|  |  |
| --- | --- |
| **自然语言处理及应用**  **实验报告** | |
|  | |
|  |  |
| **名称** | 基于朴素贝叶斯的文本分类 |
| **姓名** | 郭淇 |
| **班级** | 软件72班 |
| **学号** | 2174111345 |
| Email | 1325930068@qq.com |
| **日期** | 2021-3-23 |

1. **实验目的**

对一个有20个类别，每个类别1000个文本的数据进行分类，采用朴素贝叶斯分类器分类后再采用准确率进行验证。

1. **实验环境**

Pycharm

Windows10

Python3.8

1. **实验方法**

根据已有的20\_newsgroups数据集，首要步骤是遍历20\_newsgroups中所有的文本文件。此处可以调用os模块，利用os.walk的遍历方案进行文件的遍历。

得到文本文件内容的同时，需要对数据进行预处理，我的处理方案是首先剔除文本中的特殊字符，将非26个字母和’的字符用空格替换，对文本根据空格进行分词，对分词结果运用网上找来的停用词表进行筛选，停用词表采用百度停用词。

然后采用贝叶斯分类器进行分类。

每个文件的介绍：

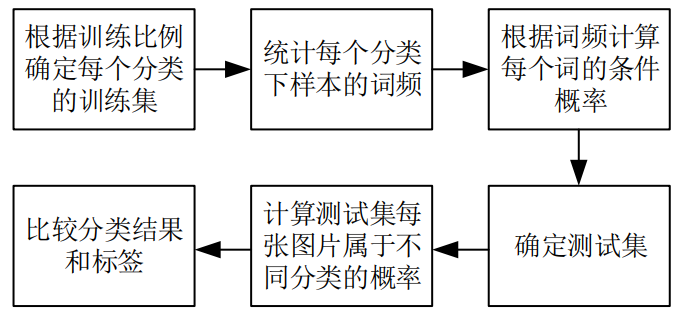
preprocess.py是文件预处理部分，将所有的文件去除停用词和提取词干后通过pickle转换为data.txt

naive\_bayes.py是用来处理转换后的文件的，先统计词频转换为向量，在通过朴素贝叶斯训练模型，最后用百分之二十的数据测试

validation.py是计算模型的准确率的

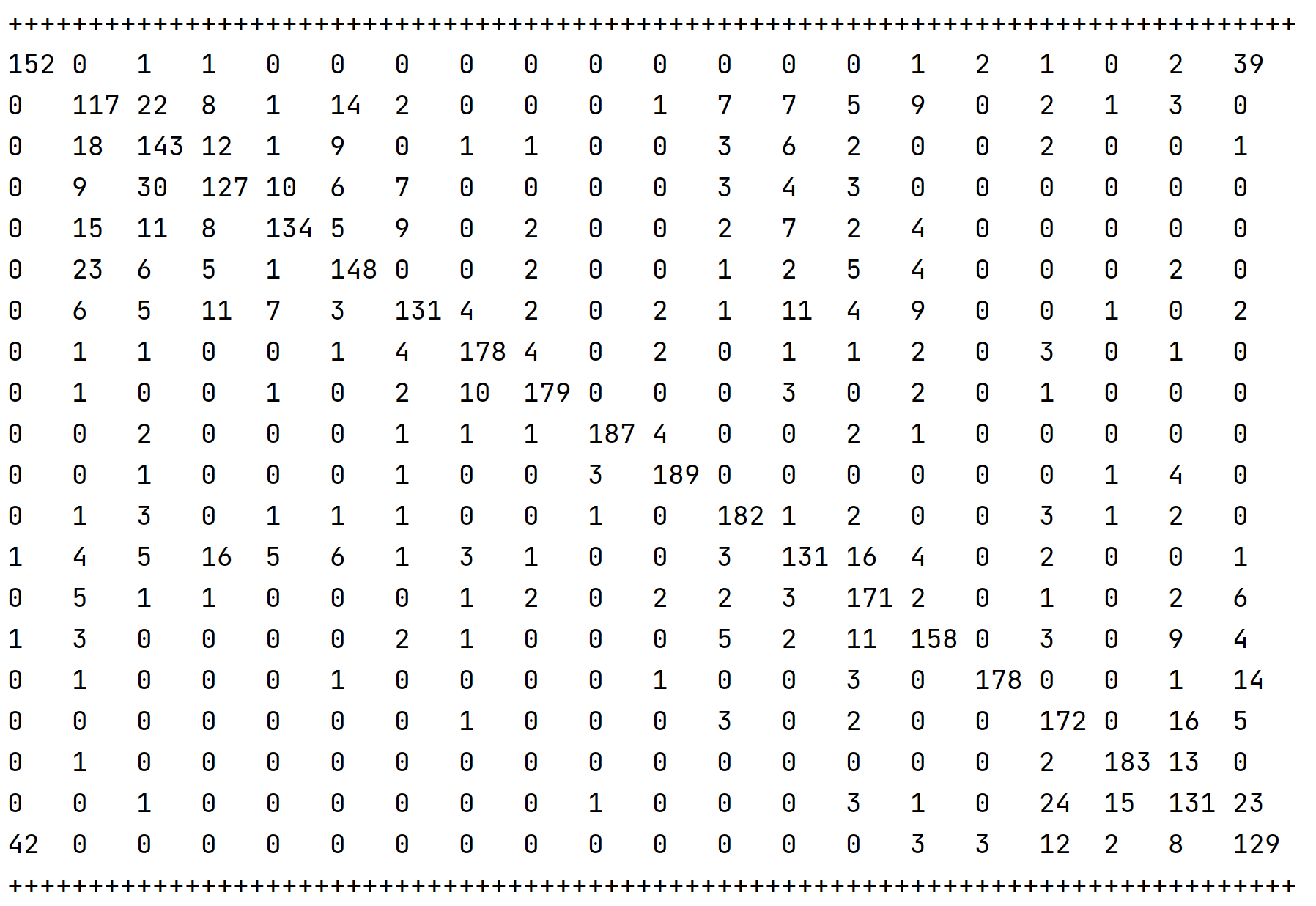
main.py是运行文件

data.txt是转换之后的文件



1. **实验结果**

将测试用的20个类别的4000个文本分类的结果，其中对角线上是分类正确的文本个数，同一行其他的是分类错误的文本个数。



每一个类别分类正确率和所有文本的分类正确率。



可以看出最终分类正确率在百分之八十左右，效果还算可以。

1. **遇到问题及解决思路**

数据量过大，运行时间慢， 采用的方法是通过pickle转换为data.txt。另外，在实验的时候发现，如果没有对数据进行统一转小写处理，反而能得到更高的精度，这有可能是因为没有随机挑选数据导致的结果，也可能是因为大小写本身携带有信息。