百度贴吧爬虫demo

1. 运行环境

* 开发语言：python3
* 系统：Windows 10 x64

1. 安装依赖库

* pip install -r requirements.txt

1. 程序设置

打开当前目录下的BDtieba\_config.json文件，看见如下内容：

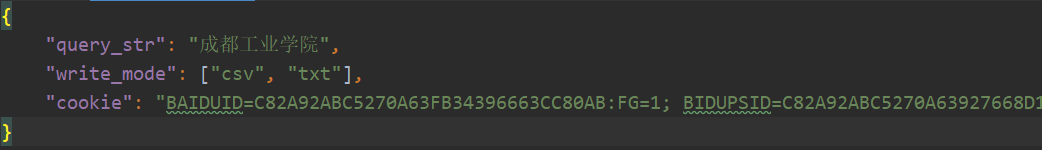


图1 设置部分截图

其中除write\_mode不提供设置，demo仅提供3项设置，其余和正式版程序相差甚微。

1. query\_str：它是代表着我们要爬取的 贴吧的名字 ，正式版中是list形式，本demo版只提供一次一个贴吧的爬取形式，当然贴吧有多少页就能爬取多少页，这个不受限。

例如：我要爬取 新型冠状病毒吧 的所有信息，仅需将“成都工学院”改掉，如下：

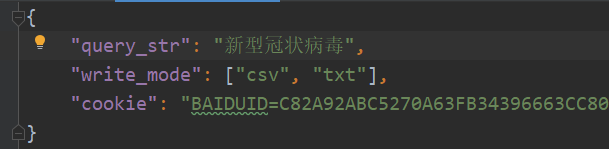


图2 获取新型冠状病毒吧信息

1. cookie：这个相当于是你的登录信息，提供的cookie可能会过期，届时需更换为你自己的cookie，具体获取方法 在 附录中的 “获取cookie” 部分讲解。
2. 按需求修改爬虫程序
3. 因为程序的主题是生成一个BDPostBar类，程序是main()调用该类实现的。
4. 为了防止百度贴吧反爬措施，在获取网页信息的函数get\_page\_info()中加入了随机睡眠“t=random.randint(1, 10);time.sleep(t)”在处理链接的函数handle\_url()中加入了随机请求头“headers=select\_header()“，可以根据需求更改
5. 简单的修改方法是：在BDtieba\_spyder.py中寻找位于代码末的main()函数中进行修改，demo能修改的地方不多。
6. 运行.py脚本

推荐使用pycharm2019.3运行，以便调试；若是使用cmd、power shell运行程序，请悉知：

* 默认是只爬取贴吧前5页内容，并保存在当前目录下，格式为.csv
* 请确认cookie能够正常使用，且配置文件已修改完毕

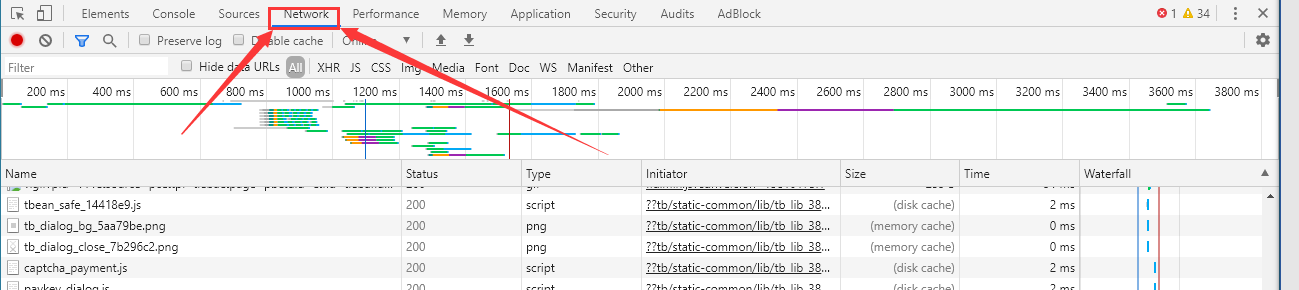
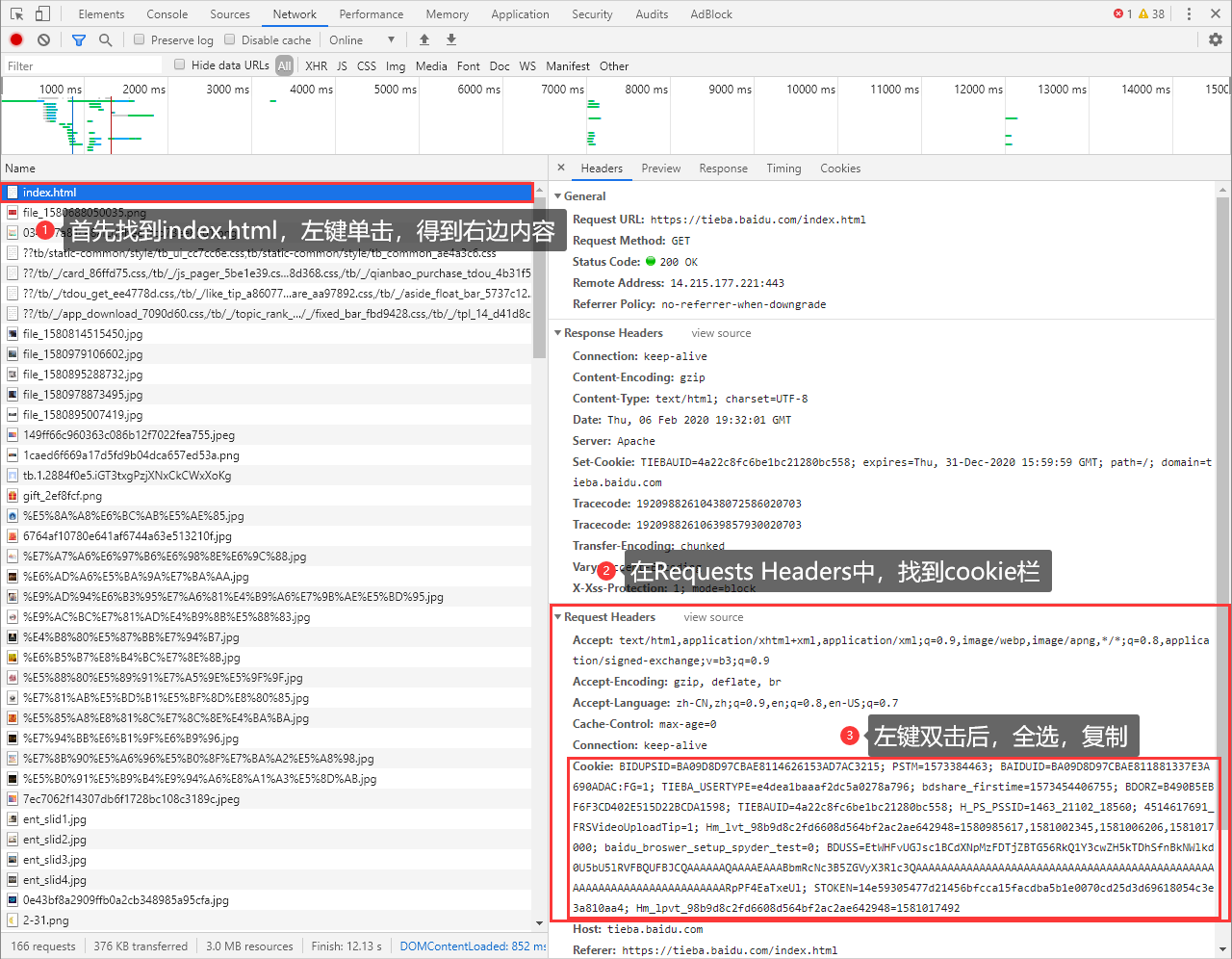
程序运行完成后，得到的信息是：

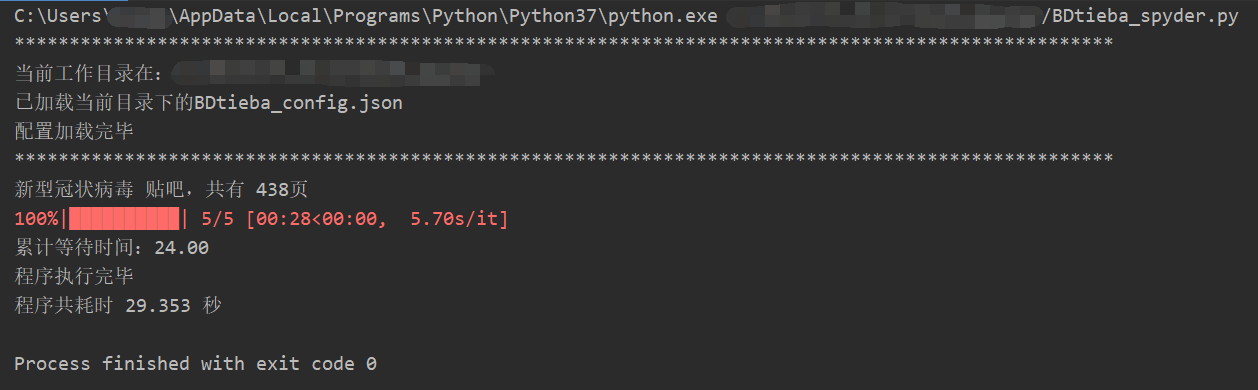
BDPostBar.user\_info：使用者的部分信息，dict类

BDPostBar.all\_pages：贴吧总共的页数，list类

BDPostBar.result：爬取到的所有信息，DataFrame类

附录

1. 如何获取cookie
2. 用Chrome或Microsoft Edge Beta打开网址<https://tieba.baidu.com/index.html>，并按下F12，选到Network栏
3. 点击上方登录按钮，输入账号密码并成功登录后，按下F5进行刷新
4. 寻找到以下图片所示位置，使用复制的内容替换掉BDtieba\_config.json中cookie的值，爬虫程序运行时就是以你的账号进行访问网页。
5. 注意事项
6. cookie必须自己设置，demo中的cookie也是账号登陆后的，所以为网络安全着想，请更换为自己的cookie。
7. cookie有时间限制，超过时间必须进行cookie的更新。
8. 在处理网页链接时可能出现偶尔的无法连接，请重试3次左右，若每次都有报错，请检查函数handle\_url()，必要时可将“headers=select\_header(),”这一行注释掉，再次尝试。
9. 爬虫正常工作预览：



1. 看懂程序了，就请忽略本篇Readme了，想咋改就咋改。