2017.06 - 2017.07 各大旅游网站的景点信息抓取系统

软件环境： Python+MySQL

硬件环境： win

开发工具： PyCharm

责任描述： 负责抓取某省份的各大网站上的城市旅游景点信息（标题，评分，评论等）。把数据进行去重之后入MySQL库，其中使用了requests、BeautifulSoup模块进行抓取，设置了代理ip池和header。

项目描述： 该项目主要是获取评论的内容，评论，时间，景区等，首先要找到自助游景点进入该链接后找到游客点评的元素，将我们所要的文本内容提取出来放入数据库中；通过对评价的信息分析作出相应的调整，随时为游客们提供最好的服务。

2016.12 - 2017.03 各大酒店网站的酒店信息抓取系统

软件环境： Python+MySQL

硬件环境： win

开发工具： PyCharm

责任描述： 设计爬虫抓取逻辑并实现代码开发,对爬取的数据进行提取，使用的模块requests，time第三方库等，使用xpath，beatifulsoup，然后对数据进行抽取，使用数据库 mysql存取数据。

项目描述： 对一些旅游网站上的酒店、航空进行数据爬取，分析。首先确定网站的入口链接然后找到抓取的页面链接在将需要采集的字段清洗过滤最后存入mysql数据库。确定目标的数据所在栏目生成相关列表url，确抓去的字段对数据进行过滤，保存到mysql数据库。实现对数据的采集 数据入库效果。

2016.10 - 2016.11 东方航空航班信息抓取系统

软件环境： Python+MySQL

硬件环境： win

开发工具： PyCharm

责任描述： 负责抓取东方航空航班信息，其中主要使用scrapy框架进行抓取，在中间键里写了代理、header等，防止爬虫中断，数据用beatifulsoup解析的，入MySQL库。

项目描述： 该项目抓取航班信息。因为现在人们的生活水平的提高，在闲暇之余出游旅行是很多人的选择，而出行的交通工具也从以前的火车变成的现在的飞机，因为飞机节省时间。所以该项目的目的是用这些航班信息来进一步调整出游的计划，让客户玩的更愉快。

2016.06 - 2016.09 股票网站信息抓取系统

软件环境： Python+MySQL

硬件环境： win

开发工具： PyCharm

责任描述： 负责抓取几个财经网站的股票系统，其中主要用selenium中的Phantomjs无界面浏览器，获取原码，用Beatifulsoup进行解析，然后清洗入库，之后用Matplotlib模块做成K线图进行一个简单分析。

项目描述： 本项目主要进行财经网站的股票数据收集，抓取股票的价格，抓取出来的数据进行制图分析，可以分析出各大持股公司的财力及综合实力。

2016.04 - 2016.05 京东商品信息抓取系统

软件环境： Python2.7

硬件环境： win

开发工具： PyCharm

责任描述： 负责抓取京东商品信息（名称，价格，产品参数，评论等等），用Scrapy框架进行抓取，中间键要有代理ip和headers信息模拟浏览器，评论用接口抓取，re解析接口页面，最后入Mongo库。

项目描述： 该项目抓取京东商品信息。近几年网购是非常火的，在家就能买东西，还有送货上门和货到付款等福利，而且价格一般也比实体店便宜很多。该项目的目的是利用这些商品的走势，进行商品的性价比分析。

2015.03 - 至今 北京优炫软件股份有限公司 （2年7个月）

python爬虫工程师 | 10001-15000元/月

IT服务(系统/数据/维护)

工作描述： 工作描述：根据项目具体的需求，承担开发任务，按计划完成项目；

负责对项目的后台程序进行更新编程，数据分析，项目的研发；

负责网页信息数据抽取、消重等工作，提升平台的抓取效率；

负责网络爬虫cookie的登录、验证码、封IP封账号采集难点攻克；

负责系统后台分布式调度系统搭建，优化，数据分析导出等。

项目经历

2016.09 - 至今 抓取中国裁判文书网

责任描述： 1.设计爬虫抓取逻辑并实现代码开发；

2.使用的selenium+time模拟浏览器延时操作，并将对爬取的网页进行提取数据源码；

3.并存在mysql数据库中保证数据库和网页能够同步更新。

项目描述： 1.使用phantomjs无页面浏览器渲染js从而得到数据源码；

2.进行url去重，并采集数据；

3.并存在mysql数据中。

4书写python脚本并监控数据的实质性；

2016.02 - 2016.07 抓取境外景点评论信息

责任描述： 设计爬虫抓取逻辑并实现代码开发,对爬取的数据进行提取，清洗，使用模块 urllib ，time，urllib2，requests第三方库等，使用xpath，beatifulsoup,正则表达式对数据进行抽取和清洗，将数据存入monogo数据库中

项目描述： 对一些旅游（携程，旅行等）的网站进行数据爬取(酒店,机票，

服务，地点景区，自驾，组团价格)。

任务目标：首先确定网站的入口链接然后找到抓去的页面链接将需要采集的字段

存入mongodb数据库。

开发流程：确定目标的数据所在栏目生成相关列表url，确定抓去的字段对可变

话数据进行过滤，保存到mongodb数据库。

结论总结:实现对数据的采集 数据入库效果以及对模块的封装

2015.06 - 2016.01 电商平台数据抓取

责任描述： 需要获取各个大型电商网站平台的所有大品牌家电商品的一些信息--包括

（商品的品牌，商品价格，商品名称，商品URL，用户的评论，商品的颜色）

店铺情况（店铺名称，店铺的URL，店铺的信誉）

确定入口链接->具体抓取的页面链接->需要采集的字段->入库：

使用的数据库 MongoDB

使用的类库 resqusts、time、Beautifulsoup、pymongo等

使用的模块 random、FormRequest等

使用xpath，beatifulsoup对数据进行抽取和清洗

使用ip池，cookie池，user-agent池 防治反爬虫技术

使用其余的模块进行相应的操作以达我们想要的目的

采用Mongo DB作为本地数据库，将我们获取到的信息存放本地数据库Mongo DB中以供公司需求的提取

项目描述： 负责整个项目的抓取，指定的电商平台的数据进行提取，

清洗与入库操作，在获取数据的过程中遇见的反爬虫机制的处理

本人具有较强的专研和动手能力，有一定的Python爬虫工作能力，从事Python爬虫技术两年时间，熟悉Python反爬虫机制，熟悉一些常用的模块以及爬虫框架的应用如：scrapy框架；并可以灵活应用很多第三方库，如requests，xpath，urllib等，熟练使用爬取动态网页工具Selenium和PhantomJs；熟悉Python的Django框架；

在工作中，有强烈的求知欲，勤奋好学，能在指定的时间内完成安排的工作，我的性格比较开朗，为人谦和，富有事业心和责任感使我能够面对任何困难和挑战。

工作经历

2015.09 - 至今 北京源石智影科技有限公司 （2年1个月）

Python开发工程师（爬虫） | 10001-15000元/月

计算机软件 | 企业性质：民营 | 规模：20-99人

工作描述： 1.根据需求制定爬虫；

2.编写爬虫计划；

3.编写爬虫；

4.后期数据处理（数据清洗、去除杂质）；

5.测试爬虫时效，编写时效文档；

6.爬虫程序维护；

项目经历

2017.06 - 2017.08 票务类爬虫（猫眼、格瓦拉、时光网、微票）

责任描述： 1．编写爬虫程序，分析站点。

2．解决反爬模拟用户请求。

3．保证爬虫数据的完整性。

4．编写票务类爬虫的说明文档

5．项目完毕，部署服务器。

6．每天定时抓取，保持数据更新！

项目描述： 票务类爬虫项目，业务：电影售票门户网站的场次数据抓取和实时票务数据抓取。此项目主要抓取四个网站（猫眼、格瓦拉、时光网、微票）每天凌晨定时清除数据库表，然后抓取最新票务信息，存入数据库。

2017.02 - 2017.05 豆瓣网数据抓取系统

责任描述： 1.分析网页的URL。

2.豆瓣网爬虫项目编写。

3.PhantomJS源码获取。

4.项目调试运行。

项目描述： 该项目主要是获取评论，评论的内容，时间等，获取用户对电影的评价，在这个项目中使用MongoDB数据库，使用第三方库如：Xpath，Css等进行对数据的抽取。

2015.09 - 2015.12 网游,单机,手游游戏相关资讯爬虫

软件环境： ubuntu

开发工具： scrapy框架、request模块、mysql数据库等

责任描述： 1,负责设计和开发游戏资讯网络爬虫程序；

2,负责分析数据爬取过程，提供技术实现方案；

3,负责分析静态、动态的网页结构、解析网页游戏资讯数据、对数据进行提取、清洗、入库；

4,负责爬虫系统的后期维护；

项目描述： 本项目运用scrapy网络爬虫, 对大型游戏门户网站如17173, 游侠网, 太平洋游戏网等进行数据爬取分析。提取相关各种游戏信息, 新闻, 图片等资讯, 存储在服务器redis数据库中, 完成定时, 去重等功能。

2016.02 - 2016.05 电脑桌面壁纸爬虫

软件环境： ubuntu

开发工具： scrapy框架、request模块、redis数据库等

责任描述： 1,负责设计和开发桌面壁纸分布式网络爬虫程序；

2,负责分析数据爬取过程，提供技术实现方案；

3,负责分析静态、动态的网页结构、解析数个网页数据、对数据进行提取、清洗、入库；

4,负责爬虫系统的后期维护；

项目描述： 本项目运用scrapy框架, 建立多线程,多进程爬虫,对zol桌面壁纸, 彼岸壁纸, 美桌网等多个主流桌面壁纸网站进行数据,图片爬取，对图片标签, 图片主题，图片内容,按不同分辨率的对图片进行处理, 爬取, 然后根据需求将图片对应信息存入MySQL数据库中。最终部署在多台服务器上,用Crontab写运行脚本达到定时爬取。

2007.03 - 2017.07 图书比价网

软件环境： ubuntu

开发工具： scrapy框架、request模块、mysql数据库等

责任描述： 1,负责设计和开发图书网络爬虫程序；

2,负责分析数据爬取过程，提供技术实现方案；

3,负责分析静态、动态的网页结构、解析数个网页数据、对数据进行提取、清洗、入库；

4,负责爬虫系统的后期维护；

项目描述： 本项目主要是采集各大销售网站的图书的数据,从而帮助需要买书的认.主要爬取淘宝图书,京东图书,亚马逊图书,当当图书网等购书网站的图书详情.然后对数据进行统计,分析数据,存储数据等

2015.04 - 2016.06 腾讯新闻爬虫

软件环境： scrapy、xpath

硬件环境： Mac

开发工具： pycharm、sublime

责任描述： scrapy框架，xpath提取数据，mongo数据库

项目描述： 本项目采用scrapy框架，运用代理池，随机取出代理，对腾讯新闻网的内容进行爬取，包括大标题，小标题，文章具体内容。对抓取的内容进行清晰，存到数据库中

2016.06 - 2017.09 汽车之家爬虫，太平洋汽车网爬虫

软件环境： scrapy-redis分布式,Beautifulsoup解析器，redis数据库，PhantomJS，Selenium。

硬件环境： Mac

开发工具： pycharm，sublime

责任描述： scrapy-redis分布式，Beautifulsoup解析器，redis数据库，PhantomJS，Selenium.

项目描述： 本项目运用scrapy-redis组件对汽车之家，太平洋汽车网进行数据爬取，对文章标题，文章内容,汽车图片，汽车价格进行爬取。由于部分网页采取了动态网页js加载，因此采用了Selenium进行了对动态html的处理

2016.06 - 2017.09 携程旅游网信息分布式爬虫

软件环境： scrapy框架、scrapy-redis分布式组件、Redis

硬件环境： Mac

开发工具： pycharm、sublime

责任描述： scrapy框架、scrapy-redis分布式组件、Redis

项目描述： 这个项目是对携程旅游网门票，酒店信息抓取（针对票价信息、酒店信息及图片）进行抓取，采用scrapy-redis分布式实现。Windows 10 做为Master端，管理多个Linux Slave端爬虫集群。分布式使用Redis做为缓存数据库，利用Redis的高并发和I/O读写来实现高速下载。抓取过程中发现票价信息及图片均为JS动态加载的，导致信息抓取不完整，故采用开启下载中间件，利用selenium + PhamtonJS的方式发送请求，待数据加载完毕后返回提取。

2017.08 - 2017.09 python爬取拉钩网招聘信息

软件环境： 跨平台

硬件环境： pc机

开发工具： pycharm

责任描述： 负责项目规划，使用python多线程和协程提升爬虫工作效率。

分析某岗位招聘状态。初步实现可视化分析。

项目描述： 该项目可分析it行业招聘行情，最后可视化使结构更加清晰

2017.05 - 2017.06 新浪网分类资讯分布式爬虫测速

软件环境： linux，windows

责任描述： 参与python的scrapy-redis实现分布式部署

项目描述： 基于新浪网分类咨询部署分布式整站爬虫，

实现高性能爬虫系统

2017.03 - 2017.04 基于PySpider框架tripadvisor(猫途鹰)自动化爬虫开发

软件环境： linux，windows，mac

责任描述： 目标站点分析，项目编写，项目测试，以及数据清洗

项目描述： TripAdvisor是全球领先的旅游网站，官方中文名为“猫途鹰”。收录逾5亿条全球旅行者的点评及建议，覆盖超过190个国家的700万个住宿、餐厅和景点，并提供丰富的旅行规划和预订功能。

爬取猫途鹰指定站点，获得最新评论以及站点信息

2017.04 - 2017.04 爬取京东女鞋的信息(爬虫)

软件环境： Linux

硬件环境： Mac

开发工具： Pycharm

责任描述： 1.网页爬取，以及数据清洗；

2.爬虫程序的后期维护工作；

项目描述： 1.网页结构分析，搭建爬虫程序的架构。

2.根据客户需求，负责网页信息抽取、数据清洗工作；编写程序完成网页的爬取；

3.对抓取的数据进行抽取、清洗、去重，并存储到mongodb数据库中；

4.爬虫程序的性能优化。

2016.10 - 2016.10 爬取京东全种类商品(爬虫)

软件环境： Linux

硬件环境： Mac

开发工具： Pycharm

责任描述： 根据公司业务需要,对途牛网的酒店居住,评价,评论数等信息进行爬去,维护爬虫项目,并将获取到的数据进行清洗,存库

项目描述： 1）采用scrapy实现增量爬取。

2）采用分布式框架实现爬虫集群，分布式使用Redis实现。

3）存储Request请求和指纹集合，并且对各个Slave端爬虫实现集中管理和控制。

4）利用Redis的高并发和I/O读写来实现高速下载。

5）通过扩展Redis集群和Slave集群来应对性能瓶颈。

6）采用MongoDB做为本地数据库，将采集到的信息保存到MongoDB中

7）同时每次下载前会检查请求指纹，防止重复下载，避免资源浪费

8 ) 使用第三方代理服务器,采用ADSL拨号定时获取动态ip

2017.05 - 至今 抓取住建委的信息

责任描述： 1.对住建委的网页进行分析；

2.获取每个区域的id、房屋类型的id、最大页数，拼接房源查询的URL；

3.请求房源查询的URL获取每个项目的URL及所有的子URL；

4.最后获取到室号的URL；

5.请求室号的URL获取相应的信息；

6.对抓取的数据进行储存。

项目描述： 该项目主要是对住建委上所有户中室的信息进行抓取，其中包含所在城市、所在版块、项目名称、项目类型、成交时间、供应时间等信息。

2017.02 - 2017.05 抓取搜房网楼盘的信息

责任描述： 1.分析搜房网页面的信息；

2.使用scrapy框架抓取页面信息；

3.利用css解析对页面进行抽取；

4.对数据整理并存储。

5.对代码的维护。

项目描述： 该项目需要抓取搜房网上新房的楼盘首页及楼盘详情，其中包括楼盘名称、咨询电话、主力户型、项目地址、近期开盘、价格等信息。

2016.10 - 2017.01 抓取房天下热点楼盘的信息

责任描述： 1.分析房天下网站URL的跳转；

2.获取房天下每个城市热点楼盘的URL；

3.使用requests请求楼盘的URL，获取源码；

4.设置时间间隔，防止被服务屏蔽；

5.使用bs4、re对源码进行抽取；

6.对数据进行存储。

项目描述： 该项目主要针对房天下的热点楼盘信息进行抓取，其中包含有楼盘名称、具体开盘时间、楼盘平米数及楼盘价格。

2016.07 - 2016.09 抓取链家二手房的信息

责任描述： 1.分析链家网URL的跳转；

2.获取每个二手房的URL；

3.使用requests添加UA、代理ip请求二手房的URL，获取源码；

4.使用bs4、re对源码进行抽取；

5.将数据整理并存储。

项目描述： 该项目需要抓链家二手房楼盘的详情页，主要包含有最新开盘时间、交房时间、供水方式、占地面积等信息。

2017.07 - 2017.08 链家房产—抓取成交数据（个人）

软件环境： Windows7

开发工具： python3、scrapy

责任描述： 爬虫工程师

项目描述： 项目要求：

1. 根据地区爬取二手房成交的相关字段信息

2. 爬取的数据根据要求的格式入到SQL server数据库里

3. 搭建分布式

实现过程：

1. 解决反爬机制

2. 抽取内容

3. 用python3开发，scrapy-redis搭建分布式

2017.05 - 2017.07 美拍—抓取图片、视频、评论（个人项目）

软件环境： linux

开发工具： python2、scrapy

责任描述： 爬虫工程师

项目描述： 项目要求：

1. 根据指定：关键字或热门

2. 爬取视频（标题，播放次数，收藏次数，上传时间，分类）

3. 视频作者（ID，名称）

4. 评论（评论人ID，评论人名称，内容，时间）

5. 搭建分布式系统框架

6. 数据精确（有些网页提取的数据是假的）

7. 实时监控爬虫的进度和报警反馈

实现过程：

1. Fiddler抓包获取评论数据

2. 提取视频url--解密

3. 精确--比较、预判

4. scrapy-redis搭建

5. 定制报警信息

6. 定时爬取，结果发送到邮箱

2016.05 - 2017.06 搭建公司通用爬虫系统

软件环境： linux

开发工具： python、scrapy

责任描述： 分析反爬以及内容的提取

项目描述： 框架设计：

 采用scrapy-redis分布式框架作为爬虫系统的原型，使用Redis作为缓存数据库。

 采用MongoDB做为本地数据库，将抓取到的数据按所属大类、子类以及标题和内容，保存在MongoDB中，多媒体文件存放在本地磁盘。

 利用Redis的高并发和I/O读写来实现高速下载。

 通过扩展Redis集群和Slave集群来应对性能瓶颈。

 同时每次下载前会检查请求指纹，防止重复下载，避免资源浪费。

 定制报警信息。

 实时监控爬虫的进度和警报反馈。

数据提取：

 处理各种反爬以及中间件的编写。

 爬取我买网、淘宝网生鲜板块以及各大电商网站的数据，提供给网站做参考数据。

 由于各大电商网站的数据加载不一样，采用Fiddler抓包；有的是ajax动态加载，规律比较复杂，采用Selenium+PhantomJS实施动态HTML抓取。

2017.03 - 2017.09 京东网图书信息

开发工具： pycharm

责任描述： 负责编写爬虫及中间件模块

项目描述： 这个项目是用Scrapy-redis分布式框架爬取京东网图书的信息：

技术关键词：Scrapy-rendis,Redis, xpath, Selenium, PhantomJs

技术描述：

1，使用Python爬虫的Scrapy-reids框架

2，使用xpath, 正则（re） 进行页面解析

3，使用Selenium ,PhantomJS 处理Cookie,click，JavaScrip

4 , 使用Redis 进行数据储存

5，使用 Scrapy-reids 进行分布式爬取

6，设置实时监督模块，Spider关闭后发送指定邮件进行通知

项目模块：

爬虫模块：爬取需要的数据

管道模块：对数据进行处理，储存

中间件模块：针对反爬虫机制 编写随机User-Agent,代理池等反反爬虫功能

2016.05 - 2016.12 唯品会美妆所有分类

开发工具： pycharm

责任描述： 负责编写爬虫及进度实时监督

项目描述： 这个项目是用Scrapy模块爬取的信息：

技术关键词：Scrapy,Redis, xpath, Selenium, PhantomJs

技术描述：

1，使用Python的Scrapy 爬虫框架进行爬取数据

2，因为页面是滚动条动态加载数据，所以在中间件内使用Selenium发送，不使用Scrapy原本的发送器，接受数据 构造url并返回给Spider继续进行爬取

3，使用xpath, 正则（re） 进行页面解析

4 , 使用Redis 进行数据储存

5，设置实时监督模块，Spider关闭后发送指定邮件进行通知

2015.11 - 2016.03 中华人民共和国国家统计局

开发工具： pycharm

责任描述： 负责编写爬虫及进度实时监督

项目描述： 这个项目是用Requests模块爬取的信息：

技术关键词：Requests,Redis, xpath, Selenium, PhantomJs

技术描述：

1，使用Python爬虫的Requests模块,进行url的发送

2，使用xpath, 正则（re）进行页面解析

3，使用Selenium ,PhantomJS 处理Cookie,click，JavaScrip

4 , 使用Redis 进行数据储存

5，设置实时监督模块，Spider关闭后发送指定邮件进行通知

2016.09 - 2017.05 基于API爬取信息

责任描述： 开发工具：pycharm+vim+liunx

使用模块：requests,time,random,hashlib

基于API接口爬取信息。

通过API接口文档信息。传入请求参数guid，token，timestamp,rand，然后通过哈希MD5，生成秘钥，带着秘钥请求服务器，服务器返回数据，把数据提取，保存起来。

项目描述： 获取API接口文档，仔细审阅接口文档需要的技术。了解APl签名认证的加密方式，获取解密 方式。

2016.09 - 2017.10 基于oauth2.0微信接口抓取数据

责任描述： 使用工具：pycharm,vim,linux

需要用到的技术，通过 docker代价web服务

注册微信订阅号，在阿里云用docker搭建web服务（django，flask），登陆微信公众号后台绑定服务器Ip，token,等信息，实现访问后获取token，，服务器通过测试令牌ID，pwd等 可以拿到想要的信息。

项目描述： 使用工具：pycharm,vim,linux

需要用到的技术，通过 docker代价web服务

注册微信订阅号，在阿里云用docker搭建web服务（django，flask），登陆微信公众号后台绑定服务器Ip，token,等信息，实现访问后获取token，，服务器通过测试令牌ID，pwd等 可以拿到想要的信息。

2016.06 - 2017.04 基于广泛新闻源信息采集

责任描述： 总结：

首先熟悉HTTP等协议和Ajax请求方式，找到请求路径。项目中遇到的问题，请求路径中的数据，是集合套列表，列表套字典，这时候生成器提取不太方便，所以我用正则匹配。

项目描述： 开发工具：pycharm+vim+liunx+xpath

使用模块：requests,lxml,等

打开新闻网站，通过页面分析得出是ajax页面加载，找到js数据中的请求路径，抓取数据，并清洗保存到数据库。

本项目用到的技术xpath，requests,time,正则等。

2016.06 - 2017.05 爬取当当网信息

责任描述： 开发工具 : pycharm.vim。xpath

项目描述 :

使用Xpath方式抓取网页信息，通过F12分析网页的请求和响应，应用请求头信息简单模拟浏览器获取网页信息，提取的信息通过生成器方式返回，

最后储存到MySQL数据库。本项目综合运用了xpath,requests,mysql,完成了项目。并运用了多进程提高爬虫效率

项目描述： 分析页面规律，制定合适的抓取方式，xpath匹配，存入数据库

2017.01 - 2017.02 爬取汽车之家、易车网

责任描述： 负责搭建Python爬虫项目，根据不同网页结构爬取汽车之家，易车网的车型数据，以及定时更新车型库数据。此项目个人独立完成，功能运转正常，可以满足生产需要。

项目描述： 通过对汽车之家和易车网的爬取，对python的使用，以及如何应对反爬虫( 动态IP库 )，如何应对简单验证码校验（图片识别），模拟登录。可以熟悉的掌握并使用Python。对于新接触Python的我来说，有编程语言基础学Python起来也比较快

2015.03 - 2015.04 数据爬去

责任描述： 1. 使用Python的scary爬虫框架；

2. 设置中间键模块：设置ip代理和User-Agent；

3. 在请求头中设置请求头headers信息为浏览器的信息将爬虫伪装成浏览器

4. 使用re，xpath，BeautifulSoup等模块对数据进行处理；

5. 使用mysql数据库对数据进行存储（使用pymysql模块）

项目描述： 当当、博库网、文轩网、天猫、京东商城图书进行爬取，东方财富网股票数据爬取，通过网站的数据能够客观反映市场需求以及变化。

对这些数据进行分析，能够让公司有针对性的进行选题，掌握最新的市场动向，把握先机，并且能够合理

化的安排项目策划的进度以及后期的生产量。

2017.02 - 2017.07 基于scrapy框架商品爬虫

开发工具： Ubuntu、Scrapy框架，Mongdb数据库,Scrapy-redis等

责任描述： 主要负责项目信息的爬取

用到的技术利用Scrapy框架。

根据网页数据在items.py中建模

定义爬虫行为，编写对应spider文件

利用xpath对数据进行抽取。

编写中间件模拟客户端访问，编写User-Agent从中随机抽取用户头。

运用第三方ip代理池，防止ip被封获取不到数据，影响效率。

打开管道，编写pipelines文件。

项目描述： 1.公司电商项目，爬取淘宝，京东等电商网站的信息。

2.将爬取的数据提交个领导，分析，进行项目的完善。

2016.07 - 2016.09 旅游信息爬虫

开发工具： scrapy+MongoDB+fiddle2

责任描述： 采用scrapy-redis分布式框架实现爬虫集群，分布式使用Redis事现。

存储Request请求和指纹集合，并且对各个Slave端爬虫实现集中管理和控制。

利用Redis的高并发和I/O读写来实现高速下载。

采用MongoDB作为本地数据库，将酒店按地区和信息保存在MongoDB中。

同时每次下载前会请求指纹，防止重复下载，避免资源浪费。

项目描述： 这个项目是对各大旅游网信息，进行抓取，采用scrapy-redis分布式实现。Windows 10 做为Master端，管理多个Linux Slave端爬虫集群。分布式使用Redis做为缓存数据库，利用Redis的高并发和I/O读写来实现高速下载。抓取过程中发现票价信息及图片均为JS动态加载的，导致信息抓取不完整，故采用开启下载中间件，利用selenium + PhamtonJS的方式发送请求，待数据加载完毕后返回提取。

2016.10 - 2016.12 大众点评酒店信息爬虫

责任描述： 采用scrapy-redis分布式框架实现爬虫集群

通过扩展Redis集群和Slave集群来应对性能瓶颈。

存储Request请求和指纹集合，并且对各个Slave端爬虫实现集中管理和控制。

利用Redis的高并发和I/O读写来实现高速下载。

采用scrapy-redis分布式框架实现爬虫集群，分布式使用Redis实现。

采用MongoDB作为本地数据库，将酒店按地区和信息保存在MongoDB中。

项目描述： 负责在大众点评抓取酒店信息,包括酒店简介,地址,价格等数据进行获取,然后对数据进行清洗后存入mongoDB中。

2017.01 - 2017.05 36kr创投助手爬虫

责任描述： 1.网页结构分析，采用scrapy框架搭建爬虫程序的架构。

2.编写程序完成网页的爬取，根据客户需求，负责网页信息抽取、数据清洗工作；

3.对抓取的数据进行抽取、清洗、去重，并做持久化存储；

4.爬虫程序的性能优化，及时处理网站升级后反爬问题。

项目描述： 该主要抓取kr36创投助手中的登记的所有公司的注册信息，包括融资动态、项目数据、融资经历及创始团队等，并将数据保存到mongodb数据库中。

2016.09 - 2016.11 o2o物流信息数据抓取

责任描述： 负责模块：

1.主要负责对O2O网页信息的爬取；

2.将爬取到的数据存储到Mongodb库中进行短期的保存；

3.爬虫程序的后期维护工作；

主要职责：

1.分析整个思路，先后顺序，获取相关的url；

2.根据老板的需求，获取相关的数据，进行筛选和爬取；

3.获取响应后保存下来，存储到所需求的数据库中；

4.整合信息，后期进行维护和更新；

项目描述： 通过对本公司的信息抓取采集汇总，从而制定相关的运营制度并开发客户。在信息化时代，网购呈现出一种不断增长的趋势，规模已经达到了空前巨大的地步，这给网购之后的物流带来了沉重的负担，对每一个节点的信息需求也越来越多。每一个环节产生的数据都是海量的，这就需要通过大数据把信息对接起来，将每个节点的数据收集并且整合，通过数据中心分析、处理转化为有价值的信息，从而掌握物流企业的整体运作情况。

2007.03 - 2007.05 联商网行业数据信息的获取

责任描述： 负责模块：

爬取购物中心模块

主要职责：

1.获取相对应的URL；

2.使用session获取登陆后的页面；

3.运用xpath获取相关的数据；

4.使用ajax原理动态掌握信息情况；

项目描述： 通过对其网站上的新零售和购物中心的信息爬取，获取相关的信息，收集整合，从而促进促进公司相关政策的更新。

以及通过随时获取资讯动态里的行业数据的信息，从而了解各行业的物流情况以及物流外包情况，随时掌握物流收入结构的信息，采取相应的措施和策略来应对前景，了解实时状况。

2017.08 - 2017.09 网络爬虫

责任描述： 爬虫整体设计与编码

项目描述： 1.爬取豆瓣近期电影，输出html文件（多线程处理）

2.定时爬取西甲/巴萨足球新闻，球队一年赛程，近期比赛，用mysql存储，在个人网站展示

2012.10 - 2012.10 网页爬虫+百度地图展示

开发工具： MyEclipse+Mysql

责任描述： 和团队一起商量采集数据策略，分工合作，每人负责爬取一个职位的保存在数据库

项目描述： 这个项目是为了分析java，大数据和web方向的热门就业城市而做的一个爬虫网站

使用技术：

◎ 利用httpClient爬取整个网页信息，利用Jsoup对爬取到的数据进行分析

◎ 利用百度地图api将爬取的数据展示在百度地图上

◎ 利用RabbitMQ作为中间件防止爬取数据过程中地宕机问题

2016.05 - 2016.07 行业信息动态爬取

责任描述： 使用python的requests模块以及scrapy框架对不同地区，不同行业动态信息进行爬取，并进行进一步的解析处理，将有效数据存储到mysql中

项目描述： 1. 参与公司爬虫系统的开发与维护

2. 负责分配站点的数据抓取，为公司其它部门提供数据支持

3. 采用requests模块实现爬取，请求目标网页的源代码，多线程请求和解析源代码

4. 在requests发起的请求头中设置请求头headers信息为浏览器的信息将爬虫伪装成浏览器

5. 使用xpath提取节点

2015.11 - 2016.04 派力人人

2017.04 - 2017.08 爬取当当网、博库网、文轩网以及相关图书信息网站数据

责任描述： 1. 使用Python的scary爬虫框架；

2. 设置中间键模块：设置ip代理和User-Agent；

3. 在请求头中设置请求头headers信息为浏览器的信息将爬虫伪装成浏览器

4. 使用re，xpath，BeautifulSoup等模块对数据进行处理；

5. 使用mysql数据库对数据进行存储（使用pymysql模块）

项目描述： 当当、博库网、文轩网、天猫、京东商城图书销量占据线上销售95%以上，且经营网站多年，尤其是当当网站，网站的数据能够客观反映市场需求以及变化。对这些数据进行分析，能够让公司有针对性的进行选题，掌握最新的市场动向，把握先机，并且能够合理化的安排项目策划的进度以及后期的生产量。

2016.07 - 2017.01 基于scrapy框架进信息爬取

软件环境： python2.7、Scrapy框架

硬件环境： win7、Linux

开发工具： pycharm

责任描述： 1.主要负责项目信息的爬取

2.用到的技术利用Scrapy框架。

3.根据网页数据在items.py中建模

4.定义爬虫行为，编写对应spider文件

5.利用xpath对数据进行抽取

6.编写中间件模拟客户端访问，编写User-Agent从中随机抽取用户头。

7.运用第三方ip代理池，防止ip被封获取不到数据，影响效率。

8.打开管道，编写pipelines文件。

项目描述： 1.公司电商项目，爬取淘宝，京东等电商网站数据信息。

2.将爬取的数据提交个领导，分析，进行项目的完善

2017.06 - 2017.06 爬取今日头条信息

软件环境： ubuntu

开发工具： pycharm

责任描述： 独立爬取

项目描述： 1.爬取今日头条街拍内容，在第一级页面抓取标题和内容链接，在顺着链接抓取内容中的图片，并下载到Mongodb中

2.今日头条的信息是使用ajax动态加载的，所以和我之前爬取的猫眼并不 一样主要是在分析ajax加载的数据花了很长时间。

3.爬取数据的时候是使用正则来进行匹配的，所以在匹配数据的时候 花了很大的功夫。

4.在进行数据清洗的时候遇到了问题，因为之前对数据清洗这一块并没有 接触太多，提取到的数据并不是标准的json格式的数据，之前对 json.loads() 和json.dumps()作用模糊不清，最后查了很多资料才搞定，之后的问题就迎刃而解了