今天为大家整理了23个Python爬虫项目。整理的原因是，爬虫入门简单快速，也非常适合新入门的小伙伴培养信心。所有链接指向GitHub，祝大家玩的愉快！

1、WechatSogou [1]- 微信公众号爬虫。

基于搜狗微信搜索的微信公众号爬虫接口，可以扩展成基于搜狗搜索的爬虫，返回结果是列表，每一项均是公众号具体信息字典。

github地址：https://github.com/Chyroc/WechatSogou

2、DouBanSpider [2]- 豆瓣读书爬虫。

可以爬下豆瓣读书标签下的所有图书，按评分排名依次存储，存储到Excel中，可方便大家筛选搜罗，比如筛选评价人数>1000的高分书籍；可依据不同的主题存储到Excel不同的Sheet ，采用User Agent伪装为浏览器进行爬取，并加入随机延时来更好的模仿浏览器行为，避免爬虫被封。

github地址：https://github.com/lanbing510/DouBanSpider

3、zhihu_spider [3]- 知乎爬虫。

此项目的功能是爬取知乎用户信息以及人际拓扑关系，爬虫框架使用scrapy，数据存储使用mongo

github地址：https://github.com/LiuRoy/zhihu_spider

4、bilibili-user [4]- Bilibili用户爬虫。

总数据数：20119918，抓取字段：用户id，昵称，性别，头像，等级，经验值，粉丝数，生日，地址，注册时间，签名，等级与经验值等。抓取之后生成B站用户数据报告。

github地址：https://github.com/airingursb/bilibili-user

5、SinaSpider [5]- 新浪微博爬虫。

主要爬取新浪微博用户的个人信息、微博信息、粉丝和关注。代码获取新浪微博Cookie进行登录，可通过多账号登录来防止新浪的反扒。主要使用 scrapy 爬虫框架。

github地址：https://github.com/LiuXingMing/SinaSpider

6、distribute_crawler [6]- 小说下载分布式爬虫。

使用scrapy,Redis, MongoDB,graphite实现的一个分布式网络爬虫,底层存储MongoDB集群,分布

式使用Redis实现,爬虫状态显示使用graphite实现，主要针对一个小说站点。

github地址：https://github.com/gnemoug/distribute_crawler

7、CnkiSpider [7]– 中国知网爬虫。

设置检索条件后，执行src/CnkiSpider.py抓取数据，抓取数据存储在/data目录下，每个数据文件的第一行为字段名称。

github地址：https://github.com/yanzhou/CnkiSpider

8、LianJiaSpider [8]– 链家网爬虫。

爬取北京地区链家历年二手房成交记录。涵盖链家爬虫一文的全部代码，包括链家模拟登录代码。

github地址：https://github.com/lanbing510/LianJiaSpider

9、scrapy_jingdong [9]– 京东爬虫。

基于scrapy的京东网站爬虫，保存格式为csv。

github地址：https://github.com/taizilongxu/scrapy_jingdong

10、QQ-Groups-Spider [10]– QQ 群爬虫。

批量抓取 QQ 群信息，包括群名称、群号、群人数、群主、群简介等内容，最终生成 XLS(X) / CSV 结果文件。

github地址：https://github.com/caspartse/QQ-Groups-Spider

11、wooyun_public[11]-乌云爬虫。

乌云公开漏洞、知识库爬虫和搜索。全部公开漏洞的列表和每个漏洞的文本内容存在MongoDB中，大概约2G内容；如果整站爬全部文本和图片作为离线查询，大概需要10G空间、2小时（10M电信带宽）；爬取全部知识库，总共约500M空间。漏洞搜索使用了Flask作为web server，bootstrap作为前端。

https://github.com/hanc00l/wooyun_public

12、spider[12]– hao123网站爬虫。

以hao123为入口页面，滚动爬取外链，收集网址，并记录网址上的内链和外链数目，记录title等信息，windows7 32位上测试，目前每24个小时，可收集数据为10万左右

https://github.com/simapple/spider

13、findtrip [13]- 机票爬虫（去哪儿和携程网）。

Findtrip是一个基于Scrapy的机票爬虫，目前整合了国内两大机票网站（去哪儿 + 携程）。

https://github.com/fankcoder/findtrip

14、163spider [14] - 基于requests、MySQLdb、torndb的网易客户端内容爬虫

https://github.com/leyle/163spider

15、doubanspiders[15]- 豆瓣电影、书籍、小组、相册、东西等爬虫集

https://github.com/fanpei91/doubanspiders

16、QQSpider [16]- QQ空间爬虫，包括日志、说说、个人信息等，一天可抓取 400 万条数据。

https://github.com/LiuXingMing/QQSpider

17、baidu-music-spider [17]- 百度mp3全站爬虫，使用redis支持断点续传。

https://github.com/Shu-Ji/baidu-music-spider

18、tbcrawler[18]- 淘宝和天猫的爬虫,可以根据搜索关键词,物品id来抓去页面的信息，数据存储在mongodb。

https://github.com/pakoo/tbcrawler

stockholm [19]- 一个股票数据（沪深）爬虫和选股策略测试框架。根据选定的日期范围抓19、取所有沪深两市股票的行情数据。支持使用表达式定义选股策略。支持多线程处理。保存数据到JSON文件、CSV文件。

https://github.com/benitoro/stockholm

20、BaiduyunSpider[20]-百度云盘爬虫。

https://github.com/k1995/BaiduyunSpider

21、Spider[21]-社交数据爬虫。支持微博,知乎,豆瓣。

https://github.com/Qutan/Spider

22、proxy pool[22]-Python爬虫代理IP池(proxy pool)。

https://github.com/jhao104/proxy_pool

23、music-163[23]-爬取网易云音乐所有歌曲的评论。

https://github.com/RitterHou/music-163