Akkuman
别聊了，造车去

javbus爬虫-老司机你值得拥有

阅读目录

回到顶部

起因

有个朋友叫我帮忙写个爬虫，爬取javbus5上面所有的详情页链接，也就是所有的https://www.javbus5.com/SRS-055这种链接，
我一看，嘿呀，这是司机的活儿啊，我绝对不能辱没我老司机的名声（被败坏了可不好），于是开始着手写了

回到顶部

构思

- 爬虫调度启动程序crawler.py
- 页面下载程序downloader.py
- 页面解析程序pageparser.py
- 数据库入库与去重管理程序controler.py

爬取入口为第一页，当页面中存在下一页的超链接继续往下爬，这是个死循环，跳出条件为没有了下一页的链接

在某一页中解析页面，返回所有的详情页链接，利用迭代器返回，然后在主程序中调用解析程序对页面信息进行解析并包装成字典返回，其中用详情页网址作为数据库主键，其他信息依次写入数据库

当这一页所有的子链接爬取完成后，继续爬取下一页。

将数据存入数据库，用的是sqllite3,失败的网址页存入一个fail_url.txt。

对于增量爬取，我是这么做的，当爬取到相同的网址时结束程序，这么做也有漏洞，才疏学浅，我没想到太好的办法，希望有好办法的给我说一声（布隆过滤正在研究之中），如果用数据库查询去重，那么势必导致二次爬取，我们都知道，爬虫更多的时间是花在网络等待上

回到顶部

问题

在写爬虫的过程中遇到了一些问题

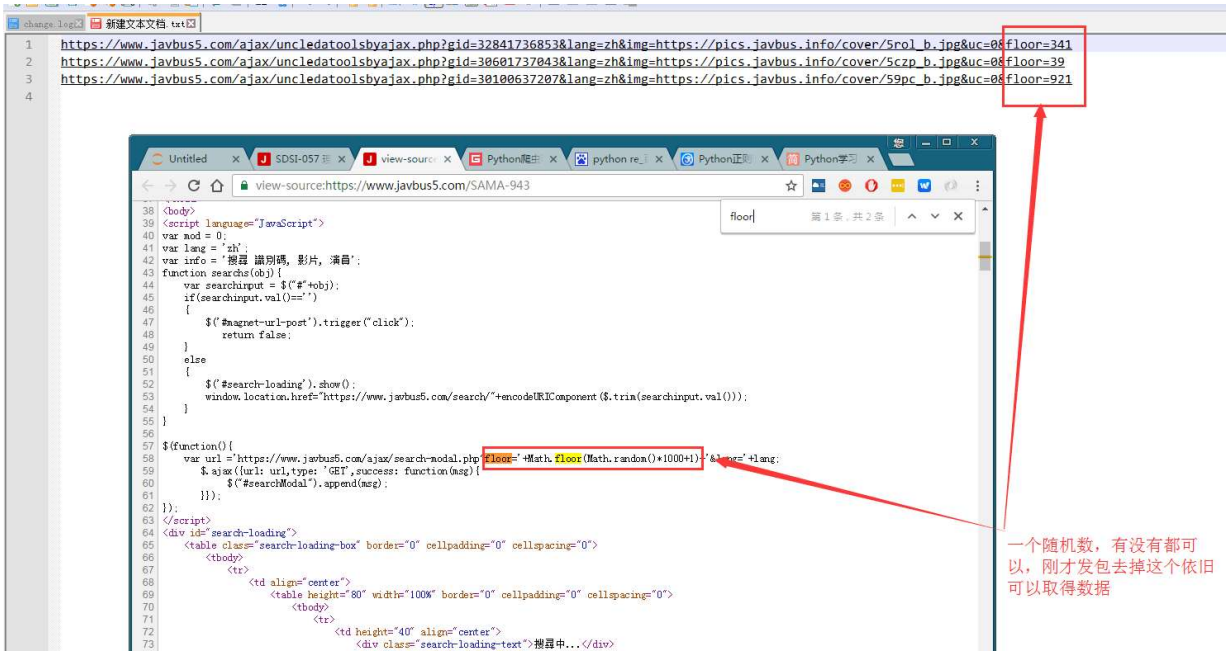
1. 在墙内爬不动，爬取几个之后就失败，这个解决方案只需要全局FQ爬取就可以了
2. 本来之前加了多线程并发爬取，但是发现爬取一段时间后会封ip导致整体无法运行，本来想搞个代理池进行并发，结果网上免费的代理太慢太慢，根本打不开网页，于是就改回了单线程
3. 就是我的那个不完善的增量爬取，导致了你一次爬取就需要爬取完成，不然数据库里面存在你之前爬到的，爬取到你已有的会直接停止
4. 存在反扒策略
   详情页中的磁力链接是ajax动态加载的，通过分析抓包，可以在XHR中找到是一个get请求，至于参数，我开始不知道怎么得来的，后来在html代码中找到了，我放几张图大家就明白了

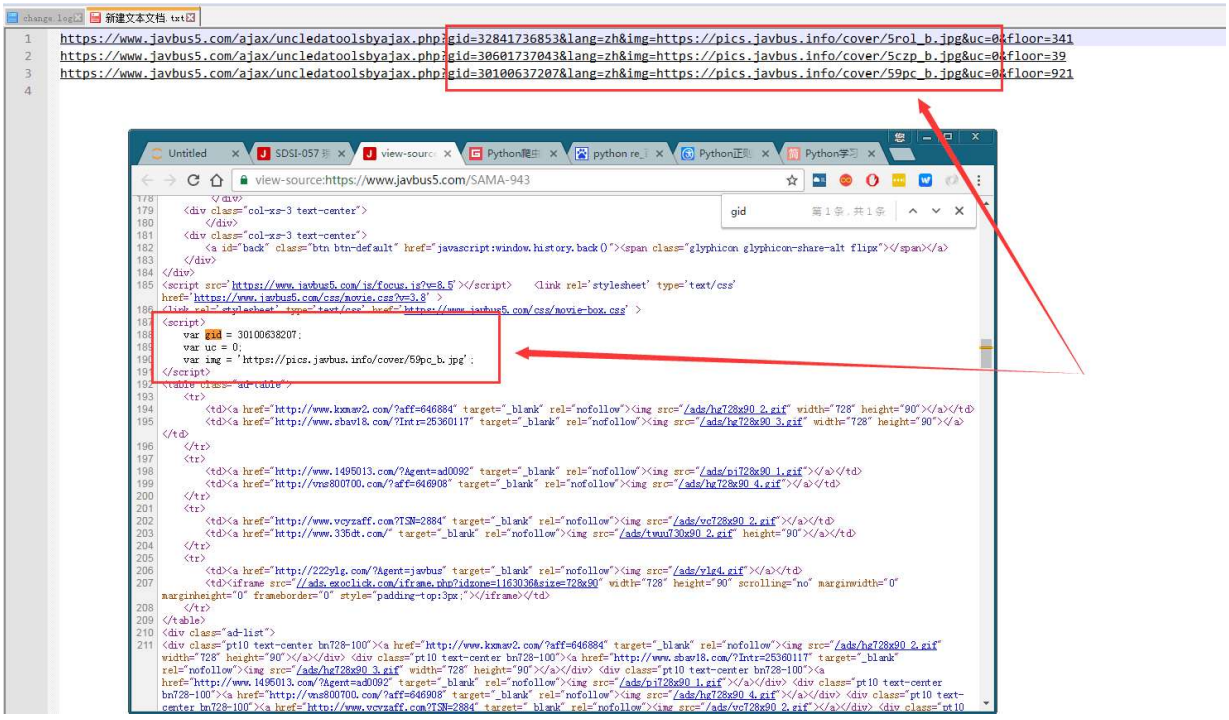
我们通过对响应内容的查看可以发现磁力的加载访问了类似于这样一个网址

```
https://www.javbus5.com/ajax/uncledatoolsbyajax.php?
gid=30100637207&lang=zh&img=https://pics.javbus.info/cover/59pc_b.jpg&uc=0&floor=921
```

那么这些get参数是从哪里来呢，这就是通过经验与基本功去发现了

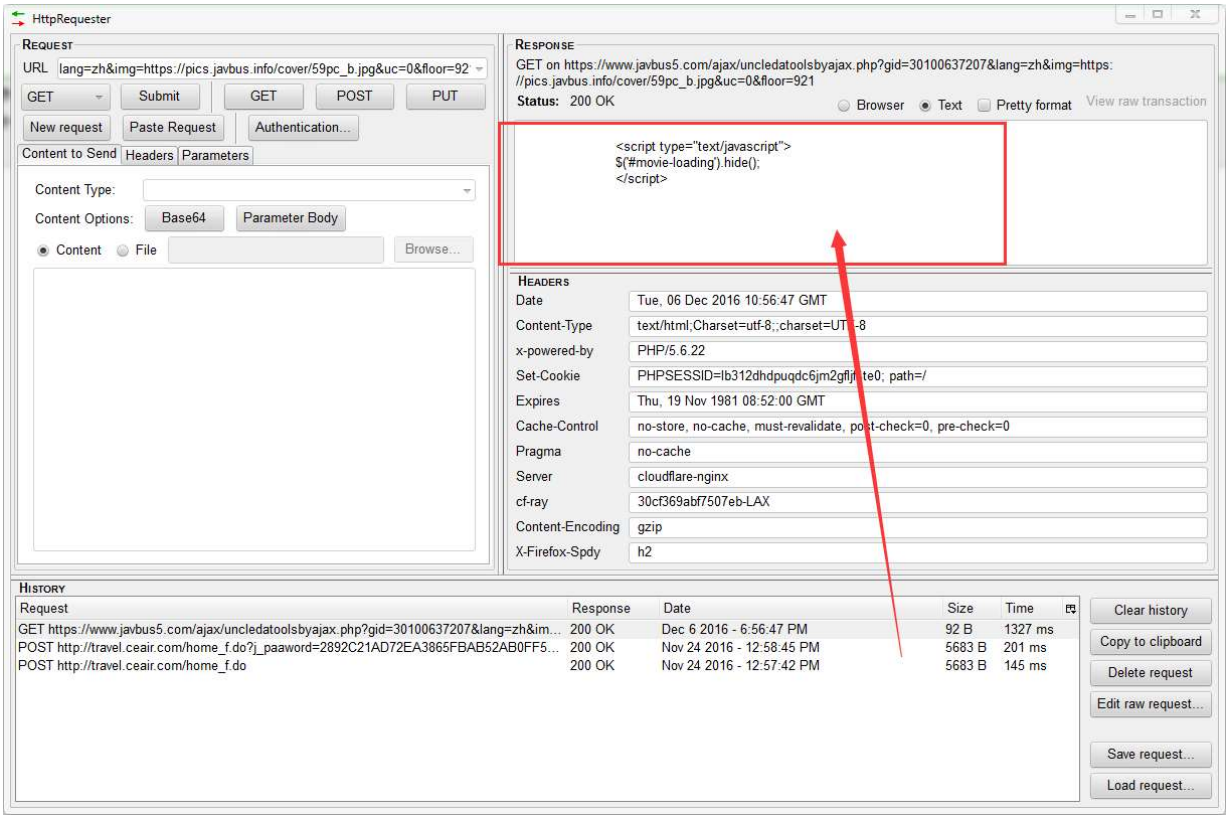
通过对html源文件的搜索，我们即可直接发现答案



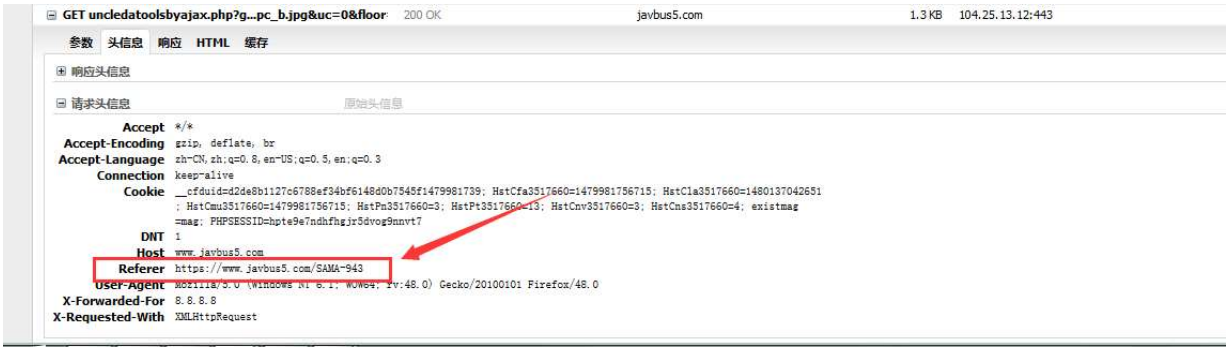


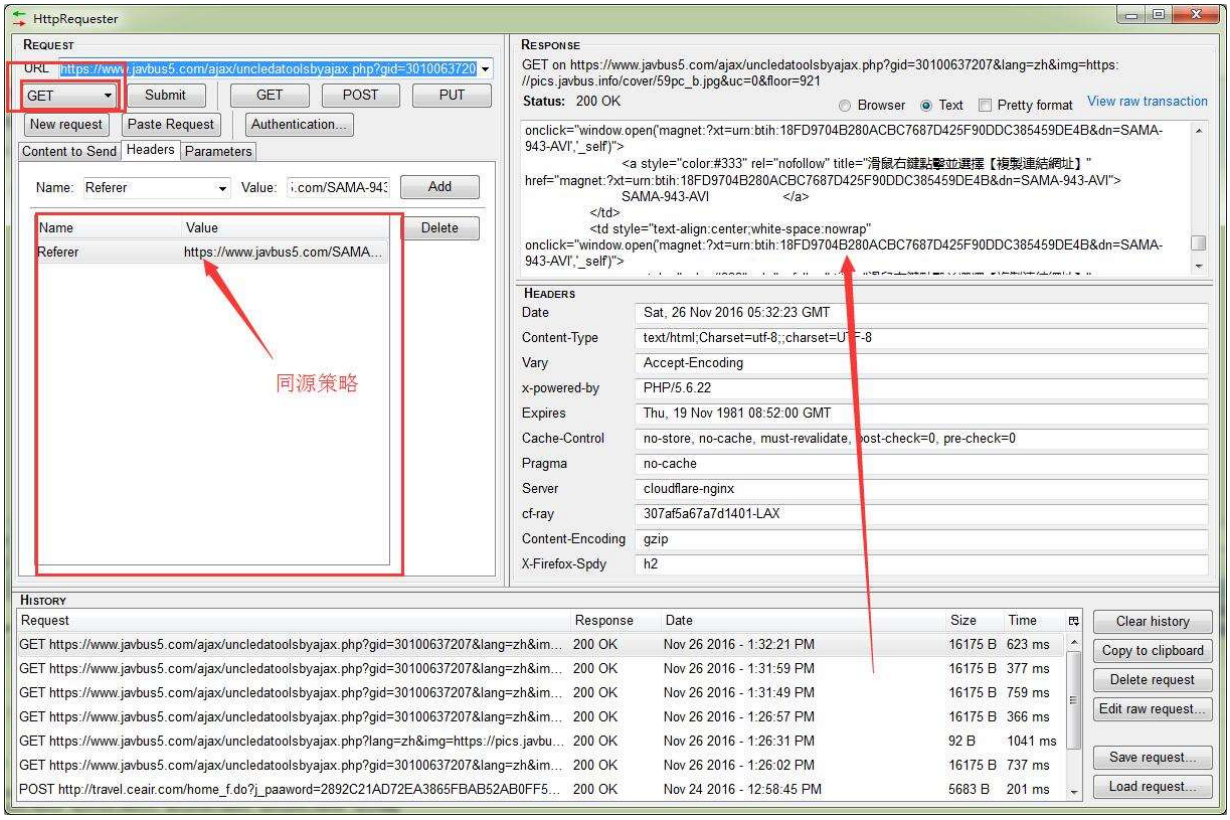
通过分析发现，后面的floor是个随机数参数，一般这种参数可以去除无影响，事实也是这样

我利用HttpRequest模拟发包，对这个请求直接get，发现所有数据隐藏



那么肯定是有反扒的策略，伪造请求头，反扒也就那么几种，通过分析发现是同源策略，对Referer请求头伪造成来源网址就可以直接获取到内容了





1. 常见的Python2.x编码问题,全部转换为unicode字节流就可以了
   这个问题在我博客中已经记录了http://www.53xiaoshuo.com/Python/77.html
   有兴趣的童鞋可以看看

2. 遇到的最闹心问题是详情页的项目抓取，有的详情页的类别不同，我开始只分析了一个页面，导致写的规则在有的页面上频频出错
   导致后面对抓取规则进行了大改,重写了分析规则，用了个笨办法，毕竟那小块的html写的十分不规范，正则规则有三种，挺烦人

識別碼: HEYZO-1329

發行日期: 2016-11-17

長度: 60分鐘

製作商: HEYZO

類別:

舔陰　家中

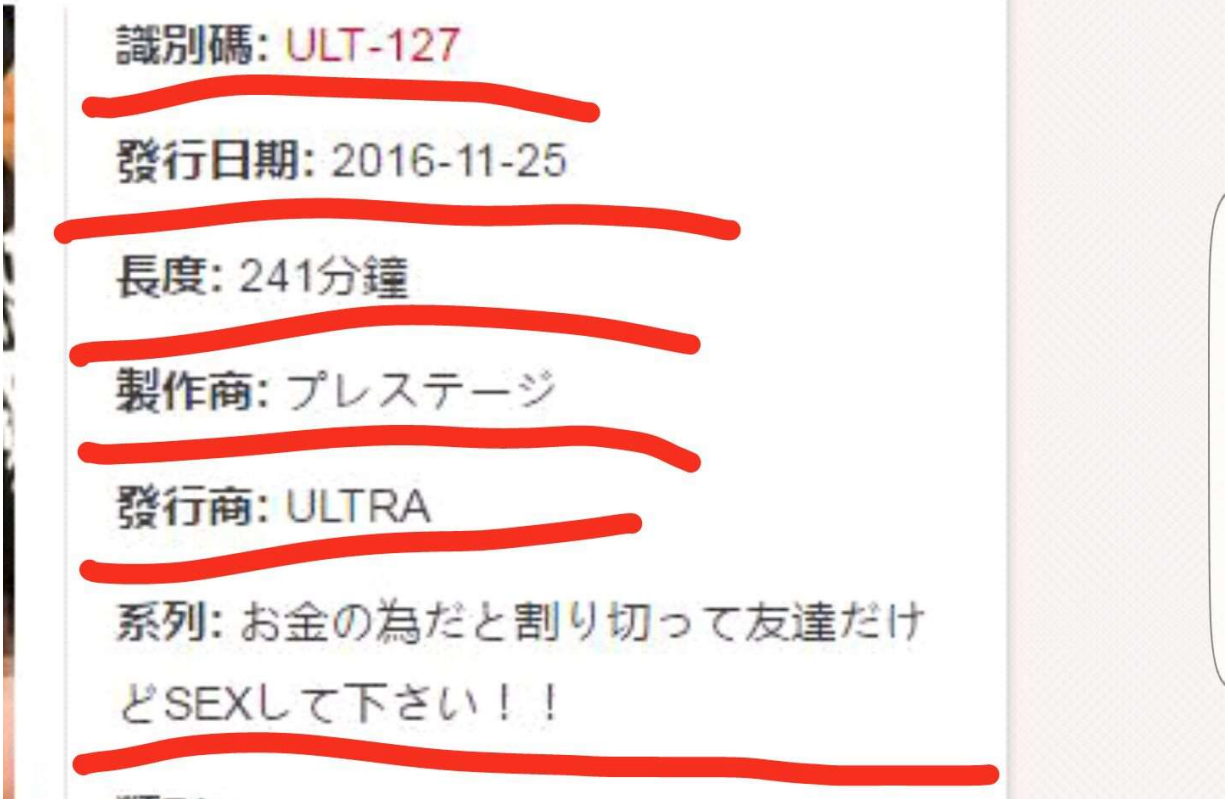
演員+

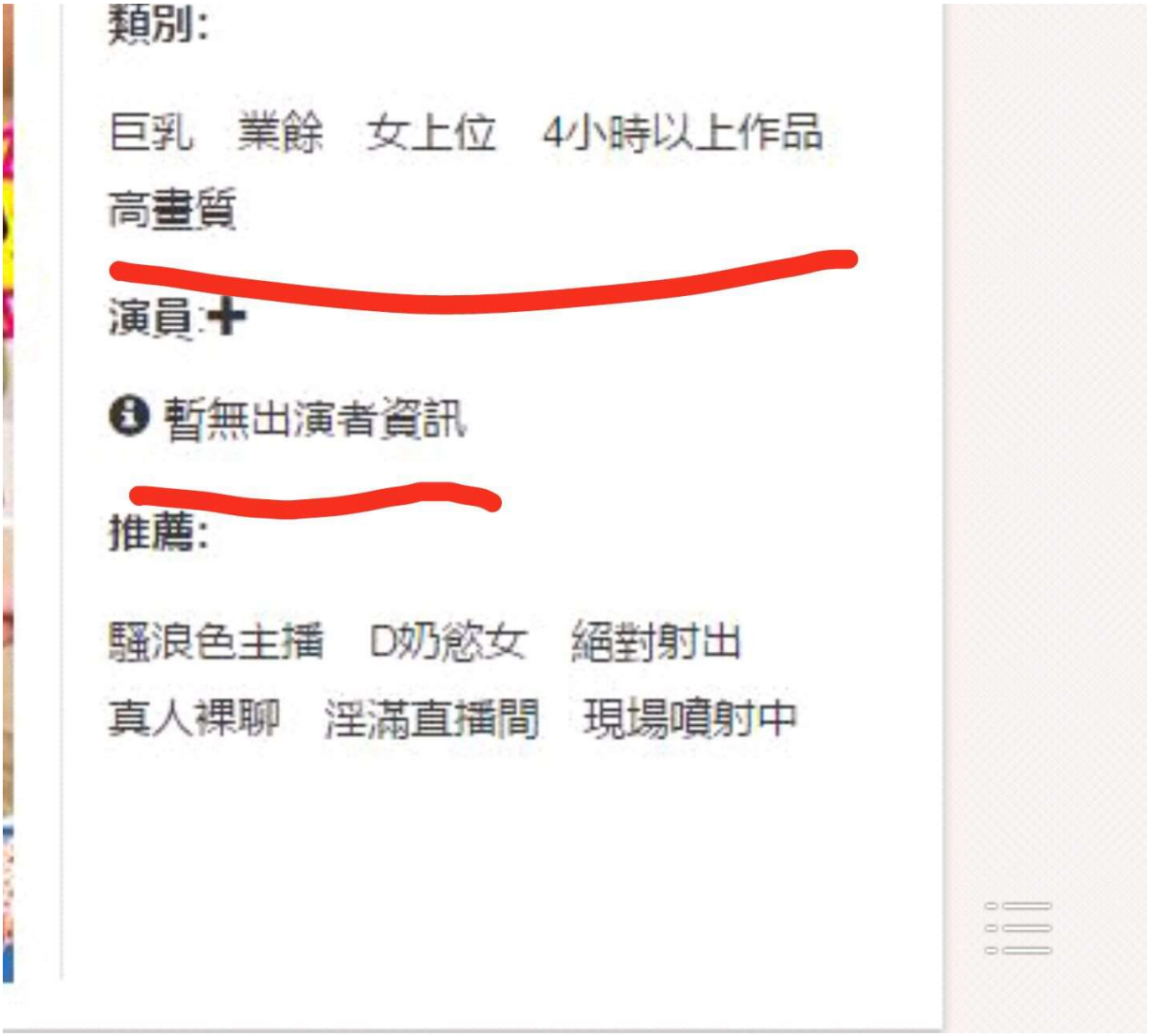
水咲菜々美

推薦:

絕對射出　真人裸聊　現場噴射中

D奶慾女　騷浪色主播　淫滿直播間





比如上图的两个就不同，html代码更是稀烂，需要判断有没有这个项，没有就设置空字节入库

在这其中纠结了一个问题

```
duration = soup.find('span', text="長度:").parent.contents[1].strip() if soup.find('span', text="長度:") else ''
print duration
```

```
duration_doc = soup.find('span', text="長度:")
duration = duration_doc.parent.contents[1].strip() if duration_doc else ''
```

就是对于这两种的比较，我想上面这种变成下面这种，毕竟第一种的话，soup.find要执行两次，但是下面这种又要比上面那个多一行，丑一点

最后我选择了第二种，所有的信息分析代码就不贴了，具体想看的直接看我的代码文件就好了

回到顶部

小Tips

1. 对于动态加载的内容的爬取，能不用selenium去模拟浏览器爬取就不用，耗费资源，更好的是自己分析网络请求，然后构造
2. 对于页面信息的解析，要多看几个页面，看是否相同，别到时候做多事情
3. 多看别人的博客学习思路

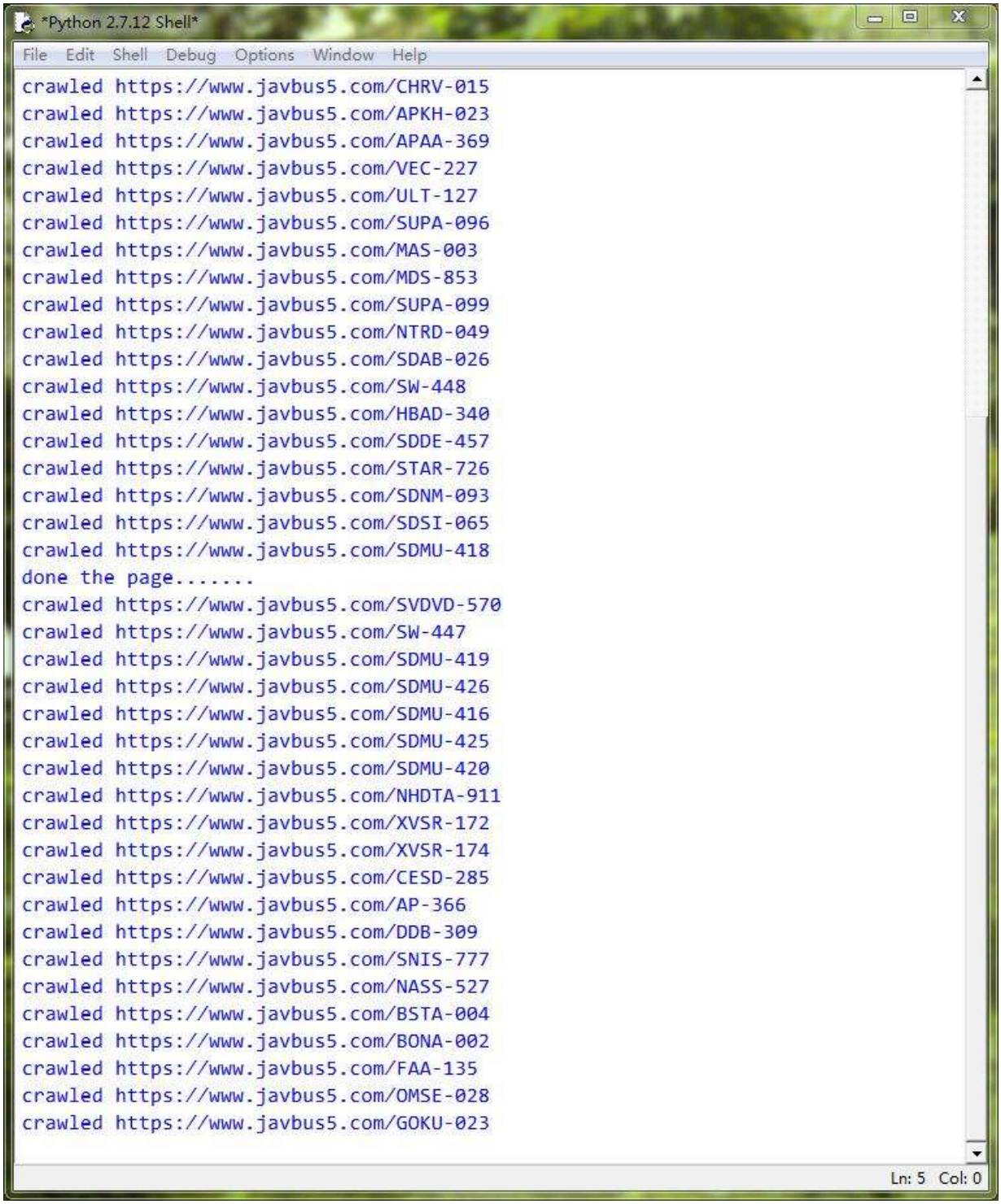
回到顶部

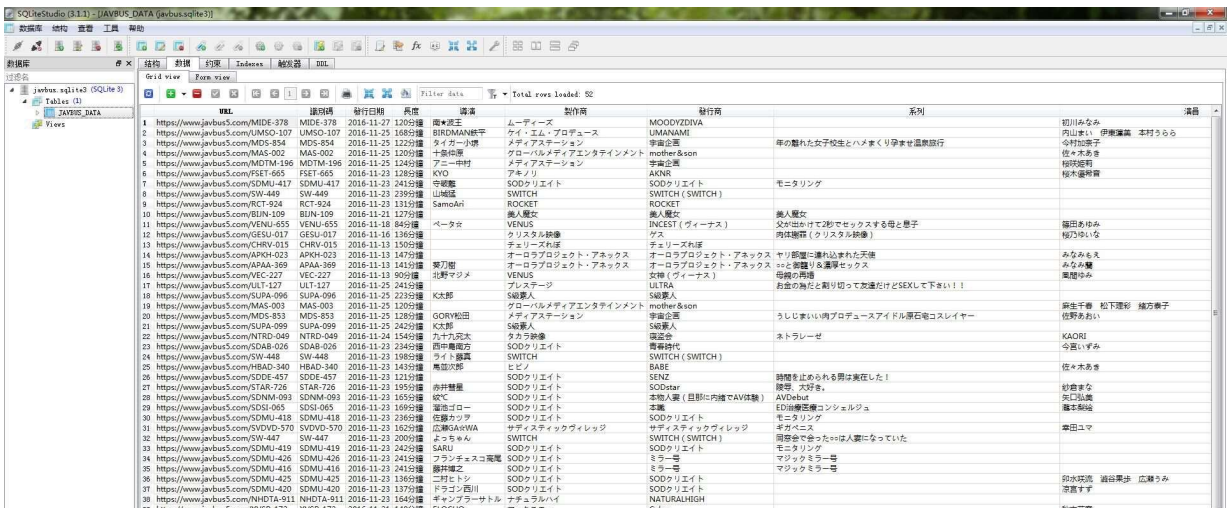
注意

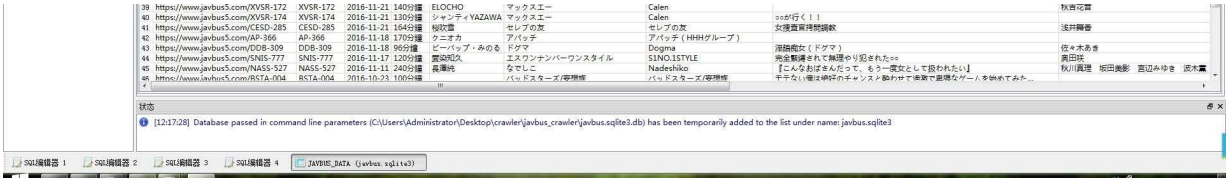
爬虫依赖的第三方库有Requests，BeautifulSoup，使用前请先pip install这两个第三方库

回到顶部

测试展与地址


Python 2.7.12 Shell
File Edit Shell Debug Options Window Help
crawled https://www.javbus5.com/CHRV-015
crawled https://www.javbus5.com/APKH-023
crawled https://www.javbus5.com/APAA-369
crawled https://www.javbus5.com/VEC-227
crawled https://www.javbus5.com/ULT-127
crawled https://www.javbus5.com/SUPA-096
crawled https://www.javbus5.com/MAS-003
crawled https://www.javbus5.com/MDS-853
crawled https://www.javbus5.com/SUPA-099
crawled https://www.javbus5.com/NTRD-049
crawled https://www.javbus5.com/SDAB-026
crawled https://www.javbus5.com/SW-448
crawled https://www.javbus5.com/HBAD-340
crawled https://www.javbus5.com/SDDE-457
crawled https://www.javbus5.com/STAR-726
crawled https://www.javbus5.com/SDNM-093
crawled https://www.javbus5.com/SDSI-065
crawled https://www.javbus5.com/SDMU-418
done the page.....
crawled https://www.javbus5.com/SVDVD-570
crawled https://www.javbus5.com/SW-447
crawled https://www.javbus5.com/SDMU-419
crawled https://www.javbus5.com/SDMU-426
crawled https://www.javbus5.com/SDMU-416
crawled https://www.javbus5.com/SDMU-425
crawled https://www.javbus5.com/SDMU-420
crawled https://www.javbus5.com/NHDTA-911
crawled https://www.javbus5.com/XVSR-172
crawled https://www.javbus5.com/XVSR-174
crawled https://www.javbus5.com/CESD-285
crawled https://www.javbus5.com/AP-366
crawled https://www.javbus5.com/DDB-309
crawled https://www.javbus5.com/SNIS-777
crawled https://www.javbus5.com/NASS-527
crawled https://www.javbus5.com/BSTA-004
crawled https://www.javbus5.com/BONA-002
crawled https://www.javbus5.com/FAA-135
crawled https://www.javbus5.com/OMSE-028
crawled https://www.javbus5.com/GOKU-023
Ln: 5 Col: 0





回到顶部

代码地址:

- coding.net javbus_crawler
- github.com javbus_crawler

司机的名声总算是没有辱没，秋名山依旧，嘿嘿

转载请注明来源作者

- 博客：akkuman.cnblogs.com | hacktech.cn
- 作者：Akkuman

分类: Python

标签: Life, Python

Akkuman
关注 - 12
粉丝 - 1

0 0

+加关注

« 上一篇：突破百度云限速与网页限制批量下载
» 下一篇：hexo在github和coding.net部署并分流（一）

posted @ 2016-12-06 18:22 Akkuman 阅读(1306) 评论(0) 编辑 收藏

刷新评论 刷新页面 返回顶部

公告

昵称：Akkuman
园龄：8个月
粉丝：1
关注：12
+加关注

搜索

找找看
谷歌搜索

常用链接

我的随笔
我的评论
我的参与
最新评论
我的标签

随笔分类(36)

Golang(5) Hacker(10)
Python(7) 读书笔记(1)
建站(2) 逆向工程(8)
一些Tips(1) 一些收藏(2)

友情链接

Akkuman's Blog

最新评论

1. Re:Windows环境下32位汇编语言程序设计笔记-基础篇
@beautifulzzzz最近对逆向感兴趣，看别人说这本书不错，就找来看看^^...
--Akkuman
2. Re:Windows环境下32位汇编语言程序设计笔记-基础篇
玩这么老的技术哈！
--beautifulzzzz

阅读排行榜

1. javbus爬虫-老司机你值得拥有(1304)

2. Golang模拟客户端POST表单功能文件上传(553)
3. Visual Studio Code配置Python开发环境(517)
4. s2-045漏洞批量检测工具(275)
5. Python之Requests的高级用法(167)

评论排行榜

1. Windows环境下32位汇编语言程序设计笔记-基础篇(2)