

基本信息

- 李勃旺-男-1996/10
- 天津工业大学(双一流一本) 复合材料科学与工程&工学学士 自学编程
- TEL: 18990284571
- MAIL: 821346679@qq.com & qq8212346679@gmail.com

求职意向

- 工作地点: 北京
- 期待职位: Python爬虫开发工程师

个人技能

- 掌握 Python 中的闭包, 装饰器, 生成器, 多线程, 多进程, 协程等并能灵活运用以解决编程问题
- 对 Pandas, Numpy, Matplotlib 等数据分析库有初步的掌握
- 熟练掌握 Re, Requests, Selenium 等爬虫基本库的使用方式
- 熟练掌握 Xpath 选择器, Css 选择器, 正则表达式 等清洗数据提取内容
- 熟练掌握 Scrapy 爬虫框架的运作原理及各个部件的功能, 掌握 Scrapy-Redis 分布式爬虫的原理
- 对网页加密数据有一定的 Js逆向 能力
- 掌握 封Ip, 验证码, Js加密, Cookie实效性 等各种反爬措施以及其应对措施
- 熟练使用 Mysql, MongoDB, Redis 的增删改查功能
- 掌握手机端爬虫爬取方式, 熟练使用 Fiddler, Mitmproxy 等抓包工具
- 掌握 Appium 手机端自动化爬虫框架
- 熟悉 Mysql 索引原理, 能手写高级的 SQL 语句
- 了解 Web 前端, 可读懂并书写简单的 js/css/html 等代码
- 能使用 Flask, Django 生成爬虫 API 供给后端调用或搭建简单的网页
- 熟悉 Linux 开发, 平时使用 Vscode 的 远程开发 功能直接在 Linux 平台开发
- 有一定 算法基础, 熟悉常用的 数据结构 与 算法, 如队列、栈、链表、二叉树、各类排序等
- 熟悉常用的网络协议如 TCP/IP 协议、HTTP(S)协议, 知道各协议的区别与作用以及网络通讯的过程
- 熟练使用 GitHub 管理代码, 参与过 GitHub 开源项目的代码贡献
- 简单掌握 Java, 能读懂 Java 代码
- 有良好的自学能力和英语能力, 能阅读英文原文官方文档并使用 Google, Stack Overflow 进行问题搜索
- 有代码洁癖, 经常为了优化代码而花费大量时间, 但是也学到了许多更高级更巧妙的办法, 代码规范整洁

工作经历

— 天津尚仪企业咨询管理有限公司

- 职位: 爬虫工程师
- 工作时间: 2019/06/10 --- 至今
- 工作内容:
 - 负责公司的数据库维护
 - 负责公司所有的数据更新, 维护和提取等工作
 - 负责公司所有的爬虫项目架构设计优化, 维护使其稳定运行
- 取得成果:
 - 把上一任留的 Java 项目改成 Python 实现
 - 通过优化代码逻辑, 使每日更新数据时间由原来的近 2小时 缩短至 5分钟 左右
 - 提供公司后端业务需要的多种爬虫接口

项目经验

— 西青区专利数据项目(长期维护)

- 技术栈: Requests + 模拟登录 + Json + Mysql + Pandas + 多线程
- 项目描述: 模拟登录目标网站, 然后通过多线程请求相关数据, 通过 Json 解析爬取到的数据并用 Pandas 分析然后生成相应的 Excel 文件, 也可以存入数据库

— 每日更新专利管家的专利相关数据项目(长期维护)

- 技术栈: Selenium + 验证码模拟登录 + Mysql + Requests + 多线程
- 项目描述: 通过 Selenium + 验证码识别模拟登录目标网站, 下载对应的 Excel 文件, 然后通过分析文件并构造另一个网站的接口请求数据更新专利的增量信息

— 天津市各个政务平台信息汇总的爬虫项目(长期维护)

- 技术栈: Mysql + Xpath + Requests + 多线程 + Django + 前端
- 项目描述: 爬取需要监视的多个网站的最新公告, 每日更新到数据库, 并通过 Django 部署到线上向公司内部展示

— 给后端提供的API (长期维护)

- 技术栈: Mysql + Redis + Selenium + Js逆向 + Flask + 分布式 + 验证码模拟登录
- 项目描述:
 - 专利信息查询API: 使用 Js逆向 + 其他技术爬取目标网站信息, 并用 Flask 部署到公司服务器上, 作为接口返回给后端, 通过专利号就可以查询到专利的所有信息
 - 企业名称查询专利API: 通过模拟登录+ 其他技术爬取目标网站信息, 通过企业名称查询企业拥有的所有专利信息
 - 商标信息查询API: 通过 Selenium + 其他技术爬取目标网站信息, 通过商标号能查询商标的所有信息