**点宽杯2022\_250\_基于东方财富股吧文本情感数据的单因子选股策略**

**程序文件说明文档**

**执行顺序：**

**量化A题第1小题→量化A题第2小题→量化A题第3小题→量化A题第4小题**

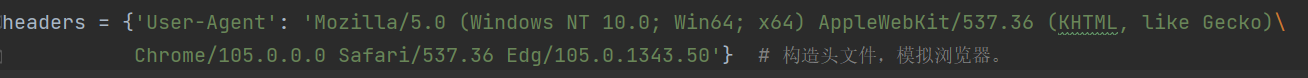
1. 运行文件夹 量化A题第1小题中的“爬虫-中国化学.py”文件，输出“中国化学.xlsx”，即股票中国化学在2021年1月1日至2022年6月30日的东方财富股吧文本数据。

想要获取其他股票的文本数据，可以通过修改“爬虫-中国化学.py”中的五处即可实现：

* 1. 修改结束页码

：495为2021年1月1日中国化学文本在东方财富股吧中的页码，应当修改为想要获取的股票的对应页码（需手动查询）。

* 1. 修改headers

：headers伪装应结合使用这电脑本身进行修改，通过浏览器打开的任意网页的源代码可以获取。

* 1. 修改起始页码

：83为2022年6月30日中国化学文本在东方财富股吧中的页码，应当修改为想要获取的股票的对应页码（需手动查询）。

* 1. 修改股票代码

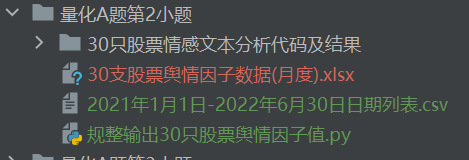
：601117为中国化学股票代码，应当修改为想要获取的股票代码

* 1. 修改输出文件名

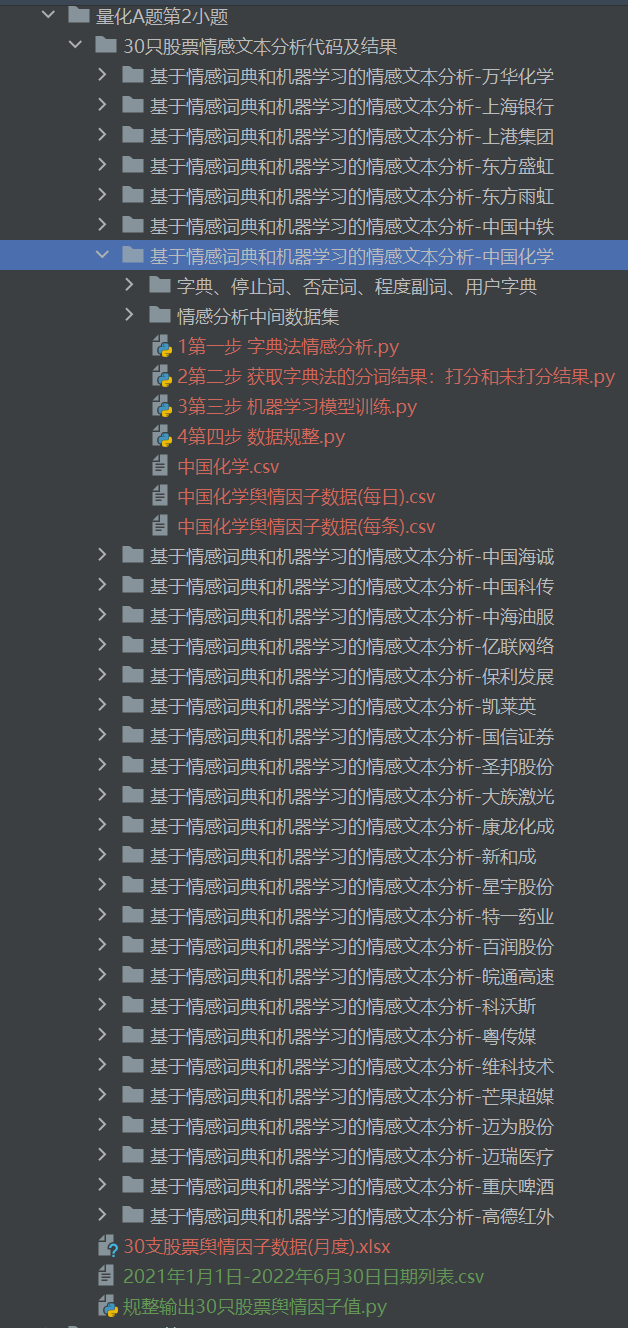
：将中国化学修改成想要获取股票的名称

以上五处修改完毕后便可直接运行获得想要的输出。

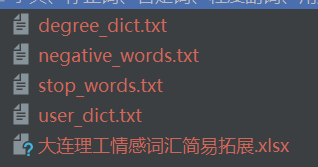
1. 打开文件夹 量化A题第2小题，可以看到以下文件：



其中文件夹 30只股票情感文本分析代码及结果 中包含了30只股票的文本数据及舆情分析结果。



以中国化学为例，程序的运行步骤如下：

（1）文件夹 字典、停止词、否定词、程度副词、用户字典 包含了四个txt文件和一个xlsx文件（每只股票文件夹中的字典、停止词、否定词、程度副词、用户字典文件夹内容相同），degree\_dic.txt和negatice\_words.txt文件为赛题方提供的程度副词和否定词文件，stop\_words.txt为否定词文件，通过网络自行搜查整理而成，user\_dict.txt为用户自定义字典，主要是赛题方提供的字典的中文单词，“大连理工情感词汇简易拓展.xlsx”中包含了赛题方提供的“大连理工情感词汇本体.xlsx”、“金融领域中文情绪词典\_姚加权等(2021).xlsx”及“中文金融情感词典\_姜富伟等(2020).xlsx”的汇总。

（2）文件夹 情感分析中间数据集 则包含程序运行过程中产生的所有数据集，“字典法情感分析初步结果.csv”为“1第一步 字典法情感分析.py”的运行结果，“bad\_result.csv”、“good\_result.csv”和“label\_dict\_result.csv”为“2第二步 获取字典法的分词结果：打分和未打分结果.py”和“3第三步 机器学习模型训练.py”的运行结果。

（3）在依次运行“1第一步 字典法情感分析.py”、“2第二步 获取字典法的分词结果：打分和未打分结果.py”、“3第三步 机器学习模型训练.py”并得到结果后，通过运行“4第四步 数据规整.py”可以得到数据集“中国化学舆情因子数据(每日).csv”和“中国化学舆情因子数据(每条).csv”，其中“中国化学舆情因子数据(每日).csv”文件包含了股票大连化学的日度舆情因子数据。

（4）文件夹中的“中国化学.csv”即为文件夹 量化A题第1小题的运行结果（文本数据）

（5）在运行完30支股票的代码文件后，通过运行“规整输出30只股票舆情因子值.py”文件可以输出“30支股票舆情因子数据(月度).xlsx”，即为30支股票的月度舆情因子数据。

1. 运行文件夹 量化A题第3小题 下的“舆情因子有效性验证.py”，可以得出因子有效性的验证结果
2. 运行文件夹 量化A题第4小题 下的“sentiment\_stock\_portfolio\_strategy.py”，可以在AT平台中得到最终的策略回测结果。