式全部转化为用 CRLF 作为换行符

解决办法：

```
rm -rf .git // 删除.git
git config --global core.autocrlf false //禁用自动转换
git init //初始化 git 库
git add . // 重新加入
```

（2）git add “文件夹”半天没有反应，出现 lock 文件，百度了一下说删除这个文件，但是在暂停的时候 lock 文件自动消失，一添加就出来，此问题没有解决。文件夹不是空文件夹，并且上传压缩包没有问题，郁闷。。。继续解决

六、实验结果

分类正确的个数： 3042

测试集总文档数： 3772

准确率： 0.8064687168610817