从11月17号开始学爬虫，到现在将近1个月

期间完成了

1. 基于Python语言的爬虫开发环境搭建（windows10/64bit和ubuntu17.10/64bit）
2. 完成了菜鸟官网(HTTP/HTML/AJAX/JSON/CSS/XPATH)以及简单的MySQL/Redis/MongoDB的学习
3. 阅读

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

的官方文档和通过例子的编写实现了一些基本功能的理解

1. 通过崔庆才的视频教程完成了基本的一些简单爬虫的编写和理解

遇到的问题和实现的项目都已经归档到博客：

<http://www.cnblogs.com/copywang/>

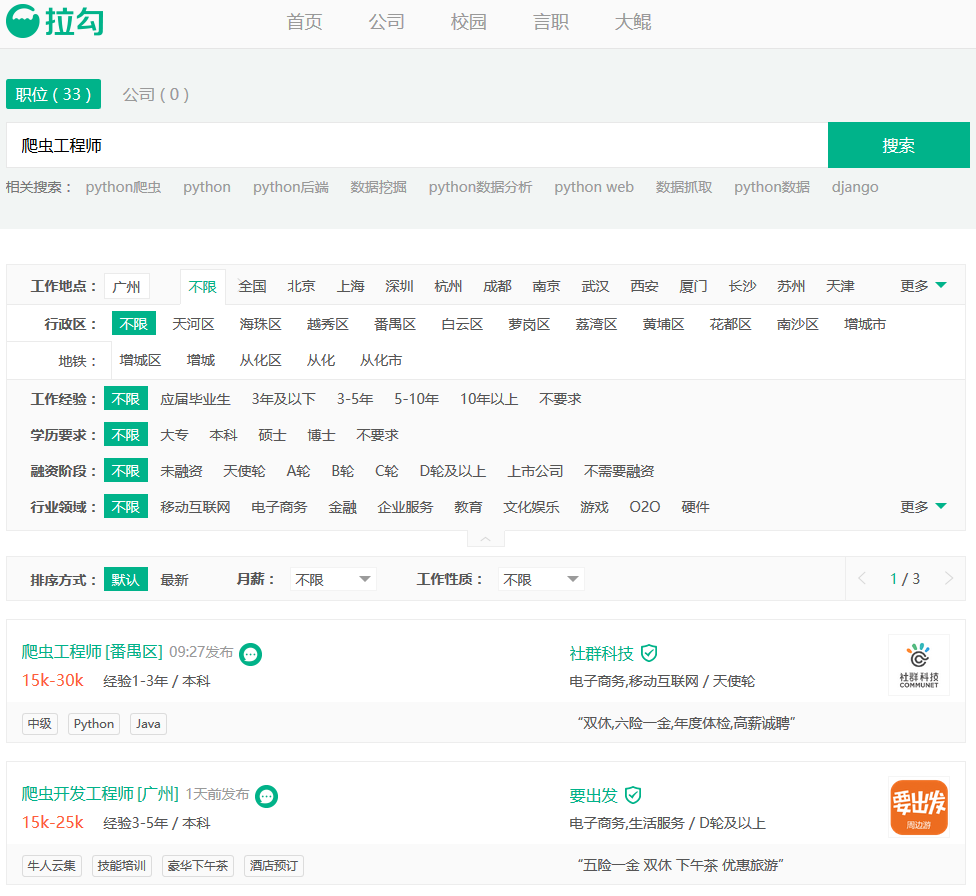
回头看看这些天的劳动，反思以下：

1. 环境问题导致的重复安装劳动（以后直接切换到Ubuntu环境），特别是框架安装时候的问题比较多（pyspider/scrapy），重复安装至少耗费了2天时间
2. 基础知识不牢固，在实现scrapy的时候，Pycharm中的external\_lib每次创建新项目和在原有项目中再打开是不一样的，出来了import错误，折腾了1天
3. 数据类型辨识错误，没有理解好各个不同的解析库解析出来的是class还是str，在调试过程每次都要打印type，浪费了比较多的编程时间
4. 不能迷信框架，在调试pyspider/scrapy的过程中，出现了比较多的调用错误，虽然框架在一定程度上减少了调用者的编程难度，在没有阅读和理解源码的情况下，没有理解到各个部件之间的联系，并不会在多大程度上给新手带来好处，如果是工作条件下的追求效率实现，倒是没什么问题，但是对于新手一上来就使用框架，并没有想象中那么好
5. 多使用基本的request和lxml
6. 动态语言写起来爽，但是在debug的时候真的要命
7. 增量爬取还没有学到任何资料

以上就是学习的内容，本吊智商捉急，无法达到大神那种一点即通的境界，只能来面向工作编程，快到年底，看看爬虫工程师的需求和岗位要求，有句话说得好，一个人的成功，纪要讲究个人努力，也要配合历史进程

1. 拉勾网的爬虫工程师需求

广州地区（本吊不能出去其他城市找工作）



全部只有33个工作岗位，而且从最近5个月我的观察来看（前面四个月去浪了~~~），一直都是这些公司在招聘

包含：电商，自媒体，金融（P2P）

经验要求：1-3年居多，大专本科起步（可见其实爬虫工程师的门槛还是比较低的）

薪资：自不用说，拉勾上出名的工资给高，所以如果标识10K-15K，基本上也就是10K（8K就算给你面子了）

技能要求：

任职要求：（我找了一个工资最高的<https://www.lagou.com/jobs/3436801.html> ）

1. 1年或以上爬虫开发相关经验，计算机相关专业本科以上学历；

2. 具有微信、微博至少一种自媒体平台的爬取经验；

3. 熟悉Python或Java语言；

4. 熟悉多线程、网络编程，精通网页抓取原理及技术，精通正则表达式，能从结构化的和非结构化的数据中获取信息；

5. 熟练使用MySQL数据库，掌握Redis、MongoDB、Hive等常用NoSQL技术并具有实战经验；

6. 精通一种开源爬虫框架，有开发爬虫框架经验优先；

1. 善于沟通，有良好的文档写作能力，口头沟通能力，良好的团队合作精神，良好的抽象思维。

再找个工资比较低的（<https://www.lagou.com/jobs/2735224.html> ）

职位职责：  
1、 负责开发网页数据的自动化爬取脚本程序及自动化解析脚本程序；  
2、 负责清洗、分类已解析的互联网数据；  
3、 负责自动化数据解析系统的部分模块开发；  
4、 参与核心爬虫系统架构设计、数据库设计；  
5、 分析爬虫系统的技术缺陷，对策略架构做出合理地调整和改进。  
任职要求：  
1、 计算机、软件、电子信息等相关理工科专业，1年以上相关工作经验；  
2、 熟悉python爬虫开发，熟悉常用的信息抓取策略和数据重组算法；  
3、 了解Mongodb、HBase、HIVE等NoSQL数据库，并至少精通一种关系型数据库的开发；  
4、 熟悉大规模网页爬取，深度网页爬取，熟悉Nutch、Scrapy、Lucene、Heritrix、Solr、Sphinx等工具优先考虑；  
5、 熟悉网络爬虫相关技术，如有分词、spider、分类、聚类、索引、检索等相关开发经验者优先；  
6、 热爱软件编程，愿意与人分享专业知识和经验。

从上面其实看得到：

1. 大公司对于专职的爬虫工程师需求没有（原因可能是，一来爬虫门槛不高，大公司的程序员，特别是前端的对于爬虫框架其实从零开始难度不大，二来大公司自己本身的数据已经足够做市场对比，不需要再从其他网站爬取），这里就不说某度了，搜索引擎公司肯定是有自己的爬虫框架的，另外，就学了2个月爬虫，想进去做搜索引擎的爬虫是不是好高骛远了
2. 爬虫设计涉及的项目经验要求比较高，一来出来要设计分布式架构，还要存储架构，还有增量更新方法，且不说去实现一个公司级别要求的爬虫系统，难度是很高的
3. 对于工资普遍都高的拉钩都能给出这么低的工资，看来爬虫工程师得真的是热爱爬虫才干得下去（个人猜测转行做后端/机器学习的不在少数）

总结：当前掌握的知识和编码水平，连最低级别的爬虫工程师都达不到

任重而道远

从职业发展看，不觉得爬虫工程师是个好的需求了。。