

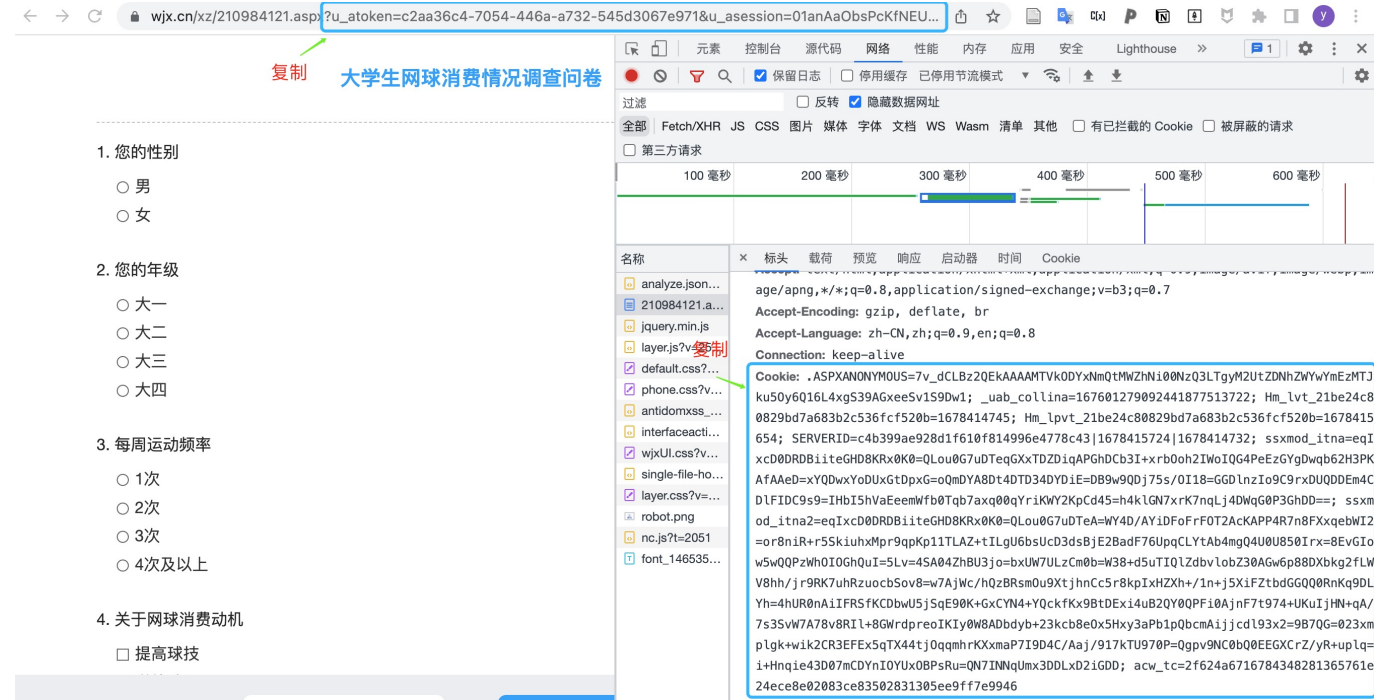
问卷数据爬取

1. 环境配置

```
pip install requests
pip install BeautifulSoup4
```

2. 设置cookie

访问<https://www.wjx.cn/xz/210984121.aspx>, 右击->检查->网络，获取cookie




替换代码中：

```
def get_response(index):
    url = "https://www.wjx.cn/xz/{index}.aspx".format(index) + "?u_atoken=64f1b01f-dce4-450a-b54b-c170896efdbf&u_asession=01c3..."
    headers = {
        "Content-Type": "text/html; charset=utf-8",
        "Referer": "https://www.wjx.cn/newwjx/mysojump/newselecttemplate.aspx?",
        "user-agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100...",
        "Cookie": "SERVERID=c4b399ae928d1f610f814996e4778c43|1676994441|1676993750;Path=/"
    }
    response = requests.get(url, headers=headers)
    response.encoding='utf-8'
    return response
```

3. 创建文件夹

```
mkdir crawl
手动创建一个新文件，如wenjuanxing_final_0.json
```

4. 后台运行代码 设置开始index, 如10000-20000，设置为自己的开始index，后续如果服务器断开了，重新开始时，选择新的开始位置也是更改该位置。

```
formatted_questionaries(209451158)  
index 
```

```
nohup python crawl.py >> output.log 2>&1 &
```

备注：如果网络断开，可以更改开始index位置，执行step4