**第二次实验报告**

1. 实验概览

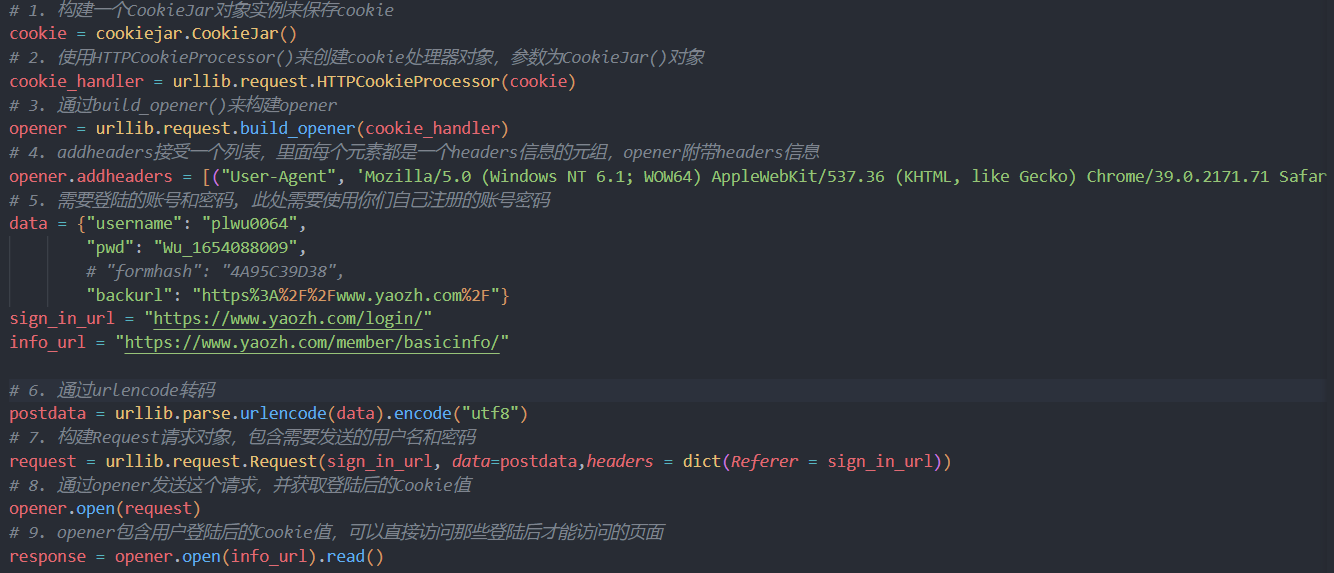
本实验旨在于理解HTTP协议的具体内容，包括HTTP请求的过程，提交的HTML表单包含内容，并利用python代码模拟实现发送HTTP请求。实验的另一部分旨在于理解爬虫的概念，学习使用爬虫技术应遵守的原则，实现爬虫抓取网页的两种策略。

1. 实验环境

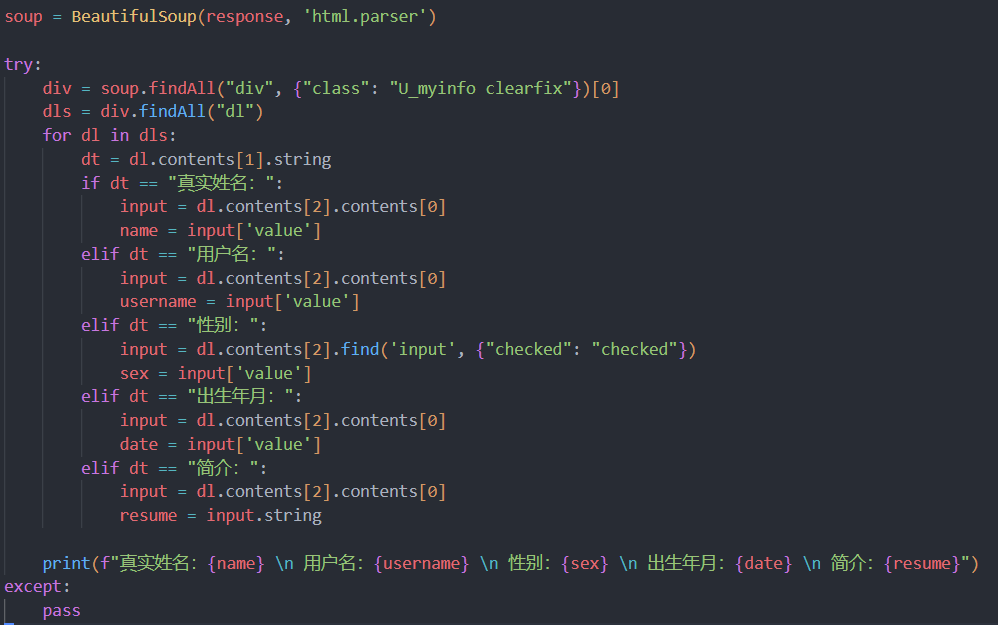
Docker: SJTU-EE208

1. 解决思路
2. 练习一

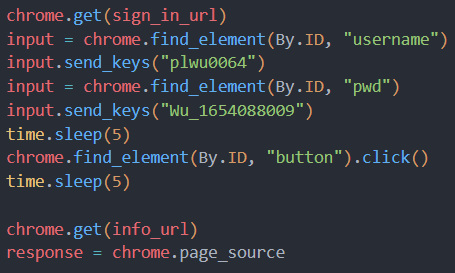
基于urllib中的request类，利用add\_header方法添加header模拟浏览器对网络进行访问；利用cookiejar库构建登录所需的cookie信息加入request中实现登录并获得网页。



基于bs4对html网页源码进行解析，通过开发者模式查看网页发现所需信息格式均在<div class=’U\_myinfo clearfix’>中，以<dl>标签分段排列，利用BeautifulSoup类的findAll方法查找对应标签得到信息。



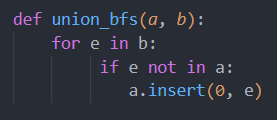
在具体实施过程中，发现登录次数一定后网页会有验证码需求，无法通过urllib直接登录，故考虑使用selenium库模拟网页图形界面解决验证码问题。



之后的解析过程一致

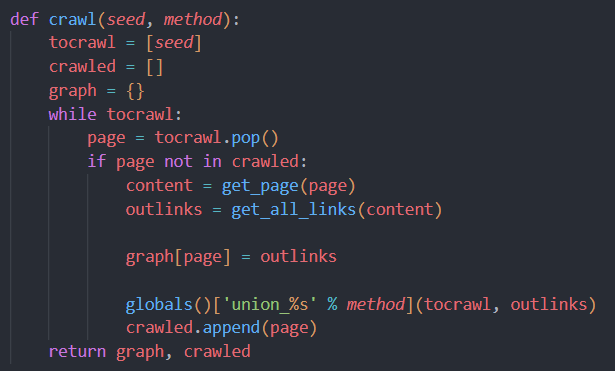
1. 练习二

利用list类的insert操作实现BFS



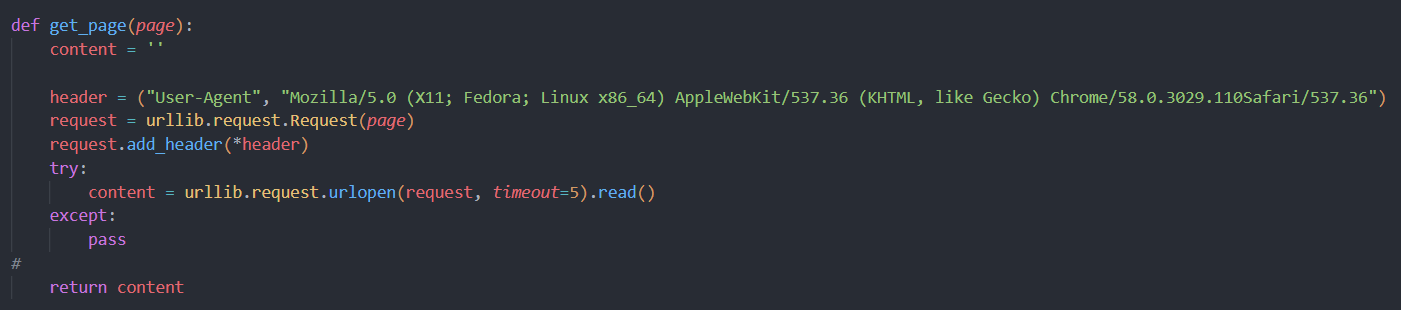
1. 练习三

分析代码，确定outlinks变量包含了每个节点下子节点，赋值实现输出graph

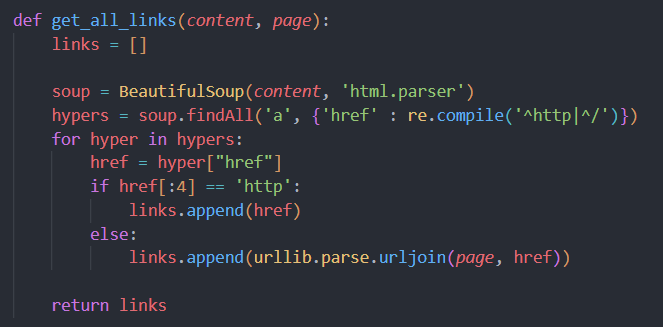


1. 练习四

基于urllib库获取网页信息并进行异常处理

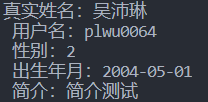


基于BeautifulSoup类和正则表达式查找超链接子网页，根据是否以http开头判断相对网页绝对网页并得到最终网页路径。

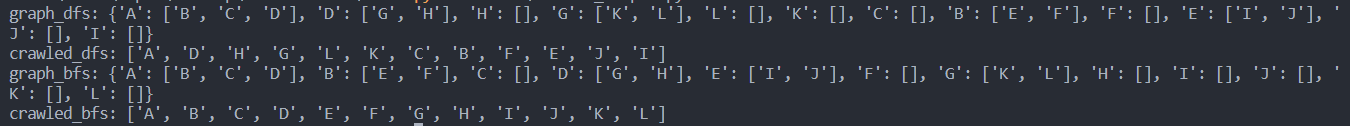


1. 代码运行结果

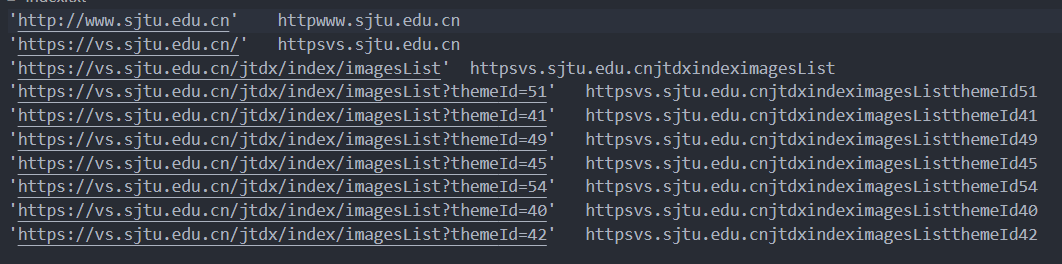
练习一：执行python example1.py，如需输入验证码在图形界面中输入，在终端中显示结果



练习二、三：运行python crawler\_sample.py在终端中得到结果



练习四：运行python crawler.py直接得到对应index.txt和文件夹，附图为index.txt，具体结果请参考对应文件



1. 分析与思考：网络爬虫的合法性？

网络爬虫需遵循礼貌性原则，爬虫对于浏览器访问频率过高，导致服务器崩溃，可能要承担对应法律责任；同时针对爬取信息应遵守robot.txt文件，爬取公民敏感信息则可能承担对应民事乃至刑事责任。