# 基于pyquery的GitHub爬虫及基于selenium的自动关注GitHub项目实现

许靖桐

1. 项目需求
2. 获得Github trending上的十个最热门项目，输出项目stars数，项目fork数，以及主要贡献者id到一个csv文件中。其中检索要求为Spoken Language=English, Language=python, Date Range=This month。
3. 利用Web Drivers/Restful API/GraphQL API，使用测试账号关注（watch）爬取到的十个项目，并选择“Releases only”选项。
4. 项目实现
5. 爬虫部分

spider.py中包含了网络爬虫函数spider()，该函数通过使用pyquery对网页进行解析，获得十个最热门项目的star数，fork数，主要贡献者并将这些数据输出到一个csv文件中。该函数返回一个包含十个最热门项目URL的列表，以便下一步使用webdriver关注这些项目。

首先对于GitHub网站热门项目页面进行观察，发现网站可以根据项目语言，编写语言，以及日期范围进行检索，在进行检索时，网页URL会根据检索项的变化而改变，因此根据要求的选项进行检索，获得URL，通过该URL可以直接检索到要求的热门项目。

通过对GitHub网页进行解析，可以发现每一个热门项目的内容都以一个<article>标签的形式出现，因此首先获得所有带有<article>标签的元素，然后对这些元素进行处理，分别根据元素class的名称获得项目名称的字段，URL，星星数，fork数的字段，以及主要贡献者的id。然后将这些数据保存在一个列表中，并创建csv文件，在文件名前加上保存日期，然后保存。

1. 自动关注部分

自动关注部分利用Selenium中的webdriver编写。该部分的代码在webdriver.py中。该部分包含三个函数，分别负责登录GitHub，关注项目以及取消关注项目。在使用Selenium前，我对我使用的Chrome浏览器下载了相应的驱动并进行了配置。

首先，使用webdriver可以自动打开浏览器，并通过URL访问指定的网页，但这样访问GitHub时需要先登录，因此通过github\_login()函数，传入GitHub登录页URL，然后输入测试账号的账号及密码，并控制webdriver找到登录按钮点击登录，这样就可以登录GitHub并完成后续操作。

然后，watch\_project()函数可以通过传入的项目URL访问项目主页，并查找到主页上的“Watch”按钮，选择“Releases only”选项并点击，以实现关注项目。同理，unwatch\_project()函数可以实现取关项目的功能。

在webdriver.py的主函数中，首先通过webdriver打开Chrome浏览器，并爬取需求的数据，然后登录GitHub并关注这些项目。最后关闭浏览器。

1. 项目总结

在项目中，我使用了pyquery进行爬虫的编写，使用Selenium进行自动化测试，由于之前并没有使用这两种工具的经验，因此在项目编写中也遇到了一些困难。比如，在利用pyquery访问包含星星数结点时，由于星星数节点与fork数节点的标签除了超链接都相同，因此只能搜索所有类名为muted-link d-inline-block mr-3的节点并用text()输出节点的内容，这样得到的内容是一个含有星星数和fork数的字符串，然后对这个字符串进行处理分别获得两项数据。同样在处理主要贡献者时，也遇到了无法准确定位节点的问题，只能通过逐步定位，然后利用查找的方式获得所有主要贡献者数据。在webdriver方面，我直接使用了输入账号密码的方式进行登录页面，这样虽然容易但是丧失了安全性，当然也可以使用控制台输入账号与密码，但这样的话又需要在控制台和打开的浏览器之间相互切换，比较麻烦。最理想的状态应该可以手动输入账号密码，然后通过Selenium判断是否登录，然后再进行下一步，但我并未实现这些功能。

我并没有为代码编写单元测试，但在测试方面，我能想到的测试包括对于网络状况的测试，对于爬虫爬取数据准确性的测试，对于写入数据准确性的测试，对于Selenium登录及关注是否成功的测试等。其中最重要的就是对于爬取数据准确性的测试。但是我对于如何测试这些数据有一些疑问，如果编写自动化测试，那么是否也需要利用爬虫获取相应的数据，这样的话，测试代码的准确性是否也需要再进行测试？这样就变成了无限循环下去的问题，目前我还没有想到如何对这些数据的准确性进行验证的方法。

在这个项目中，我也学到了很多之前不了解的知识，比如pyquery与Selenium的原理，其中Selenium的使用非常方便，令我对其产生了极大的兴趣。总的来说，我收获了许多。