

DATA Mining

-Homework2

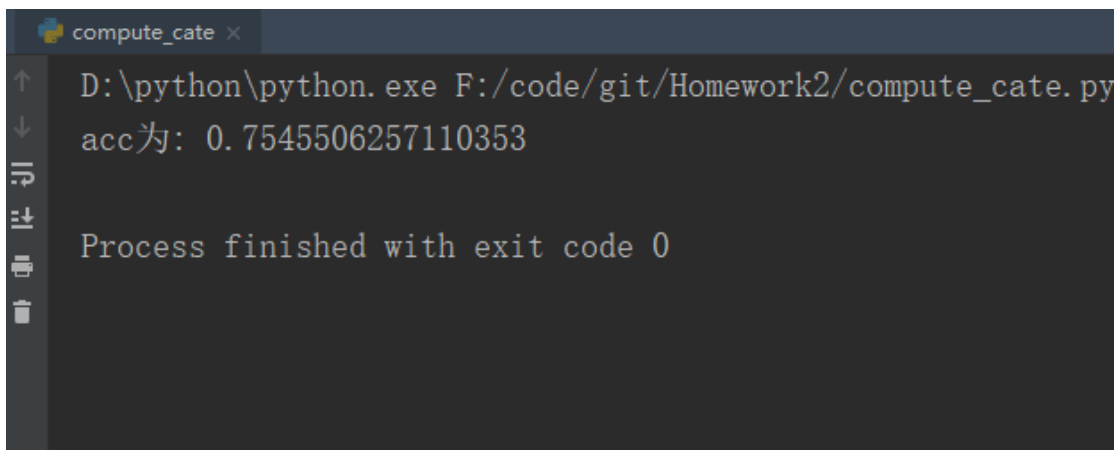
Task :

1. 实现朴素贝叶斯分类器，测试其在 20 Newsgroups 数据集上的效果。
2. The 20 Newsgroups dataset is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.
3. 20news-18828.tar.gz - 20 Newsgroups; duplicates removed, only "From" and "Subject" headers (18828 documents)

Work :

- !!!! 先解压 20news 文件 (是 HomeWork1 处理之后的文件)
1. 首先运行 `creat_testFile()` 创建 test (测试集) 和 train (训练集), 生成比例按照 2:8 生成。
 2. 运行 `NBCprocess()` 对数据集进行朴素贝叶斯分类。
 2. 运行 `computer_acc()` 计算分类的准确率, 测试集和数据集分开。
 3. 最后的分类的准确率为: 0.754

截图:



```
compute_cate x
D:\python\python.exe F:/code/git/Homework2/compute_cate.py
acc为: 0.7545506257110353

Process finished with exit code 0
```