黄新

15311886138
1731893469@qq.com
https://github.com/Seriainme
工作经验: 3年
职位: 爬虫工程师

教育背景

安徽财经大学 金融学 (本科) 2014.09.01 ~ 2018.06.30

个人作品

个人博客 访问链接

用GitHub page搭建，记录了平时的学习成果

自我介绍

零基础自学转码，自学能力出色，热衷学习新技术。有责任心，做事认真踏实，抗压能力强。
曾经在外贸行业工作过，英语沟通无障碍可作为工作语言
自驱型爬虫工程师，三年爬虫开发经验。
负责公司爬虫框架的开发和维护，对逆向工程有深入了解和实战经验。

个人技能

- web逆向，安卓逆向 ★★★★★
- 爬虫开发 ★★★★☆
- Python，sql ★★★★☆

工作经历

拓尔思信息技术有限公司 数家产品部 2022.01 至今

逆向工程师

负责处理舆情大数据采集开发中的逆向工作，以app端为主也有一些网页站点的采集。
同时也负责头部境外媒体的数据采集程序开发

药渡经纬信息科技公司 源数据中心 2020.06 ~ 2021.12

后端工程师

负责采集医疗大数据数据的爬虫程序的开发维护，以及数据清洗等后端工作。

其他工作经历 2019.01 ~ 2020.06

2019.12-2020.02 中国平安 保险代理人 销售区域客户的挖掘以及公司产品的销售

2019.8-2019.10 北京21世纪房地产（阳光上东店） 对外租赁 开发在京的外国人租房市场以及维护客户关系

2019.2-2019.6 北京中诚致信网络科技有限公司 金融电销 电话销售贷款等金融服务

项目经历

逆向工程以及爬虫框架维护 2022.01 - 至今

核心开发者

项目描述：在公司采集工具"雷达"的采集基础上，处理有加密无法采集的app客户端和站点。破解站点加密逻辑后，根据其域名为爬虫框架编写对应的配置文件，供雷达采集调用。

主要工作：熟练掌握app抓包的技巧，知道如何绕过检测代理的问题。

熟悉frida的常规hook操作，以及对app的脱壳处理 ，frida-dexdump,blackdex,apktool 都有使用过。

会使用jadx分析java源码，也会用ida来分析native层源码并用unidbg调用。

熟练使用油猴脚本对web端的加密进行hook。熟悉一般补环境的流程，有处理过瑞数，抖音头条等网站的经验，会使用协议过验证码。

新华社深度数据采集 2022.10 - 2023.03

独立开发者

项目描述：采集新华社各个国家脸书和INS的贴文和用户数据

主要工作：独立设计了爬虫框架完成定时采集。
使用代码操作浏览器来采集下载相关的Excel，并用pandas去解析数据，同时使用xpath解析页面的数据。使用了selenium-wire库来获取页面所加载的xhr请求。

采集新闻app的推送消息 2022.07 - 2022.08

独立开发者

项目描述：评估了主要的工作流程，根据app的推送消息是否能抓包将他们分成两类。对于能抓包走http协议的app，进行抓包和相关的破解加密参数处理。对于不能抓包的，进行了多次hook尝试并总结除了方法来hook到对应类所推送的消息。

主要工作：对app脱壳分析源码，对于有推送开关设置的app，找到设置页面的对象，分析即可知道推送sdk的种类。继续使用frida去溯源追踪，结合网络找到资料，就可以定位到推送内容相关的代码部分。

对于没有推送开关设置的app，在源码里搜索notificationopen 或者openNotification等关键字，也可以定位到推送代码的位置。

在上述测试中，我还使用了objection来hook类的所有方法。定位结束后，开发了相关的Xposed插件来对app推送就行采集和解析入库。

微信公众号文章采集 2020.05 - 2020.07

独立开发者

项目描述：根据需求采集指定微信公众号的发文。前期通过购买第三方数据的方式，编写代码从接口取数据，实现每日定时采集并去重更新到数据库。后期自己在GitHub寻找开源项目，进行代码优化和改进，实现了自主采集。

主要工作：阅读了开源项目的代码和文档（https://github.com/lixi5338619/weixin-spider），知道了其主要的工作原理，在服务器布置了相关环境并对采集到的数据进行解析清洗入库。

代码主要的原理为：对于每个公众号，在微信浏览器使用pyautogui库来模拟点击url，使用mitmproxy去截获相关的请求。拿到其中有时效性的token之后，再根据公众号的biz和其他参数去api请求，获得该公众号的发文。

大数据关键词标定 2020.04

独立开发者

项目描述：将所有的文本和关键词标定，但是由于量级过大笛卡尔积达到几十亿，所以匹配效率低下，使用正则或者sql去做都很慢。

主要工作：通过自己寻找资料，多次实验最终选定使用ac automation 算法来对数据进行匹配，并对原来的算法进行了多进程的优化。将原来需要两个星期才能完成的标定过程压缩到只需要两天。具体代码逻辑在博客