爬虫自学之路

# 简介

理论上来说，任何支持网络通信的语言都是可以写爬虫的，爬虫本身虽然语言关系不大，。但是，总有相对顺手、简单的。

目前来说，大多数爬虫是用后台脚本类语言写的，其中python无疑是用的最多最广的，并且页诞生了很多优秀的库和框架，如scrapy、BeautifulSoup 、pyquery、Mechanize等。但是一般来说，搜索引擎的爬虫对爬虫的效率要求更高，会选用c++、java、go(适合高并发)，

因为我们的目标不是搜索引擎，因此我选用python来学些爬虫。

学习之路如下

1. 了解爬虫
2. 学习Python基本语法，并熟练使用
3. 学习Python中关于爬虫的几个重要的内置库：urllib/http/Cookie等
4. 学习正则表达式，beautifulsoup等解析网页的工具或包
5. 利用上几步学习的至少爬取比较简单的网站，比如一些应用市场等等，不需要登陆
6. 学习利用工具分析网页请求流程、学习模拟登陆，拿新浪微博、知乎等需要登陆的网站进行练习
7. 学习Python中关于多线程、多进程的东西，将以前写的代码改为多线程版本，提高效率 7. 学习Python中的爬虫框架，或者自己写一个爬虫框架。 更高级的，还有验证码识别、js动态请求、js执行、代理IP等等。 推荐一个我自己的专栏：撸代码，学知识 - 知乎专栏，里边有一些爬虫的基础，以及自己尝试写的一个小型爬虫框架。

# 了解爬虫

爬虫简单的理解，就是通过程序模拟人操作网络发送请求，获取数据返回，清洗，筛选，整理出有用的数据，结构化保存数据，方便数据的浏览，计算，可视化，最终实现数字的描述型价值与预测型价值

[【爬虫二】爬虫的框架和基本议题](https://zhuanlan.zhihu.com/p/22002718)

一篇文章了解爬虫技术现状：http://blog.csdn.net/albenxie/article/details/72638114

零基础如何学爬虫技术？：https://www.zhihu.com/question/47883186

# 技术准备

python 基础是学习python爬虫不可少的,python基础学习资源：

Blog：

[廖雪峰的官方网站](https://www.liaoxuefeng.com/wiki/0014316089557264a6b348958f449949df42a6d3a2e542c000)

视频：

python 视频下载 http://blog.csdn.net/weixin\_39415084/article/details/78256803

python 视频在线 http://video.1kejian.com/itjishu/python/

# 环境搭建

python 2.7.14 + PyCharm

# 入门实战

1. 爬一爬[小猪的租房信息](http://cd.xiaozhu.com/)

仿照下面地址的介绍抓取第一批title列表数据

https://www.zhihu.com/question/47883186 （搜索套路：Requests+Xpath）

全部code查看：learn\_title.py

1. 五分钟入门

https://zhuanlan.zhihu.com/p/29433134

1. Python爬虫入门

<http://cuiqingcai.com/927.html>

# 开源库/框架学习

常用库和框架如下：

1. lxml

pip install lxml 安装

用于数据解析

1. requests

pip install requests 安装

用于http请求

1. beautifulsoup

pip install bs4 安装beautifulsoup

这个包可以很好地从网页代码中提取想要的数据

1. scrapy

pip install scrap 安装开元框架

官网：https://doc.scrapy.org/en/latest/

[基于Scrapy网络爬虫的搭建](http://www.lining0806.com/%E5%9F%BA%E4%BA%8Escrapy%E7%BD%91%E7%BB%9C%E7%88%AC%E8%99%AB%E7%9A%84%E6%90%AD%E5%BB%BA/)

1. PhantomJS

没有UI的webkit内核的浏览器

1. Pyspider

<http://www.pyspider.cn/>

pyspider all (启动PhantomJS) 可以加载js动态解析

# 数据存储

* Mysql

Mysql下载：

<http://www.runoob.com/mysql/mysql-install.html>

<https://dev.mysql.com/downloads/file/?id=473605>

mysql [命令行](https://www.baidu.com/s?wd=%E5%91%BD%E4%BB%A4%E8%A1%8C&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1dBnWbkPyc3my7bPWfznAn10ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3En163rHcsn1Rv)里执行 show variables like '%datadir%';

文件放在：C:\ProgramData\MySQL\MySQL Server 5.7\Data

常用图形化工具：https://www.cnblogs.com/emanlee/p/4471011.html

随笔- 50  文章- 0  评论- 1

[python操作mysql——mysql.connector](http://www.cnblogs.com/vathe/p/6964086.html)

<https://www.cnblogs.com/vathe/p/6964086.html>

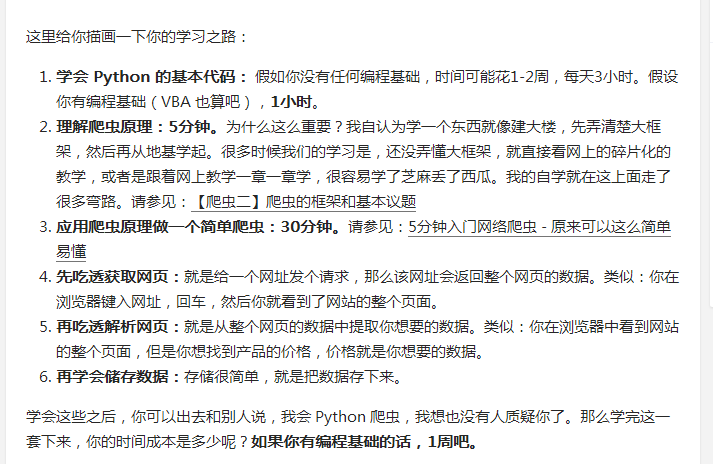
# 项目实战

1. 友盟数据

[友盟：教你如何使用统计分析工具做运营](http://tech.163.com/14/0327/11/9OBCICQV00094ODU.html)

[友盟页面爬虫工具](http://blog.csdn.net/henni_719/article/details/51900065)

百度搜索：友盟的统计数据 openapi



－－－－－－基础篇：－－－－－－－－－－

－－－基础库**urllib,urlllib2**,通过 openurl('url') 发送请求，接受返回数据。注意大部分网会对url请求做安全验证

－－－毕竟安全第一，生产第二，验证方法复杂多样[防盗链，cookie，登录验证，各种密钥，动态token]，这正是数据爬虫的难度所在，也是爬虫工作者存在的价值。通过各种验证顺利完成请求，后面python正则分分钟，格式化出你想要的数据。

－－－－－－进阶篇：－－－－－－－－－－

－－－取长补短，善假于物－－－－－－－－－－－－学习**Requests**库，主要是简化请求工作

－－－**Beautiful Soup**取代枯燥易错的正则表达式，简化数据匹配，快速清洗

－－－**Xpath语法lxml**主要了解文档结构，知己知彼，好下手

－－－**PhantomJS** 处理网页中的动态js请求，模拟浏览器完成js的渲染，等到异步情节，保证文档结构数据的完整性，所见即所得

－－－**Selenium** 模拟浏览事件，支持多步连续请求，减少人工手动触发［登录表单填写，触发表单发送，调整登录成功页等］

－－－**pyQuery**顾名思义，用python语言复制一版jquery,做前端的小伙伴都懂

－－－－－－高级篇：－－－－－－－－－－

－－－就是学会用框架，能站在巨人肩膀上的人，能力往往都会会太差，推荐学习**PySpider**，体验后真的很赞