爬虫学习阶段性总结

爬虫的基础知识我打算就先学到这里了，以后需要用起来的时候再去看看相关文档和谷歌，做一个小量级的爬虫程序问题不大，对于分布式的和增量更新去重等需求就直接上框架，用别人的轮子还是蛮爽的。

简单小量级：requests+pyquery

JS渲染太多的：selenium+Phantomjs

框架：Pyspider或者Scrapy，个人比较喜欢Scrapy，主要是pyspider的文档真的少，两者框架差不多的，前者有WEBUI，后者是命令行模式，看喜欢哪个就用哪个吧。

学习资料总结：

第一部分基础

1. 环境搭建：

建议直接用虚拟机ubuntu（自带python2和3），terminal敲代码就可以了

1. python基础（0基础的看这个）：

<http://www.runoob.com/python3/python3-tutorial.html>

1. 看完上面这个之后看下面这个：（感觉作者写到错误调试那一章之后就不是太好了，感觉是一股脑东西给你砸过来）

<https://www.liaoxuefeng.com/wiki/0014316089557264a6b348958f449949df42a6d3a2e542c000>

扩展阅读：

《简明python教程》

<https://molun.net/byte-of-python-2017-new-translation-edition-release/>

1. Git教程

<https://www.liaoxuefeng.com/wiki/0013739516305929606dd18361248578c67b8067c8c017b000>

这个写的蛮好的，看这个就够了，后续需要的项目代码可以从github上拷贝，在上面也有很多的爬虫项目可以参考，注意看star和fork数量，可以参考时间比较新的（push时间），算是一个挺好的资源来源（比自己百度去搜索的爬虫很多都过期了，或者网站策略 更新了，对新手不太友好，会浪费比较多时间）

第二部分爬虫：

1. 崔庆才博客和教程视频：

<http://cuiqingcai.com>

教学视频在youtube有，我也存到本地网盘了。

视频内容都挺好的，一步步做下来，基本也就入门了爬虫了

1. 阅读相关的基础知识：

上面做爬虫的时候会经常遇到一些概念（前端），崔是做前端的，所以有些讲的很快，不懂的时候建议把视频暂停下来，看看下面这些概念

HTTP/HTML/AJAX/JSON/CSS/XPATH

教程都可以在这里找到：

<http://www.runoob.com/>

看重点就可以了，不需要全部都看，比如HTML就看一下基础元素属性和总结就好了，需要用到其他的时候再补

1. 数据库的基础知识：

教程都可以在这里找到：

<http://www.runoob.com/>

MySQL/MongoDB/Redis

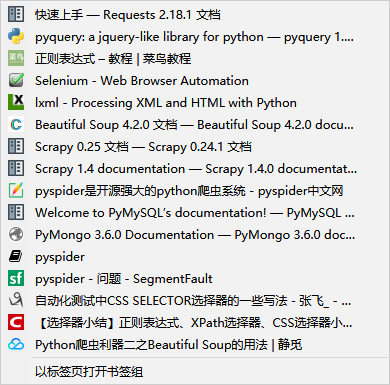
同理，看一下基础应用就好了，简单爬虫要求并不高，后续再深入学习。

1. 阅读相关的库文档，按照库文档的理解敲敲例子

这里包括请求库，解析库，存储库和工具库

大概是下面这些（资料通过百度或者谷歌都能搜到）

requests/re/selenium/lxml/beautifulSoup/pyquery/pyspider/scrapy/pymql/pymongo



重点觉得应该看一下几个解析库，比如bs4/pyquery/lxml和用一下css/xpath语法，熟悉一下每次筛选之后是什么类型的数据，以及怎么遍历，取出自己想要的数据，每个例子最好都使用print(type())打印一下类型，加深理解。

第三部分爬虫练手资源：

1. 知乎分享的合集（有挺多不能用得了，但是可以参考一下名字，搜一下最新的文章）

<https://zhuanlan.zhihu.com/p/27938007>

1. 崔庆才视频中的几个项目

不需要太多。

第四部分其他经验：

1. 爬虫容易写出满足感，所以不要在这上面花太多时间（其实我需要数据的时候还是找其他人要的，所以加个QQ群，程序员都是乐于助人的），适可而止就好了，除非要在爬虫这个职业上做长期发展
2. 加多几个靠谱的python或者爬虫QQ群
3. 多做笔记，多写代码（写个爬虫可能只需要1个小时，但是写个文章就需要2个小时，在写文章的过程中加深理解，跟读书的时候做笔记是一样的）
4. 至少要做到以下几个爬虫例子：
5. requests+正则表达式爬取静态网页（最好是加入搜索关键词的），并加入多进程，数据库存储，文件下载（图片和文本）
6. requests+lxml+xpath爬取静态网页，其他同第（1）点
7. requests+bs4+css/xpath爬取静态网页，其他同第（1）点
8. requests+pyquery+css爬取静态网页，其他同第（1）点
9. selenium+Phantomjs爬取静态网页，其他同第（1）点
10. pyspider+ selenium+Phantomjs爬取静态网页，其他同第（1）点（静态网页用pyspider爬感觉大材小用）
11. scrapy爬取动态网页，其他同第（1）点
12. 找一个封IP和cookies的网站（比如微博），用scrapy爬取，把几个pipeline都用起来，然后加入分布式爬取（找3个云服务器就ok了，一个发布任务，两个爬取），其他同第（1）点
13. 坚持，适可而止

第五部分 其他资源：

1. 反爬文章参考：

这个写的很全面也挺棒的：

<http://www.freebuf.com/articles/web/137763.html>

这是一个实例演示，也不错：

<http://www.freebuf.com/news/140965.html>

这个破解例子不错：

<http://blog.csdn.net/bone_ace/article/details/71056741>

1. 一个完整的爬虫+前端界面（真的挺棒的！）

<https://github.com/GuozhuHe/webspider>

