杨开群

求职意向：python 爬虫工程师

基本信息

出生：1997.06（26 岁）

民族：汉

性别：男

群面：共青团员

电话：17385722018

邮箱：lzj155@foxmail.com

籍贯：贵州遵义

住址：附近

学历：大专

教育背景

遵义师范学院

电气自动化

2021. 3. 10–2023. 7. 10

大专

重庆互联网学院

华信智原

Python 技术开发

2020.12 – 2021-12

培训内容：

HTML5 & CSS3 & JavaScript

Python 爬虫 & Scrapy 框架

MySQL & Redis

数据分析 & 机器学习

网络编程 网站搭建 服务器部署

Linux & 渗透测试

正则表达式 & 数据采集工具....

工作经历

2021.6 – 至今 爬虫工程师

公司：重庆衡科大数据科技有限公司

职位描述：

1. 负责多平台信息获取过程中遇到的验证类反爬机制（验证码）进行算法&平台打码。

2. 解决各类技术疑难问题，包括网络问题、网站搭建、服务器配置、app 分析、反反爬虫等。

3. 负责设计和开发分布式的网络爬虫，以及策略持续优化。

项目经验

2020. 05–至今

- 目标站点：
  汽车之家 全站口碑数据（https://www. autohome. com. cn/）
  数据量 200w +80 多个字段。
- 目标站点：
  某品牌汽车项目（https://www. autohome. com. cn/）
  汽车之家 易车网 太平洋汽车 懂车帝 爱卡汽车 全站口碑点评数据爬取 数据清洗 入库 数据量 80w+。
- 目标站点：
  赛盈分销平台(跨境电商网)（https://www. saleyee. cn/）
  全站商品信息(**)爬取 入库。
- 目标站点：
  中国政府采购网（http://www. ccgp. gov. cn/cggg/dfgg/gkzb/）
  实现程序，提供关键字，自动搜索仅限当日发布的招标公告 并弹窗提示。
- 目标站点：
  北京科技园拍卖招标有限公司（http://www. bkpmzb. com/bidding/）
  提供关键字，自动搜索仅限当日发布的招标公告 并弹窗提示。
- 目标站点：
  得捷电子网（http://www. bkpmzb. com/bidding/）
  实现程序，指定货号后，每日定时（周一到周五 22：00 开始放货），自动填写数量，生成订单发送至指定邮箱。
- 目标站点：
  小程序 – CiE 美妆创新展
  根据商展列表，排名从高到低，依次访问详情页，获取展商信息。数据保存到 csv 格式的 Excel 文件中。
- 目标站点：
  点点数据（https://app. diandian. com/）
  提供关键字，获取评分趋势-新增评价趋势图中的数据。数据保存到 Excel 文件。提供代码。
- 目标站点：
  ANNUAL REVIEWS（https://www. annualreviews. org/）
  输入关键字，爬取标题、作者、发布时间、大纲等…对应的内容以及 PDF 下载链接。数据排版写入 TXT 文件，并提供代码。

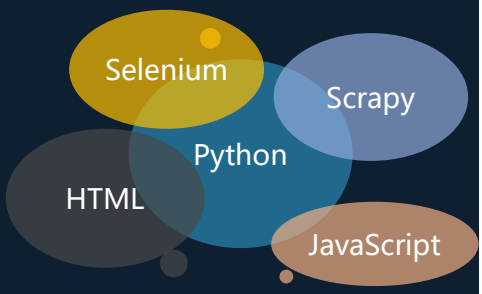
所获证书

- Python 高级技术开发
- CEAC 数据分析师
- CEAC 大数据挖掘工程师

技能掌握

Python	80%
Redis	75%
JavaScript	50%
MySQL	70%
HTML	80%
jQuery	49%
Scrapy	60%

个人特长



兴趣爱好

阅读　旅行　骑行　游泳

- 目标站点：
  中国奥委会官方网站（http://www.olympic.cn/）
  定时爬取今日要闻，实时动态更新数据（时间差在 5s 内）。写入服务器内 Redis 数据库保存，并指导相关人员调用数据发布至指定微信小程序。
- 目标站点：
  沃尔玛官网（https://www.walmart.com/）
  提供关键字，监控 产品价格升、降价提醒，并编辑信息发送至指定邮箱。
- 目标站点：
  TikTok（抖音海外版 APP）
  通过搜索话题 爬取每个话题底下热度最高的 40 个视频信息。
  字段包括点赞转发评论数量，还有用户名。
- 目标站点：
  中国知网（https://www.cnki.net/）
  根据关键字，搜索发表的论文，爬取其被引文献直到 2021 年，数据包括，发表时间，期刊名称，期刊影响因子，被引量，下载量。数据入库 Redis。
- 目标站点：
  MINDMAP（https://www.azuki.com/mindmap）
  1:1 仿目标站点网页，加入功能：同时播放音乐 切换其他版面的时候音乐还在放
- 目标站点：
  U.S. DEPARTMENT OF STATE
  （https://fam.state.gov/Fam/FAM.aspx?ID=05FAM）
  爬取 FAM 和 FAH 分类下所有子目录内容。数据包含标题以及内容，入库 Redis。
- 目标站点：
  拓者设计吧（https://www.tuozhe8.com/forum-122-1.html）
  根据分类，采集整站图片，图片根据分类-编号命名。数据保存本地目录。
- 目标站点：
  伟海精英-证券从业人数（https://www.weihai.com.cn/data/person）
  采集 12 年到 18 年合计、一般证券从业、保荐代表。数据保存本地 csv 文件。
- 目标站点：
  LLINOIS STATE POLICE firearm Services Bureau
  （https://www.ispfsb.com/Public/Home.aspx）
  解决此网站的 ReCaptchaPress 谷歌验证码。
- 目标站点：
  豆瓣电影（https://search.douban.com/movie）
  根据关键字，采集豆瓣电影搜索后所有内容。关键字以及详情、链接保存 csv 文件。

掌握技能

熟悉 Python、HTML、CSS、JavaScript、MySQL、redis、正则表达式、jQuery 从结构化和非结构化数据中解析数据

熟悉网页的结构特点及规律、以及网页、APP 抓取原理及技术

熟悉 Tableau、Excel 分析工具，js 逆向..

有短视频、评论、小说、图片、商品价格、客户端、APP、抓取经验。

自我评价

我对计算机有着十分浓厚的兴趣。

能熟练使用 fiddler,Hbuiderx,pycharm,charles,python 环境、八爪鱼等工具、

经过互联网，我不仅仅学到了很多在日常生活中学不到的东西，并且坐在电脑前轻点鼠标就能尽晓天下事的欢乐更是别的任何活动所不及的。

其他信息

部分项目及源码已上传至博客和 GitHub..

个人主页：demo443.com　demo520.com