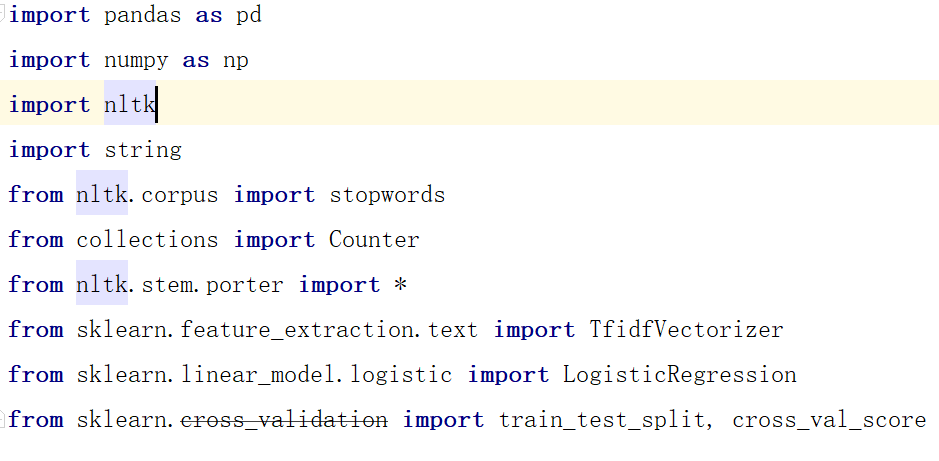
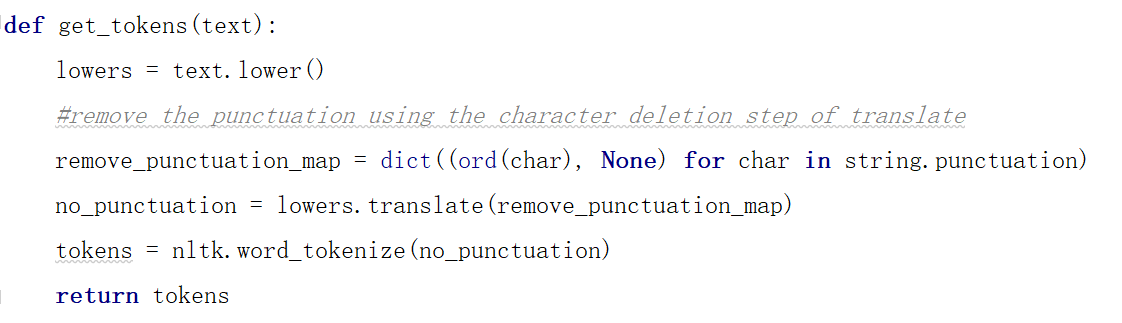
实验报告

姓名：洪语晨 学号：10161511322

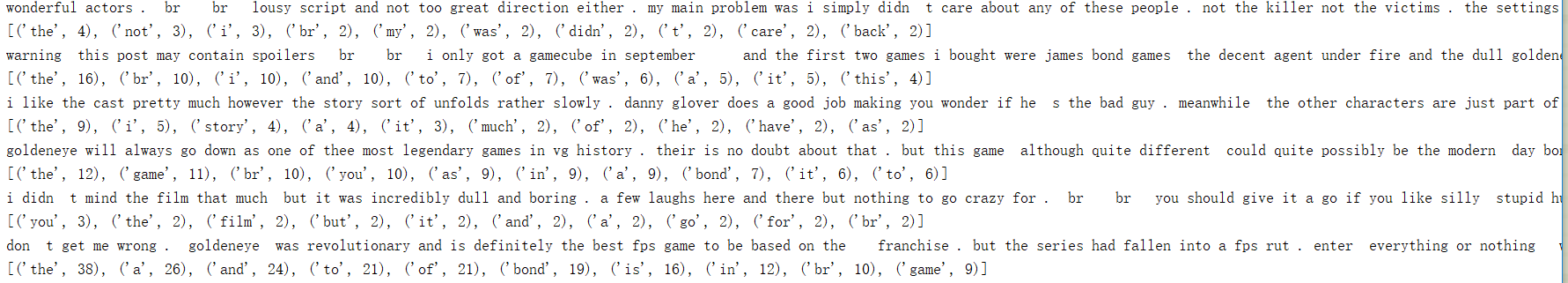
完成该实验引用的包：



1. 对文本进行去标点符号，并将文本转化为均为小写字母：

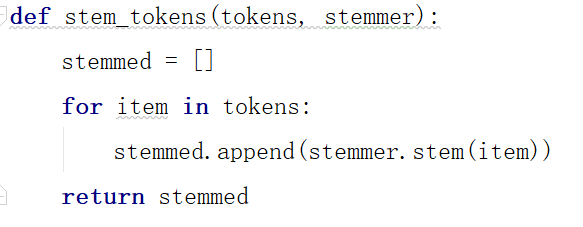


运行结果：

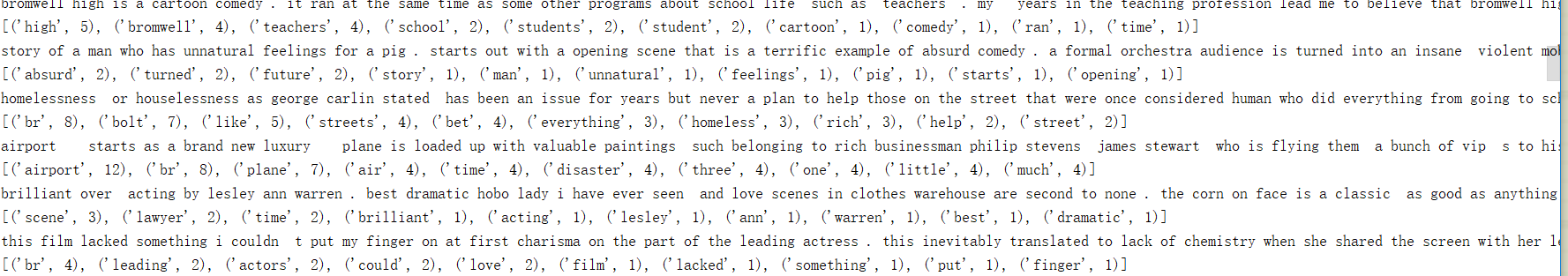


观察可以发现其中包含大量the，it 等大量无实际意义的停用词

1. 去停用词：

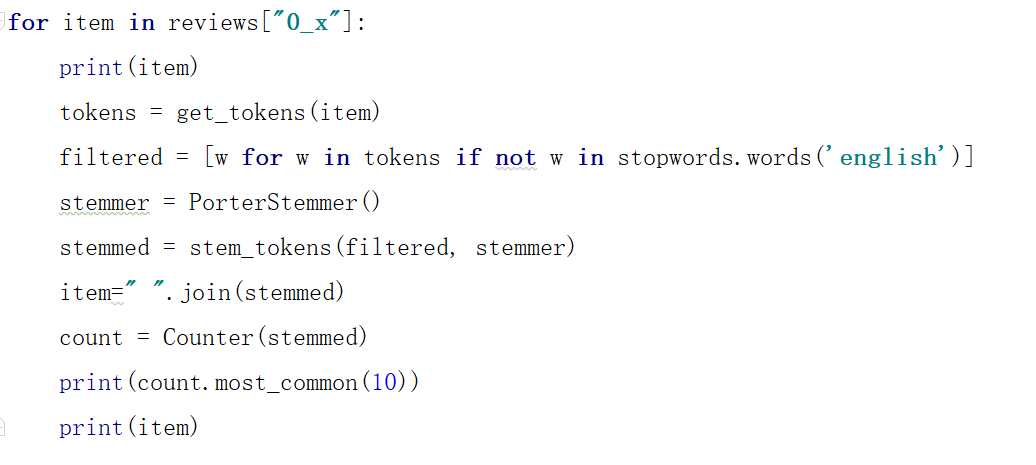


运行结果：



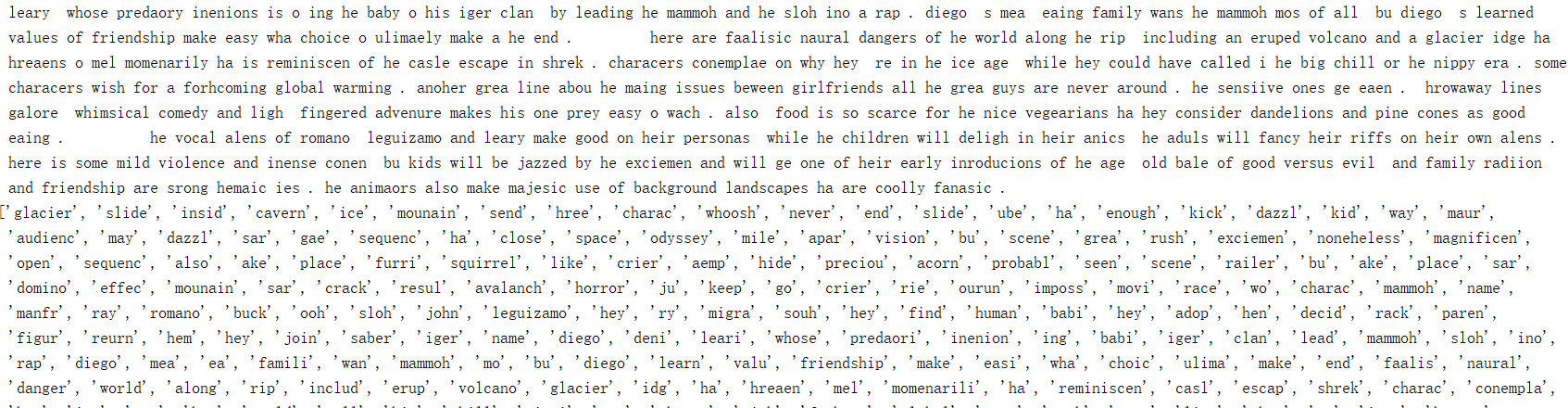
其中有’br’，’t’这种不知道是什么的无意义词，并且同一单词的不同词形会被分开处理

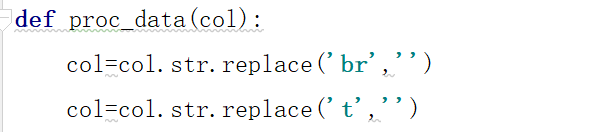
1. Stemming，并且去除’br’,’t’



在代码中，将处理完的单词进行合并，变成句子，并且输出每个样本的前十个数量最多的单词。

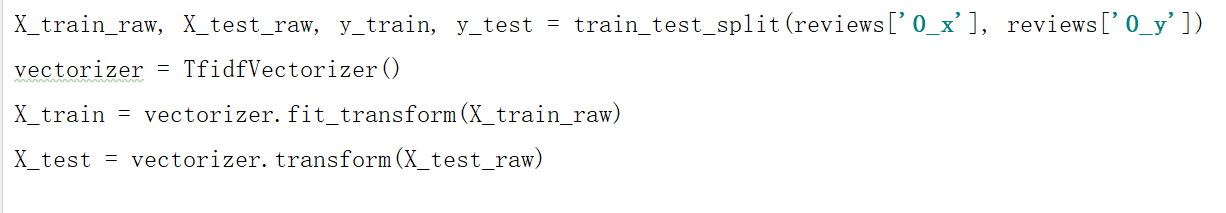
运行结果：



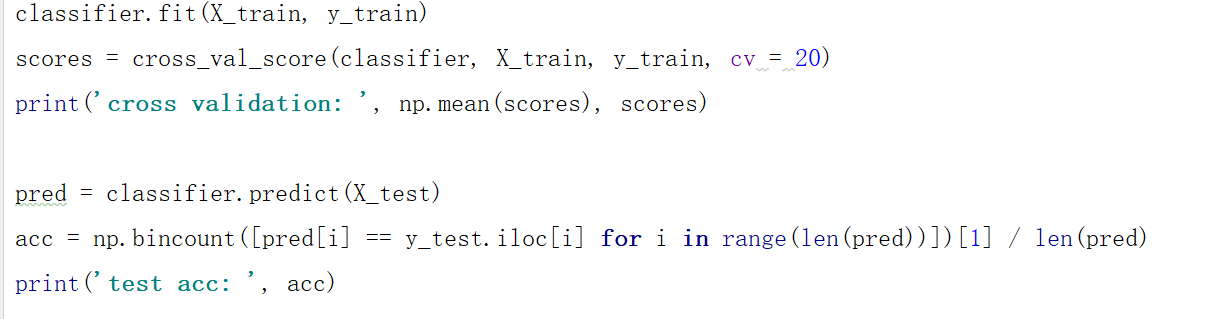


完成预处理

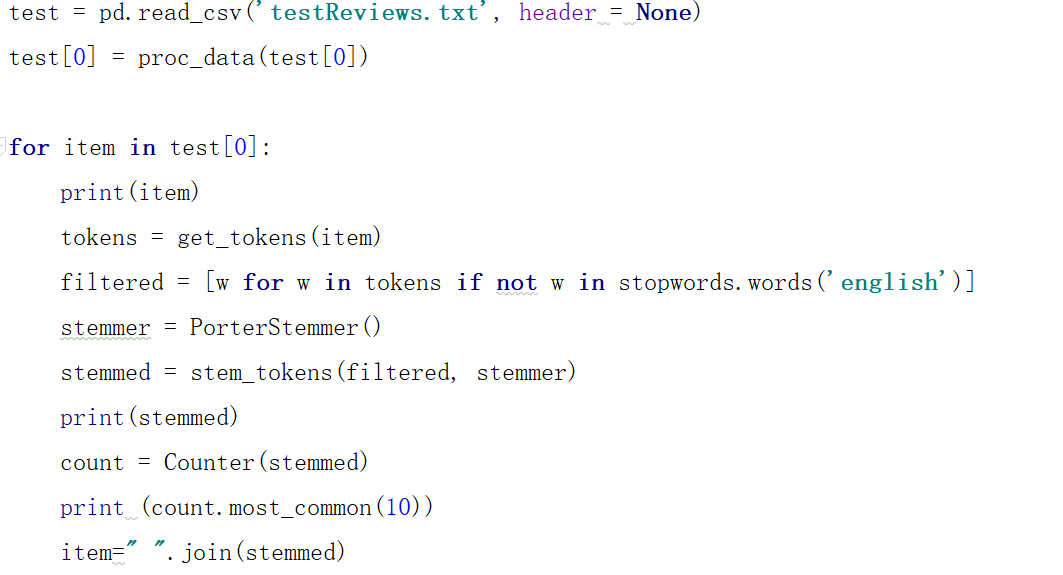
1. 分割训练集，并且对得到的句子进行TF-IDF处理：



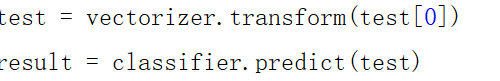
1. 逻辑回归



1. 处理测试集（同训练集处理方法相同）：

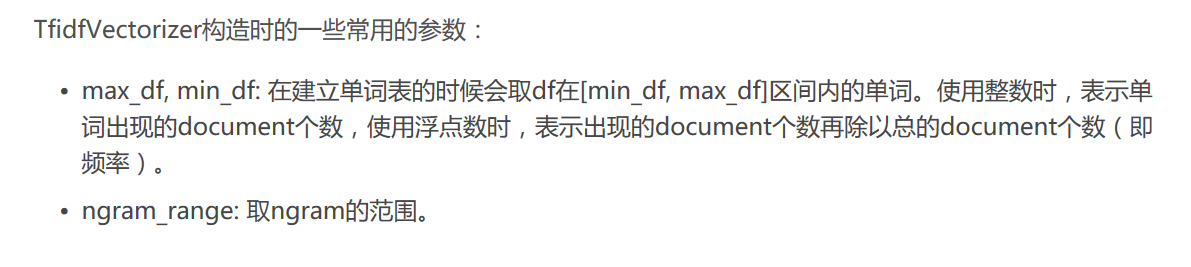


1. 完成评估



1. 改进

1).代码当中引用tfidf时，没有写参数…



修改后的函数：

