DataMining

朴素贝叶斯分类器

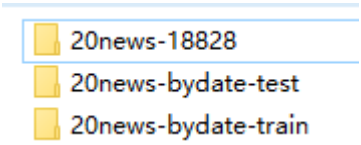
姓名：董潇
学号：201834861
班级：2018 级计算机专硕班
指导老师：尹建华
时间：2018 年 11 月 18 日

一、 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。

二、 实验内容

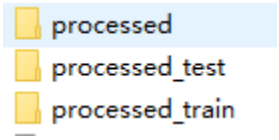
1. 准备好实验所用数据集：



数据集目录

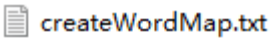
上图中，20news-18828 为总的数据集，20news-bydate-test 与 20news-bydate-train 为以 2:8 分好的测试集以及训练集。

2. 借助 NLTK 的 Python 库对文件进行预处理，包括通过对文本数据分句、分词、去 stopwords、去非字母字符、转化小写等。
在本实验中，借助 processing()函数对上述三个数据集分别进行了预处理，分别生成测试、训练、以及为建立词典做准备的预处理文件即 processed_test、processed_train、processed

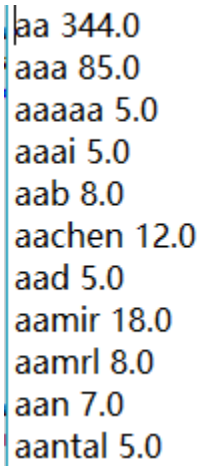


预处理文件目录

3. 创建词典
借助 createWordMap()函数，对预处理生成的为建立词典做准备 processed 文件进行过滤词频处理，生成词典 createWordMap.txt

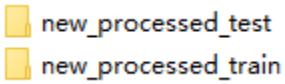


词典文件



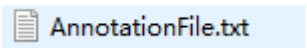
词典小部分内容截图

4. 借助 tokenWords()函数,对训练和测试预处理文件按照词典选取 token，分别得到 new_processed_test、new_processed_train



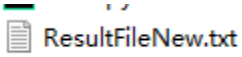
Token 文件截图

5. 创建测试文件的分类标注文件 AnnotationFile.txt，为计算准确率做准备。
   标注：序号 所属类



标注文件

6. 借助 Bayes(traindir,testdir,ResultFileNew)函数,用贝叶斯算法对测试文档分类得到分类结果文件 ResultFileNew.txt



分类结果文件

其中，分别计算条件概率、先验概率：

条件概率 =（类中单词 i 的数目+0.0001）/（类中单词总数+训练样本中所有类单词总数）

先验概率 =（类中单词总数）/（训练样本中所有类单词总数）

最后返回该测试样本在该类别的概率。

7. 将测试集的真实类别与算法册数的类别相比较，计算准确率。

三、 实验结果

```
In [31]: runfile('D:/DataMining/HomeWork2/main.py', wdir='D:/DataMining/HomeWork2')
trainTotalNum:
2065381
accuracy: 0.762917
```

实验结果截图

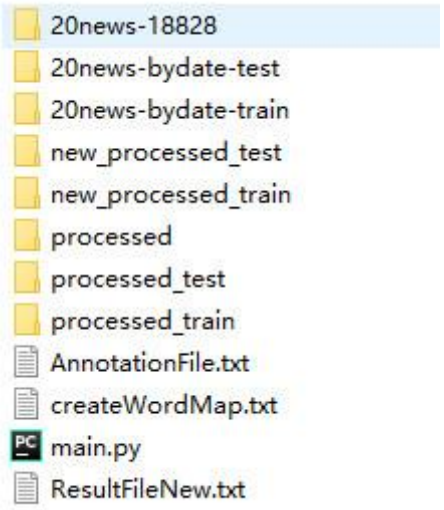
四、 实验总结

因为有了实验一的基础，此次实验的进行顺利了许多，实验一直接借助 sklearn 库函数进行的预处理，本次实验的预处理尝试手写，在此过程中也收获颇多。

此次实验中需要注意的就是数据规模大，在计算概率时，数据相乘容易溢出的问题。通过询问同学，发现可以借助 log 来避免。

五、 附录

总的文件目录截图：



文件目录截图