一、3种爬取数据的方式

|  |  |
| --- | --- |
|  |  |
| Requests+beautifulsoup | 对于反爬机制一般的网站适用（网址参数较为固定、非动态生成）如果存在动态参数，如雪球网的md5\_1038参数，其参数值需要在线运行Javascript文件得出，破解办法只有实施JS逆向工程，找到参数值的计算逻辑并本地化部署（难度很高） |
| Selenium | 自动化操作浏览器，可绕过动态参数限制。但可能会被反爬机制严格的网站直接识别，如雪球网对selenium库生成的浏览器对象直接触发反爬验证（这个我没解决掉）；反爬机制一般的网站可用 |
| pyautogui | 全局自动化操作，可绕过动态参数限制。缺点是对屏幕分辨率有一定要求。建议使用时直接查找网站评论数据API网址，如雪球网为：<https://xueqiu.com/query/v1/symbol/search/status?symbol=SZ002570&page=45&count=100>，可高效复制粘贴JSON格式化数据到本地 |

二、数据的格式

建议数据整理为包含以下属性的CSV文件：

['股票代码', '评论id', '评论用户id', '评论日期', '评论内容', '回复量', '转发量', '点赞量', '收藏量', '浏览量', ‘来源’]

注意：帖子本身的API网址很可能和评论区的API网址不同，如雪球的评论区API网址为<https://xueqiu.com/statuses/v3/comments.json?id=310600040>。

三、进度汇总日期安排

暂定4月1日（已参考部分组员需求）。