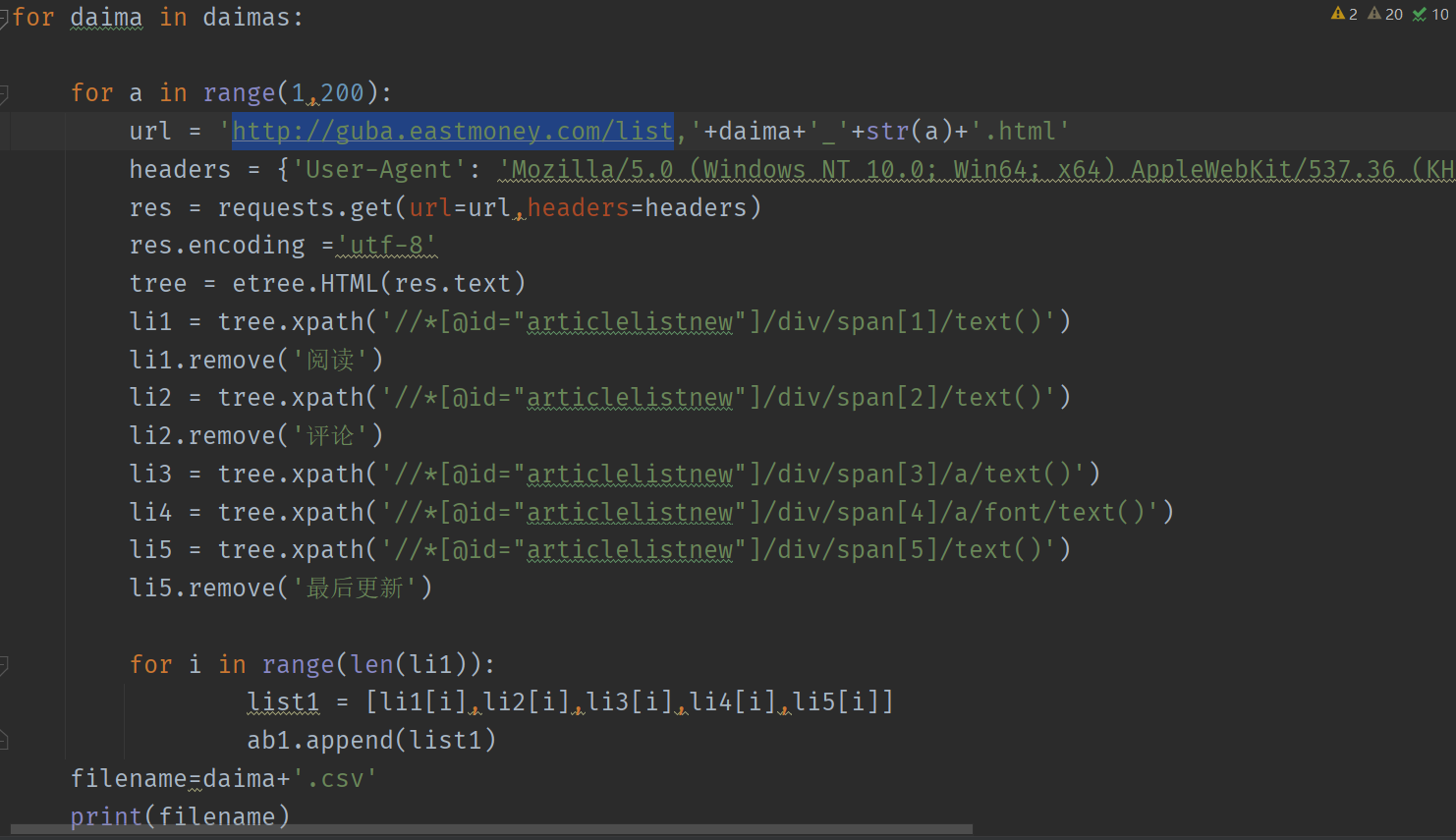
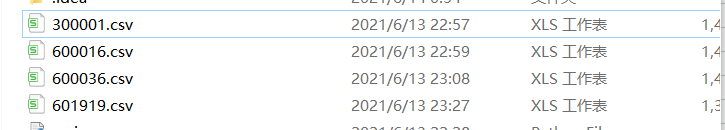
**第一部分：爬虫代码**

**爬取网页[http://guba.eastmoney.com](http://guba.eastmoney.com/)上特定4支股票的舆情，并保存本地，格式为csv**

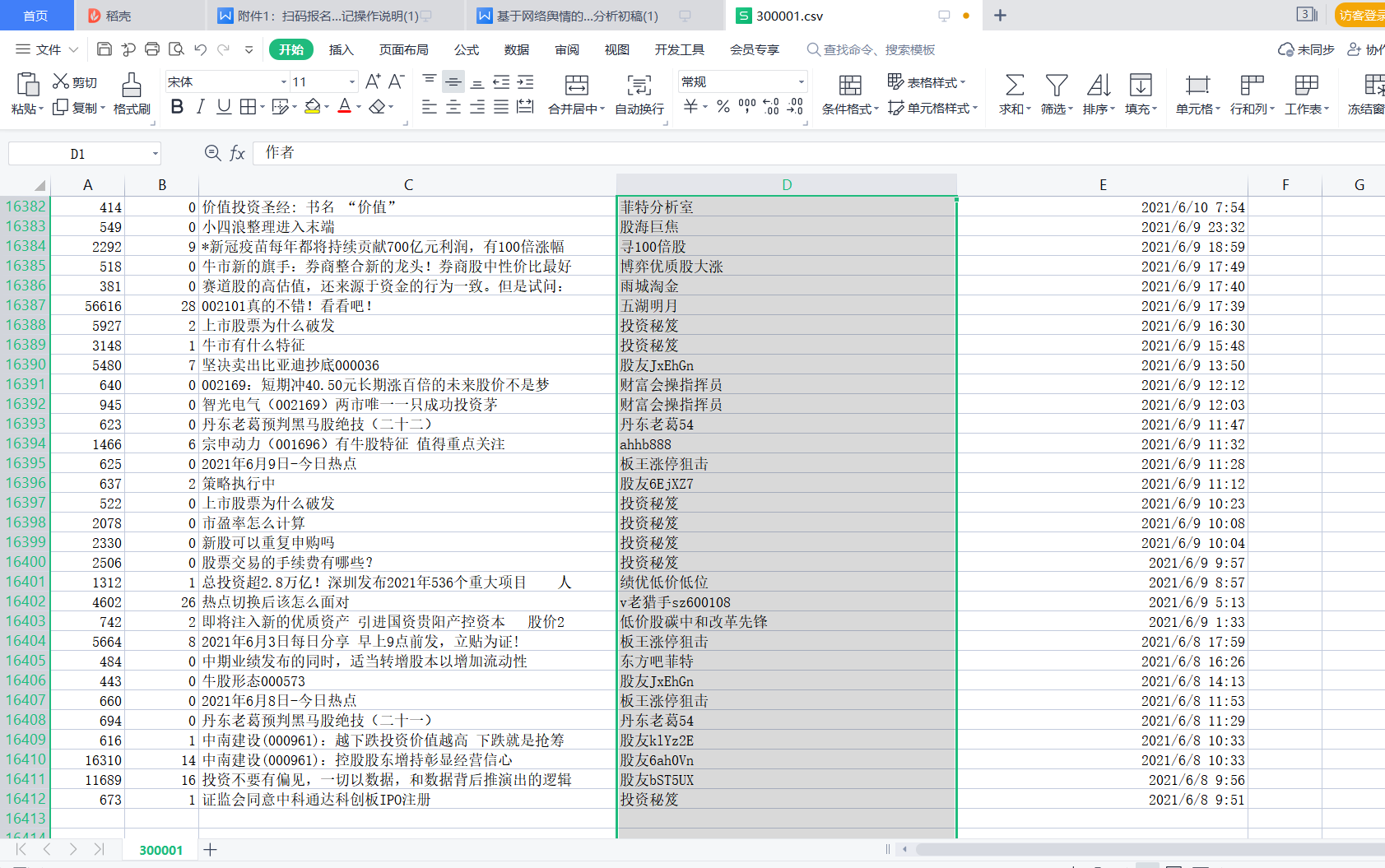
**爬取前200页的数据**







**每一只股票评论数有1W6**



**分析部分：**

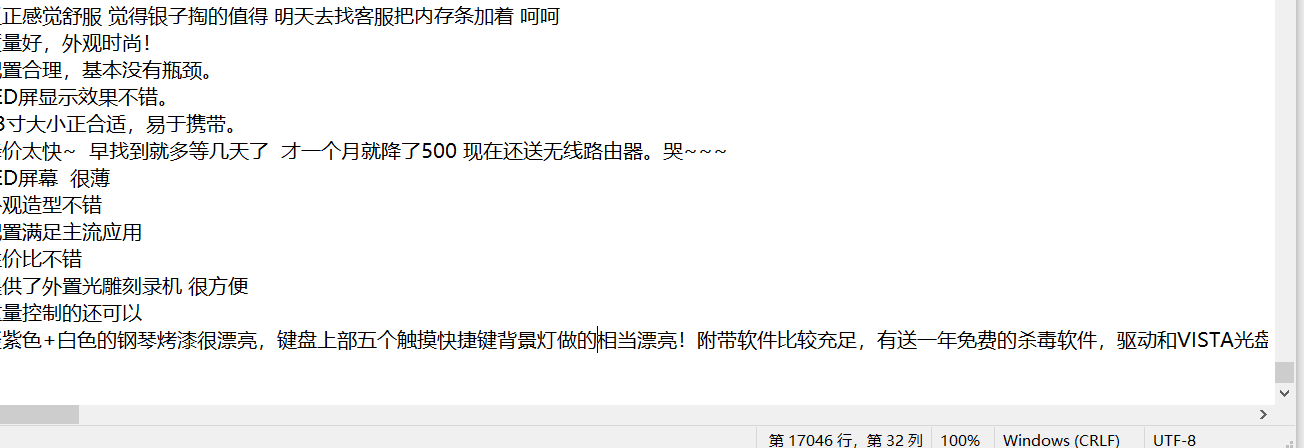
**建立新的语料库：**

**另外爬取几支股票，对其中的部分积极情绪和消极情绪进行人工划分**



分别保存到两个文本文件中，再从网上下载一些评论语料库





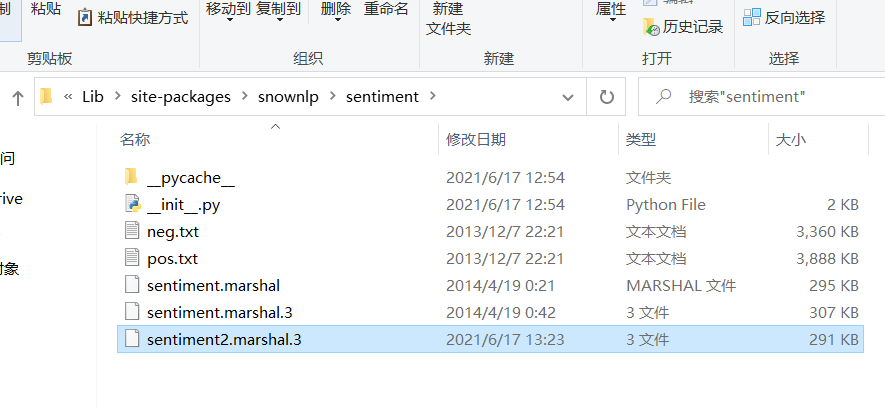
做出新的语料库，语料库中积极与消极的词汇分别都有1W7条左右

根据语料库来训练snownlp模型，再导出模型

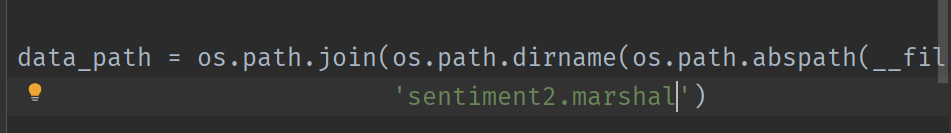




复制到snownlp文件的配置文件中



修改snownlp配置文件，使用训练好的模型进行训练



**读入刚刚保存至本地的csv文件，保存为dataframe，用pandas进行分析**

**对包含“？”的标题进行提取，发现大多数为新闻或者提问，是不包含情绪的无用信息**

****

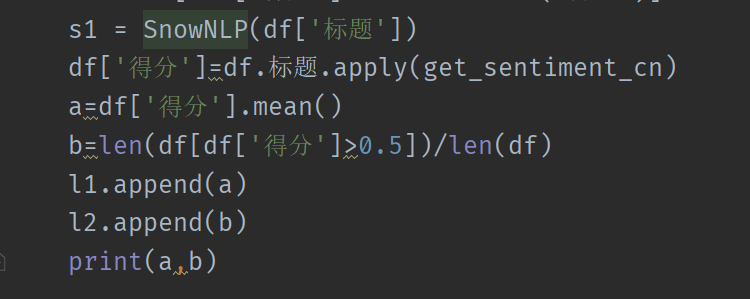
**对数据进行预处理，除去含有疑问词的无用信息**



**处理后，过滤了三千条无用信息**



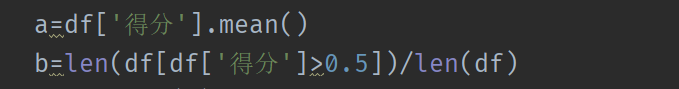
**使用snownlp包对标题进行文档分析，并将得分增加至新的一列**



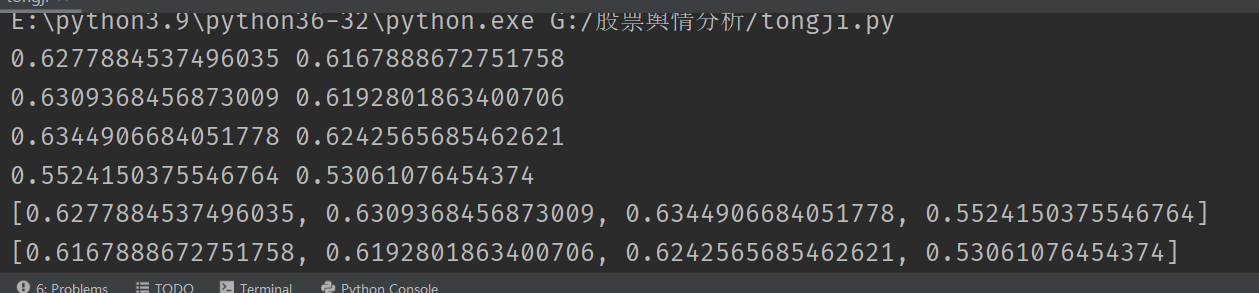


**评分>0.5即为 积极情绪，评分<0.5即为消极情绪**

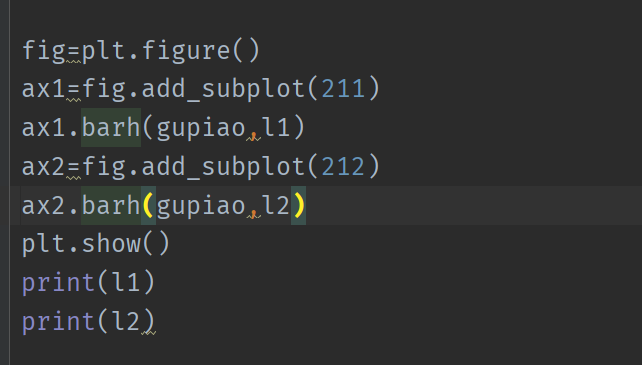
**计算每个股票的平均得分，以及积极情绪占总的比例，来得出实验结果**

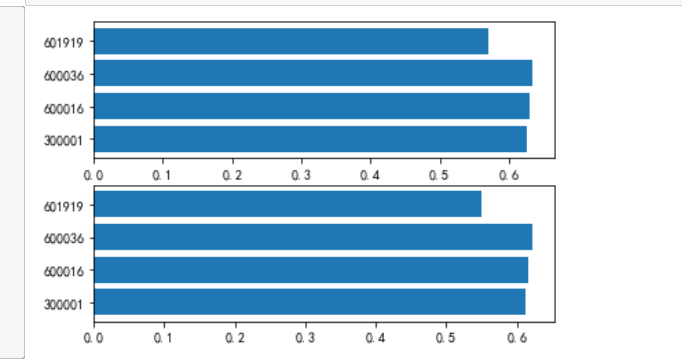


得到的平均分，积极情绪占比数据如下：



**再进行可视化：**





**第一个图为四个股票的平均得分，第二个图为四个股票的积极情绪占比。**

**结论：这四只股票中 600036 前景最好**