爬虫项目总结

一：项目一：名称《爬取微小宝数据https://data.wxb.com/rank》

《爬去微口网http://www.vccoo.com/》

1. 项目环境：python3.6，windows
2. 项目框架：scrapy 或 scrapy+selenium(请求限速使用)
3. 爬取需求：1)爬取微信公众号（公众号名称，阅读量，点赞量，头条数，发文篇数，小宝指数，粉丝数）

2)爬去微口网(文章标题，公众号名称，阅读量，类 型)

1. 应对的反爬【a) 需要模拟登陆，登录以后获取headers请 求头信（Accept,Cookie,Host），
2. 限制访问速度：用uniform（3,7）随机一个时间，

C) 下载中间件里面使用(随机User-Agent，随机 ip(构建ip池)】

D) 数据请求放在script标签里面

5.拼接url，解析json数据，通过item返回。

6.通过pipelines(管道)存储数据到mongodb，或 csv文件

7.将定时爬去写入定时爬去(apscheduler)

二：项目二：名称 《爬去大众点评https://www.dianping.com/》

1. 项目环境：Python3.6 + windows
2. 使用scrapy + selenium
3. 爬取需求：抓取大众点评（店名，电话，地址，好评数，中 评数，差评数，以及评论信息）
4. 遇到的反爬机制以及应对【 a) . 需要模拟登陆拿到cookie

B).模拟登陆时，找到隐藏在iframe 标签里面的url，跳转到真实的登录的url】

5. a). 在中间件里面使用selenium+headless(无头请求)，使用随机 ua,和随机ip

b). 用selenium模拟登陆判断如果是登陆页面的url，就做模拟 登陆的过程，如果不是直接请求url并返回 Htlmresponse(url,boby=driver.page\_source.encode)

c). 然后在parse函数里面进行解析。

d). 返回item ，通过pipelines(管道)将数据以json格式的字符 串存到txt文件里面，

e). 在run.py使用定时爬去调用（apscheduler）

项目三：爬去赶集网二手车信息（http://xa.ganji.com/ershouche/）

项目环境：Python3.6

项目框架：scrapy

爬去需求：(公司名联系人，联系电话(车商),车名,颜色，价格,车 龄,公里数)

反爬机制以及应对：a).获取电话的时候属于(点击获取)，属于ajax 的post请求，并且post请求的参数通过 哈希编码。【这里我采用爬取手机端】

b).获取到的详情页的url需要进行拼接，首先 需要获取到id然后在进行拼接。用 xpath拿出的href标签里面的url格式不对

正确的格式 【http://xa.ganji.com/ershouche/36419119657233x.htm?type=jx】

1. 请求start\_url获取详情页的url，将函数回调到下个函数解析 （）
2. 请求详情页的url 解析获取puid 和以上需要解析的信息，在 通过meta函数将参数传递到下个函数
3. 请求点击获取 url=”http://xa.ganji.com/ajax.php?\_pdt=che&module=get\_phone400”,

发送post请求 formdata= {‘puid’,36419119657233,’key’:””}

使用json.loads()

将json格式的字符串转换成python的对象或者字典。

再用response.meta将上面的上个函数的参数接出来，放入 item返回。存储到mysql数据库

四：项目四：爬取企查查等上面企业的信息(pdf文档)

|  |  |  |
| --- | --- | --- |
| 1.亿欧网-1 | https://www.iyiou.com/intelligence/resource?date=2018 | 报告名称，报告正文PDF文件，分类及来源，发布时间 |
| 2.亿欧网-2 | https://www.iyiou.com/intelligence/report/ | 报告名称，报告正文PDF文件，分类及来源，发布时间 |
| 4.艾瑞咨询 | http://report.iresearch.cn/ | 报告名称，报告正文PDF文件，分类及来源，发布时间 |
| 5.清科免费报告 | https://free.pedata.cn/ | 报告名称，报告正文PDF文件，分类及来源，发布时间 |
| 6.艾媒报告 | http://report.iimedia.cn | 报告名称，报告正文PDF文件，分类及来源，发布时间 |
| 7.企查查行研库 | https://www.qichacha.com/elib\_research | 报告名称，报告正文PDF文件，分类及来源，发布时间 |
| 8.招股书-企查查库 | https://www.qichacha.com/elib\_prospectus | 股票名称，股票代码，招股书名称，招股书正文PDF文件，发行日期 |
|  |  |  |

1. 这里我们依然使用windons+python.3.6 scrapy urlib.request

2.拼接分页的url，

3.请求拼接好的url，获取pdf文档的链接，使用item返回通过管道解析，获取pdf文档存到本地

*for* i *in item*['pdfwei']:  
 file\_path = 'D:\数据\报告'  
 file\_name = i.split('/')[-1]  
 *try*:  
 *if not* os.path.exists(file\_path):  
 os.makedirs(file\_path)  
 filename = '{}{}{}'.format(file\_path, os.sep, file\_name)  
 urllib.request.urlretrieve(i, filename)  
 *except* Exception *as* e:  
 print(e)