# 谭浩宇

166-1991-3869 | thy_0417@163.com | 北京 昌平

25岁 | 中共党员

在职 | 求职意向：python爬虫工程师 | 期望薪资：15k-25k

## 教育经历

**华东交通大学** 南昌

软件工程 本科 2012年9月 - 2016年6月

**伊尔库茨克国立交通大学** 俄罗斯

交换生 2014年3月 - 2014年7月

## 工作经历

**北京传智播客教育科技有限公司** 北京

修正校区 python助教 2018年3月 - 至今

北京传智播客教育科技有限公司：是一家致力于高素质软件开发人才培养的新三板挂牌公司，旗下已涵盖黑马程序员及博学谷等五大子品牌，目前已在19个地区成立分校并开设13门课程。

- 负责python23期同学的技术辅导工作，在最近一次教师评分中拿到了99.25分的最高分；
- 负责完成对各大招聘网站进行数据抓取，为同学提供就业信息。

**北京健康有益科技有限公司** 北京

AI中心 爬虫工程师 2018年1月 - 2018年3月

北京健康有益科技有限公司： 一家探索健康管理、健康产品消费升级领域，通过人工智能、生命科学等技术在健康管理应用层的研究及实践，为有健康需求的人群提供健康管理、健康产品及健康服务的企业。

- 隶属于AI中心，负责对数据进行抓取以此来为后续的图像识别（一图一物、一图多物）、语音识别、知识图谱的本体构建等人工智能工作提供数据。目前已完成的爬虫抓取项目为堆糖网图片抓取项目，中国商品信息服务平台抓取项目；
- 进行机器学习NLP工程师的转岗工作，使用RNN的LSTM算法负责了一个体重预测的项目。

**中交三航局交建分公司** 上海

技术开发部 软件工程师 2016年7月 - 2017年11月

中国交通建设集团： 中国最大的港口设计及建设企业，核心产品包括基础设施设计和建设业，以基建疏浚和环保疏浚为主的疏浚业，2017年世界五百强103位，拥有34个全资子公司和7万多名员工。

- 完成数据提取、解析、过滤、入库和监控等研发和优化工作；
- 探寻网页反爬虫的特点规律，参与设计反爬虫来提升平台的抓取效率；
- 爬取全国材料信息价格项目，实时为每个项目部以及招投标提供材料设备采购指导价；
- 负责BIM项目的搭建，完成了BIM模型，并撰写了相关技术论文。

**南昌麦虹网络科技有限公司** 南昌

技术部 实习 2015年6月 - 2016年6月

南昌麦虹网络科技有限公司：整合南昌市内的超市资源，提供线上线下配送服务，目标客户是社区和学校的用户。

- 对网页、链接进行特征分析； 对酒水相关网站进行酒水种类数据爬取；
- 负责即将上线商品信息的二次校对和录入，并保证上线商品的错误率不超过1%；
- 负责保证后期数据的准确性以及整体流程的更新和迭代。

## 技能

- 了解常规验证码的处理方式，了解滑动验证码的处理思路；
- 了解并能够处理常见的反爬方式；
- 能够使用Fiddler进行App端抓包；
- 掌握HTTP/HTTPS协议、TCP/IP网络协议等；
- 了解前端知识，会HTML、CSS、JS的常规使用；
- 掌握爬虫技巧，会使用scrapy框架、urllib、urllib2、Requests等，Selenium+Chrome/Firefox/PhantomJS进行动态数据抓取，并使用正则表达式、XPath、BeautifulSoup从结构化和非结构化数据中进行信息提取；
- 熟悉常用数据库MySQL、Redis以及MongoDB的常规使用；
- 了解机器学习的常规知识，了解PaddlePaddle、TensorFlow框架。

## 项目经历

**招聘信息抓取**

项目描述：对多个招聘网站，如拉勾网、腾讯社招、智联招聘、BOSS直聘进行数据抓取，并可以通过关键字查询爬取到相关的数据，最后数据保存为json数据，之后教学中使用再转换为CSV文件。

关键字：Requests、urllib、json、XPath、Fiddler、csv

技术简介：

1. 用urllib.urlencode将POST请求参数进行封装转换；
2. 使用requests模块进行POST访问；
3. json.dump()将数据转换后存储到磁盘；
4. 读取文件，json.load()将数据转为python类型；
5. 创建一个csv文件读写对象，获取所有的数据部分，最后关闭文件。

**体重预测模型的搭建**

关键字：TensorFlow,PaddlePaddle,Pandas

1. 使用Numpy，Pandas对数据集进行预处理，完成缺失值的填充和处理；
2. 分析特征值并保留，无关因素去除形成新的完整数据集；
3. paddle.layer建立LSTM模型；
4. 构造使用随机梯度下降的trainer并设置优化算法，采用AdaGrad算法；
5. 使用trainer.train进行模型训练并不断优化参数；
6. 使用paddle.infer进行预测。

**中国商品信息服务平台数据采集**

项目描述：根据需求文档提供的关键字，采集中国商品信息服务平台该关键字的相关数据

关键字：urllib, Selenium+Chromedriver, time, random, csv

1. 配置chrome浏览器无用户名密码代理并创建浏览器对象；
2. 使用driver.find_element_by_xpath完成第一次模拟操作，进入人机身份验证；
3. 人工点击身份验证，完成第二次模拟操作，并进行反爬设置等待时间为10s（经测试）；

4. 留接口用来获取汉字并添加到搜索框内，完成第三次模拟操作；

5. 爬取数据并存为csv格式。

**堆糖网图片采集**

项目描述：采集出现食物的场景图片，图片用于训练一图一物和一图多物模型

关键字：Requests，urllib, threading

1. 使用requests模块进行数据抓取；

2. 实现入口、url处理、发出请求、下载并处理图片的函数功能；

3. 其中处理url函数中利用urllib模块实现中文转为url编码，并且提取一个页面的所有图片链接；

4. 设置最大线程锁，实现多线程爬取下载图片，之后对图片进行编号存储。

**采购材料设备的信息采集**

项目描述：爬取造价平台上各省份每月材料价格

关键字：Scrapy，XPath，re，Redis，MongoDB，User-Agent

技术简介：

1. 使用scrapy框架抓取网页，并采用scrapy-redis分布式策略；

2. 搭建分布式爬虫环境，核心服务器为master，跑爬虫程序的三台机器为slave；

3. 在master上搭建一个redis数据库用作url的存储，通过设置slave上scrapy-redis获取url的地址为master地址，以此保证有多个slave获取url的地方只有一个；

4. 使用XPath（lxml），正则进行页面分析并获取数据；

5. slave从master的redis中取出待抓取的request，网页下载完之后就把内容发送回master的redis；

6. 各个slave在完成抓取任务之后，把获取的结果存储到MongoDB上。

模块介绍：

1. spider模块：用来定义特定网站的抓取和解析规则。

2. 中间件模块：设置代理和User-Agent。

3. 管道模块：检查是否是重复数据，如果重复就删除；数据库交互并保存数据。

**BIM技术推广新闻信息采集**

项目描述：对BIM技术新闻信息（中国BIM论坛，BIM中国网）进行提取

关键字：Scrapy，XPath，re，User-Agent，MongoDB

1. 使用Python的爬虫框架scrapy框架，开启多线程爬取；

2. 获取URL请求，当请求返回后调取一个回调函数。第一个请求是通过调用start_requests()方法。从start_urls中的Url中生成请求，并执行解析来调用回调函数；

3. 在回调函数中使用XPath（lxml），正则解析网站的内容；

4. 使用MongoDB进行信息存储。

## 其他

- **证书/执照：** 机动车驾驶证（C1）、 CAD绘图师资格证书
- **语言：** 英语（CET-4），俄语（简单交流）

- **兴趣爱好：** 架子鼓，手风琴

## 自我评价

- 快速学习能力，喜欢新技术，乐于从事有挑战性的工作
- 善于在工作中发现问题，解决问题，有较强的分析能力，并进行独立思考
- 乐于与用户以及同事和领导进行沟通，有一定的组织领导能力