

陈**

1801552**** xtchen.pitt@gmail.com <https://github.com/worldwonderer>
北京市朝阳区 后端工程师 在职

教育背景

2014.9-2018.6 大连理工大学 工商管理 本科

技能评价

- ◇ 熟悉 Python，能写出 Pythonic 的代码。熟练使用 Shell 处理日常需求
- ◇ 熟悉 App 抓包，Xposed/Frida Hook，熟悉 App 脱壳、逆向 so，使用 IDA 等工具进行静、动态分析
- ◇ 熟悉 Flask 网页框架和 Celery 任务调度框架的使用，编写过复杂的调度模块
- ◇ 熟悉根据业务需求设计数据库表结构。熟悉 Redis/Kafka/MySQL/MongoDB
- ◇ 有独立设计并落地大型爬虫系统的经历

工作经历

2017.5-2019.4 ** 技术平台部 数据研发工程师

- ◇ 担任长短视频、PGC 视频和机器剪辑视频业务线的爬虫和收录工作
- ◇ 完成视频处理流程调度的重构，提高视频处理的透明度和稳定性
- ◇ 负责机器剪辑视频业务表字段设计、视频处理调度设计和相关代码编写

2019.4-至今 ** 内容研发部 后端工程师

- ◇ 负责爬虫整体架构和基础组件的设计，主导外站自媒体库，全网内容池等项目
- ◇ 攻坚高反爬强度的网站、App，快速给出应对方案。如推送监控、某信公众号、某条、某小红书、某音等
- ◇ 从 0 到 1 搭建反反爬基础模块，验证码识别、Android 设备群控、浏览器集群，优化代理池和 Cookie 池服务

项目经历

2018.9-2018.12 机器剪辑视频 核心成员

- ◇ 独立负责机器剪辑项目中的爬取和流程设计部分，完成机器自动生产视频和辅助人工制作视频
- ◇ 设计视频信息存储和视频的处理流程，包括长视频精彩看点、横屏短视频剪竖屏、短视频合成等类型，接入视频处理平台，日均生产 1000 余条视频
- ◇ 负责突发事件剪辑平台项目的表设计、爬虫及后端编写，实时/离线爬取热点话题及其文字、图片和视频素材，实现辅助人工制作视频、文本转视频、自动生产热点话题视频

2019.12-至今 全网内容池 Owner

- ◇ 爬取高日活平台、垂类 TOP、传统新闻门户在内的 30 多个站点，提供海量内容，每日入库量 5000 万+，主流站点自媒体作者覆盖达 95%
- ◇ 基于 Celery 和 requests 自研爬虫脚手架，落地单元测试和监控。将爬虫所需的资源生产/获取部份抽离，形成独立的代理 IP 模块、Cookie 模块、验证码识别、手机群控和浏览器集群，并抽象出对外的接口层，实现反反爬基础模块和爬虫逻辑的高度解耦，大幅提高开发效率和可维护性
- ◇ 拆解传统爬虫为细分任务和组合链路，细分任务为最小粒度的爬取目标，如获取某文章的评论，组合链路将细分任务灵活地拼接，满足具体的爬取需求，增加可复用性
- ◇ 利用各类站外指标如粉丝数、点赞数等，制定自媒体作者以及内容的评分机制，对优质内容挖掘有深入的理解

2019.11-至今 Android 设备群控 Owner

- ◇ 基于 Frida、Auto.js 和 STF 二次开发，支持远程控制，批量自动化操作和 Hook，提供加解密算法等功能
- ◇ 分布式部署 STF 设备管理平台，共操纵 30 余台手机，实现了手机网络和进程监控，异常重启和自动重连
- ◇ 通过 websocket 实现服务端与手机 Auto.js 的通信，支持批量下发和运行模拟操作脚本和捕获日志
- ◇ 支持 Frida 批量执行注入，通过 hook 加解密函数提供 RPC 调用，hook 内容入口函数截获内容
- ◇ 研究过主流风控的实现，并针对风控的检测项实现修改设备参数

2019.12-2020-3 浏览器集群 核心成员

- ◇ 支持 worker 节点监控，渲染任务状态记录，缓存和负载均衡等特性
- ◇ 修改 chromium 源码的字体、插件、webkit、canvas 等接口，实现属性随机，规避浏览器指纹的检测
- ◇ 基于 cef 二次开发，在更底层操纵 chromium，绕过自动化测试工具带来的特征暴露问题

个人评价

- ◇ 善于通过沟通进行需求拆解，结果导向解决工作中各类难题
- ◇ 良好的自我驱动力，乐于钻研前沿技术，不断强化自身软硬技能。所带团队 8 人，氛围积极活泼，重视团队共同成长和整体贡献