黄一天

- **概况**：`2-3年工作经验` | 男 | 1990年11月 | 硕士 | 计算机科学
- **电话**：`+1-(703)-395-3877` | *目前人在北美，可视频或电话面试，六月底回国*
- **邮箱**：`wfgydbu@163.com` | **微信**：`wfgydbu` (2020)
- **技术博客**：Ethan's Journal | **Github**：https://github.com/wfgydbu
- **期望岗位**：`Python爬虫工程师` / `Python工程师` | **期望城市**：`上海市`

教育经历

- **乔治华盛顿大学** | 201年9月 - 2017年5月 | 计算机科学 | 硕士 | GPA：3.97/4.0 | 专业方向：操作系统、网络安全和数据库
- **南京工业大学** | 2008年9月 - 2012年6月 | 计算机科学与技术 | 学士

工作经历

- **Acuty LLC** | *软件工程师（实习）* | 2017年10月 - 至今
- **天加能源数据管理有限公司** | *软件工程师* | 2013年7月 – 2014年4月
- **中软国际资源服务有限公司南京分公司** | *测试工程师* | 2012年3月 - 2013年5月

项目经历

Python爬虫项目 | 所有代码均已上传到Github

- **对kuku漫画进行分布式全站爬取**，主要使用scrapy和scrapy_redis，消息队列和结果均保存在redis数据库；9个实例（本机8个+云主机1个），每个实例16个并发数，处理请求时峰值大约为27W/H。
- **使用pyspider对kuku漫画进行全站爬取**，单实例，15个并发，处理峰值大约为7W/H。
- **使用celery对空气质量指数进行分布式爬取**，爬取acicn.org上2500左右个监测点的空气质量指数，在两台虚拟机上部署，20条协程并发，大概160s跑完一次爬取任务；构建定时任务定时爬取，绘制空气质量指数趋势图。
- **模拟登陆新浪微博并抓取关注页面和个人信息**，模拟登陆新浪微博（计算加密后的用户名和密码），获取成功登陆的Cookies；从登陆用户开始，递归爬取每个用户的关注列表和个人信息页面；使用Bloom Filter和索引双重保障保证不会有重复的数据；预留接口便于扩展爬取其他页面；对结果进行分析。
- **模拟登陆淘宝并抓取所有历史订单**，模拟用户输入账号密码和滑块拖动 (selenium, PhantomJS) 或扫描二维码登陆淘宝并获取登陆Cookies；登陆成功后，抓取所有历史订单并展示结果(PrettyTable)。
- **其他验证码处理相关**， 涉及到一些图片处理或识别的算法：灰度二值化、连通域、字符分割、倾斜校正和样本训练等和库：tesserocr, numpy, PIL, matplotlib和cv2等
- 其他一些爬虫项目，通常都使用了requests、bs4和re库，项目包括爬取**空气质量指数 (aiohttp, asyncio) 、天气信息 (sqlite3, ast) 、12306余票和票价 (json, sqlite3, colorama) 、斗鱼视频下载、搜狗微信文章(tomd)、百度贴吧、虾米音乐 (json, exejs) 、网易云音乐热评 (json) 、今日头条、猫眼电影榜单、糗事百科。**

Python项目 - ohHTMLToMarkdown库 | 代码已上传Github | 可以从PyPI上安装，搜索`ohHTMLToMarkdown`

- 一个HTML转Markdown的Python库，将一部分或整个HTML页面转化为Markdown文档。
- 利用bs4将HTML文档解析成块状结构，然后通过自定义方法对HTML标签进行处理，三种选择：转化成md语义、进一步递归解析或直接舍弃，目前支持几乎所有主流HTML标签。转化效果基本满足需求，提供基本的测试代码，测试对象包括简书的文章、知乎的回答和我博客中发布过的文章。

Python项目 - ohIPPool代理IP池 | 代码已上传Github

- 一个基于Python、Redis和Flask的动态代理IP池。
- 提供配置项，可以轻易按需求调整并部署到服务器；通过接口或爬虫大量获得免费代理IP，使用Redis的有序集合对IP进行存储和管理，采用计分机制保证池中IP的质量，定期清理无效IP；对外开放一个Web接口服务，通过简单的GET请求就可以获得一个有效且高质量的代理IP，可供任何语言代码直接使用（只需该代码支持HTTP协议）。

其他 - 轻量级线程库 | 代码以上传Github

- 使用*C语言和汇编*实现了一个面向x86-32处理器的用户级线程库。提供线程创建/跳转/阻塞/结束的接口；线程的调度； 通过通道(channel)或集(group)实现线程同步/异步通信；最后将其移植到CompositeOS并支持操作系统的抢占式调度。。

其他 - TIME任务管理系统 | 代码已上传Github | Demo

- 实现了一个面向个人的任务追踪管理系统。基于Wordpress开发，前端使用*CSS*，*HTML*，*Javascript*，后端使用*PHP*，数据库使用*MySQL*，最终以插件形式发布。系统功能包括任务和项目进度追踪、统计报表生成（各种图、表）、数据导入导出等。

工作期间项目

ESM能源管理系统

2013年7月 - 2014年4月 | 天加能源数据管理有限公司 | 10人开发团队 | 职位：软件工程师 | 部门：云计算中心 | 行业：互联网/能源 | 语言：C/C++, SQL

- 实现一个能源管理系统，功能主要包括对定点能源消耗的实时监控和定期生成能源报告，报告交由能源专家进行分析以发掘节能潜力。
- 负责：
  - 设计和开发数据处理模块和DBA模块，前者按照自定协议接收来自工控设备的能源数据（水、电、温度湿度等），处理之后传输给DBA模块；后者在接受到数据后写入到数据库。
  - 部署并维护了一个MySQL集群（2主1备），前端部署负载均衡模块，对外提供数据服务。集群主要用来存储系统采集到的数据，同时向Web界面提供数据；
  - 设计并编写表结构、视图和存储过程的SQL脚本。

WISG-SCG产品线测试

2012年3月 - 2013年5月 | 中软国际资源服务有限公司南京分公司 | 项目组约100人，属于10人测试小组 | 职位：测试工程师 | 部门：Consumer | 行业：通信 | 语言&工具：shell，自研测试工具

- 外包到华为南研所移动宽带价值增长业务解决方案下的SCG项目组，项目主要为通信运营商提供第三方的计费及各类通信增值服务。
- 负责：
  - 几个特定版本的计费功能测试：在实验环境(suse 11)下手动执行测试用例；
  - 几个特定版本的自动化测试：使用华为自研的自动化测试工具执行和维护各版本自动化测试用例，维护各版本的自动化测试环境；
  - 负责几个特定版本的性能测试，部署镜像环境进行功能和性能测试。 支持海外工程师的局点项目部署。

语言水平

英语 CET-6 425|TOELF 92|留学经历 日常英语交流，擅长英语阅读和写作。《Adventures in Minecraft》中文译者之一，《Minecraft Modding For Kids For Dummies》和《Ruby for Kids For Dummies》独立中文译者，三本书均已由中国邮电出版社出版。

自我描述

- 有较丰富的编程经验，熟悉多种编程语言，目前最擅长的语言是C和Python。语言只是工具，用最合适的技术解决问题才是一名工程师应该做的事。
- 有扎实的计算机知识体系，热衷于操作系统、网络和安全三个方面。希望能够融汇贯通，学以致用。
- 很强的自学能力，能很快的上手之前没有接触过技术，遇到问题能迅速从互联网上定位到需要的信息并构建初步的解决方案。同时善于总结，博客上共计发布技术文章100余篇以及一部关于密码学的WIKI。
- 敏捷开发理念的追随者。

技能（排名不分先后）：

- **爬虫相关**：scrapy, scrapy_redis, pyspider, bloomfilter，requests, bs4, celery, re, json,等
- **语言**：Python 3, C/C++, HTML5, CSS, JavaScript, PHP，Shell
- **系统**：Windows, Linux, Amazon Web Service/EC2
- **数据库**：MySQL, MySQL Cluster, Oracle, sqlite3, redis