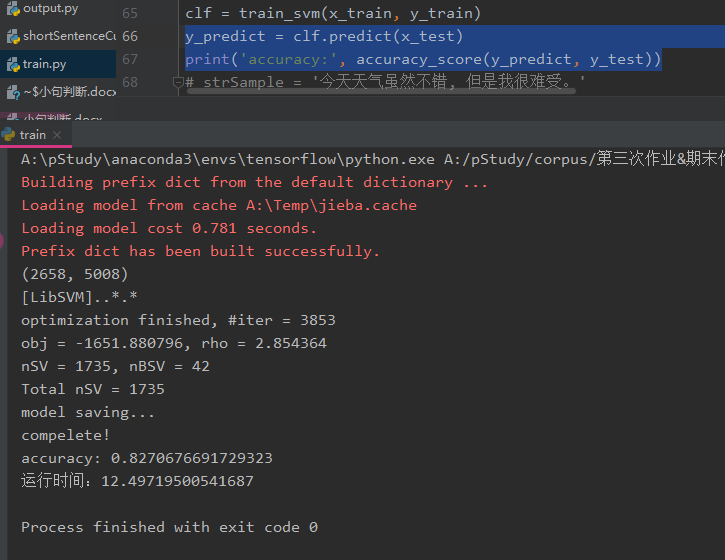
语料库期末-任务1：小句切分

游畅：20181003005

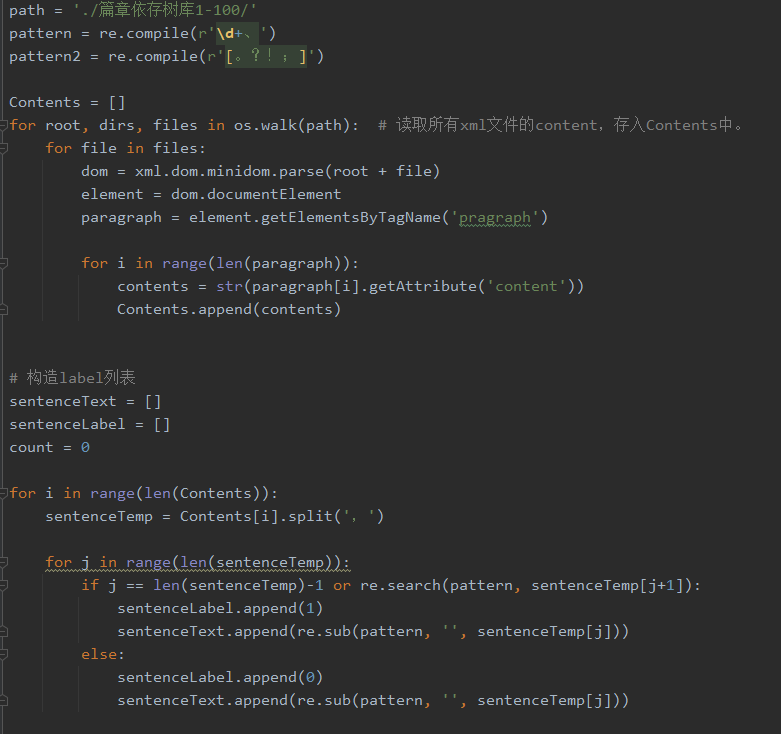
本次任务选择的是SVM分类，其中准确率为：



分类结果输出格式如下：



1 预处理与建立数据集。根据前两次作业给出的xml文档对每一个待切分的逗号所在的小句进行标注：



其中，Contents列表内容为docx的单位对象paragraph；恰好对应xml文档中paragraph父标签，表示一个含有数个小句的自然段。

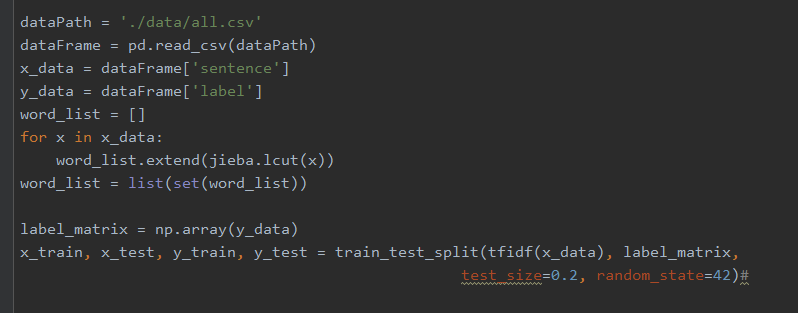
sentenceText为切割了小句标号之后的小句，sentenceLabel为该小句句末是否为分界点的标识。

将其写入csv文件存放于./data，共有pos、neg与all三个csv文档，分别是：句末是分界点的小句、句末非分界点的小句与全部小句及其标注。

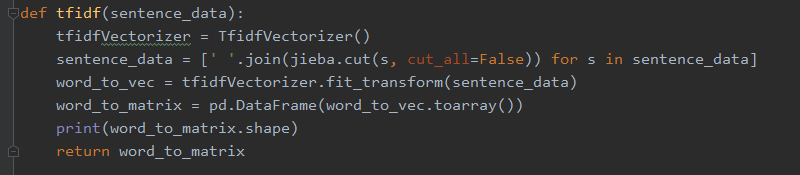


2 搭建SVM模型

2.1 将数据集转化为输入模型的格式，然后划分训练集与测试集：



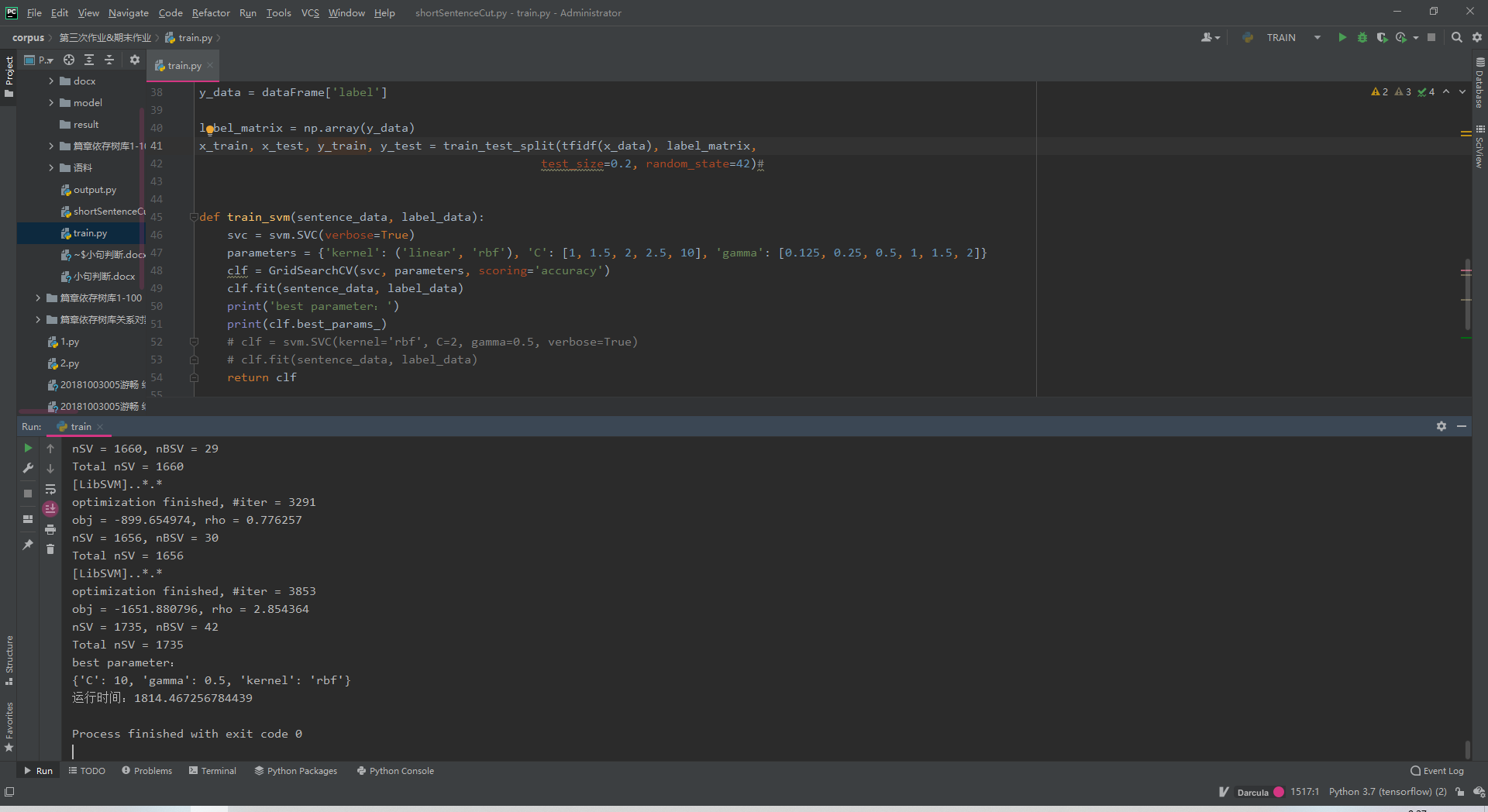
2.2 首先对词进行tfidf特征提取，并转化为矩阵：



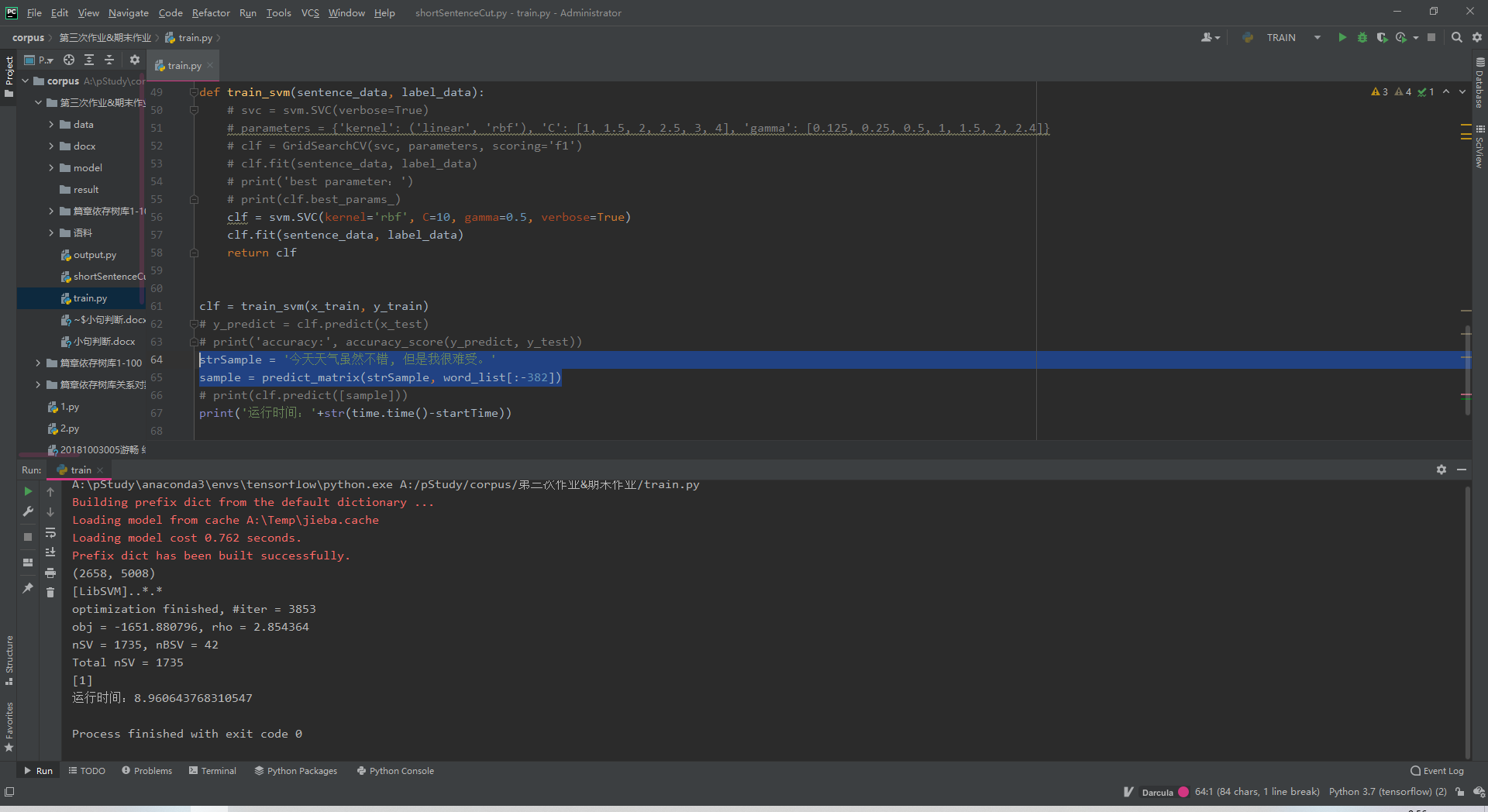
2.3 使用GridSearchCV寻找合适的svm超参数：

预设参数字典：parameters = {'kernel': ('linear', 'rbf'), 'C': [1, 1.5, 2, 2.5, 10], 'gamma': [0.125, 0.25, 0.5, 1, 1.5, 2]}

输出最优参数：

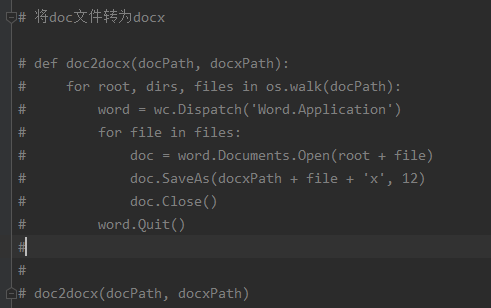


2.4 测试分类情况：

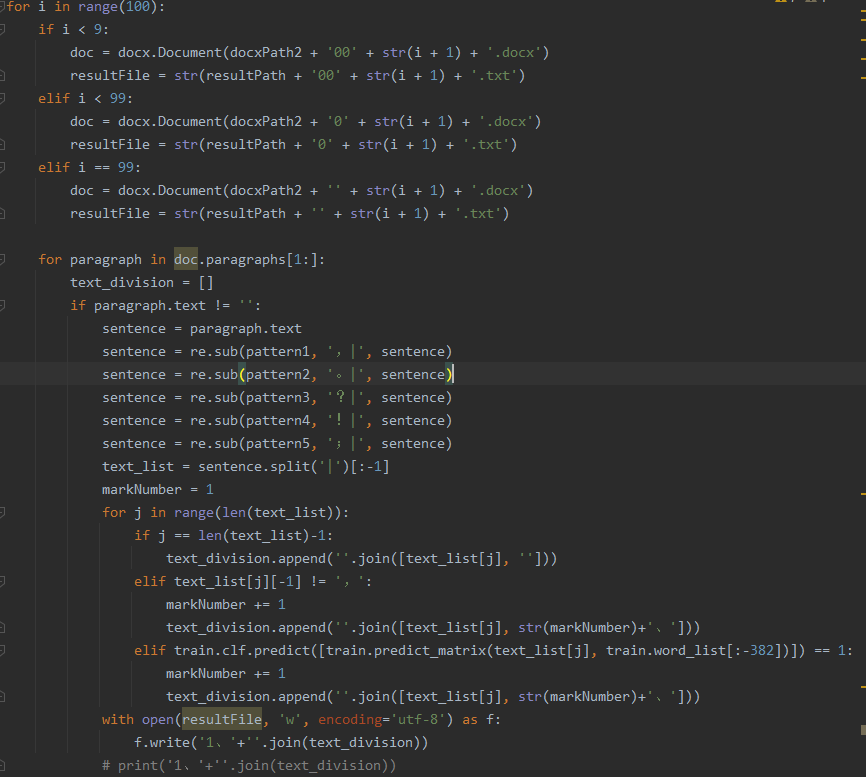


3 具体实现

3.1 将doc转化为docx格式：



3.2 循环读入文档并根据预测划分：



3.3 print查看输出结果：（此处缺了少句首的“1、”，写入txt时补充即可。）

