Python网页爬虫

1. 网站：https://www.taoguba.com.cn

注意，首先要登录，如果没有要先注册再登录。

登录后，在右上角搜索框输入股票名称，比如“新开源”，在自动下拉列表选择【包含“新开源”的标题】下面的“新开源”，如下图的灰色栏。



1. 点击后进入跳转网页，显示为：



想要的就是【讨论】下面发帖内容，筛选范围选择默认参数**半年内**即可，可能有好几页内容，仅需当前第一页的内容。

1. 需要提取的内容显示示例为：

|  |  |  |  |
| --- | --- | --- | --- |
| 新开源 | 2019-11-28 00:29 | 探讨扫板新开源的核心逻辑与背日K线和分时图图形的必要性 | [摘要] 新开源前面有一个二连板，然后调整两天，第三天也就是今天反包涨停！我扫板的逻辑是这是一只人造肉和基因测序概念的股票。今天早上大幅低开盘后快速拉到红盘，后面一路涨停，从分时图上看也适合扫板。跟当时的新通联大跌后反包第一个反包板很相似。而且今天新开源的成交量突破了前面几天的…… |

1. 把当前页显示内容都按照上述格式整理成一个表格
2. 按照给定的一个股票名称数组，循环遍历搜索下一个股票名称，比如电广传媒，把搜索内容汇总成表格
3. 汇总以上所有表格，导出到csv文件中。

注意几个问题：

提取网页内容时需要验证登录身份，因此需要实现如何通过登录验证。