

**《Python-某招聘网爬虫项目实训》**

**课程设计报告**

**---信盈达联合指导**

|  |  |
| --- | --- |
| 二级学院 | 计算机与信息工程学部 |
| 课题名称 | 某招聘网爬虫项目实训 |
| 学生姓名 | 伍尚腾 |
| 班 级 | 17计算机7班 |
| 学 号 | 511721050575 |
| 指导教师 | 方坚 |
| 起止时间 | 2020年06月20日-28日 |
| 设计地点 | 腾讯课堂直播+录播 |

计算机与信息工程学部2020年制

**目录**

[第1章 Python-某招聘网爬虫实验报告 3](#_Toc5513)

[1.1 概 述 3](#_Toc25366)

[1.2 制作IP代理池（五号黑体） 4](#_Toc27732)

[1.2.1 隐藏身份信息(五号黑体) 4](#_Toc24218)

[1.2.2 爬取IP代理数据 4](#_Toc29367)

[1.2.3 IP代理池的制作 4](#_Toc30405)

[1.3 爬取首页的岗位信息（五号黑体） 4](#_Toc26077)

[1.3.1 爬取数据(五号黑体) 4](#_Toc12668)

[1.3.2 提取有价值数据 5](#_Toc9178)

[1.3.3 存储结果 5](#_Toc11584)

[1.4 爬取所有岗位信息（五号黑体） 5](#_Toc3836)

[1.4.1 分析所有页面url并拼接网址(五号黑体) 5](#_Toc20912)

[1.4.2 分页爬取数据信息 5](#_Toc27846)

[1.4.3 提取所有岗位信息 6](#_Toc24863)

[1.4.4 保存数据信息 6](#_Toc9654)

[1.5 课程设计体会与总结（五号黑体） 7](#_Toc23483)

[1.5.1 课程设计体会 7](#_Toc13377)

[1.5.2 个人总结 7](#_Toc26909)

[1.6 参考文献（五号黑体） 8](#_Toc17581)

[1.7 附录 9](#_Toc25408)

# Python-某招聘网爬虫实验报告

## 概 述

随着互联网的发展壮大，人们获取信息的途径逐渐被网络所替代。互联网发展初期，人们主要通过浏览门户网站的方式获取所需信息，但是随着Web的急剧发展，用这种方式寻找自己所需信息变得越来越困难。目前，人们大多通过搜索引擎获取有用信息，因此，搜索引擎技术的发展将直接影响人们获取所需信息的速度和质量。

爬虫就是模拟浏览器去访问各个网站，并把看到的有价值数据信息取回来的探测机器。爬取某招聘网提取需要的信息并对其进行整理分析，海量数据人工难以收集整理并分析，爬虫可以快速提取海量有用信息，快速对其整理成客观的分析表，使用python语言，pycharm软件，爬虫技术，对某招聘网站爬取内容并保存下来，然后把保存的信息进行提取整理，最后制成可观的图表，让大家都能一眼看懂并了解。

技术爬取：①爬虫技术模块：urllib、requests、lxml、re；②HTML标签、XPath、json、Excel表格；③做数据分析、numpy/pandas/matplotlib/xlwt；④访问频率限制/time/random等库。

利用上述这些方法，我们可以对很多网站进行类似的爬取想要的信息，爬虫和数据分析结合起来。

## 制作IP代理池（五号黑体）

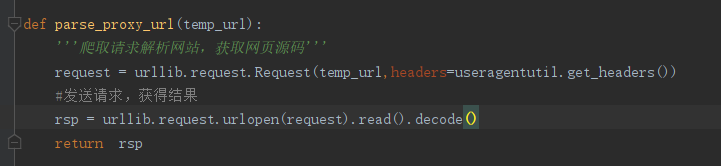
下面通过隐藏真实身份、爬取IP代理数据、IP代理池的制作这几个方面进行详细掌握。

### 隐藏身份信息(五号黑体)

# 添加请求头,隐藏身份信息

headers={ "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.61Safari/537.36" }

### 爬取IP代理数据



获得网页源码后提取信息：

1. 解析器对象； --用于解析 html 代码
2. ）获得所有的 tr 列表；

（3）遍历 tr 列表，tr 单个元素；

（4）提取数据内容；

（5）-->拼接数据、处理数据



保存获得的ip数据



### IP代理池的制作

代理池：验证当前 IP 代理是否可正常通信？ 能正常通信，存入到一个文件池中；若不能 正常通信，则丢弃该 IP 代理。

有用，添加。 --->响应码！ --当响应码为 200 时，表示请求成功！

-->思路分析： （1）获取到昨天保存的 IP 代理数据；

（2）测试一个一个的 IP 代理，有用，则添加到代理池中；

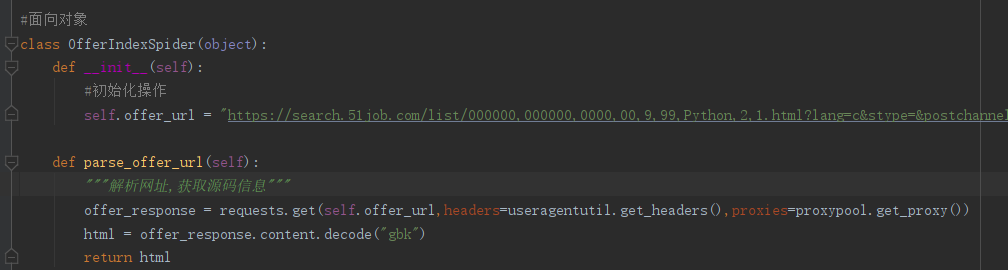
（3）最终保存文件。 --有用的 IP 代理



## 爬取首页的岗位信息（五号黑体）

爬取首页的所有源码信息、提取岗位数据并保存

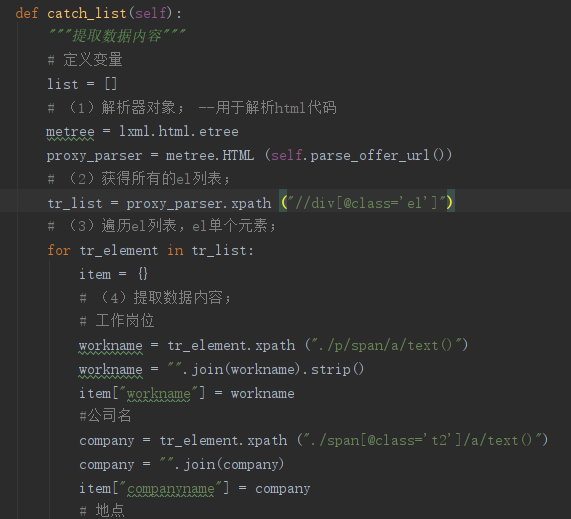
### 爬取数据(五号黑体)

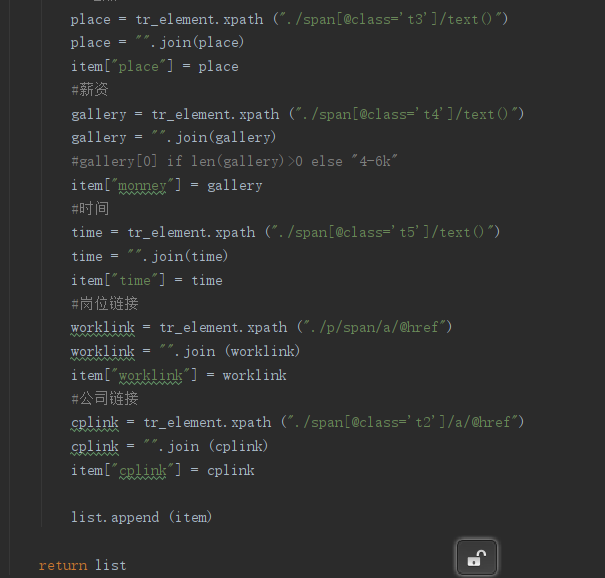


### 提取有价值数据

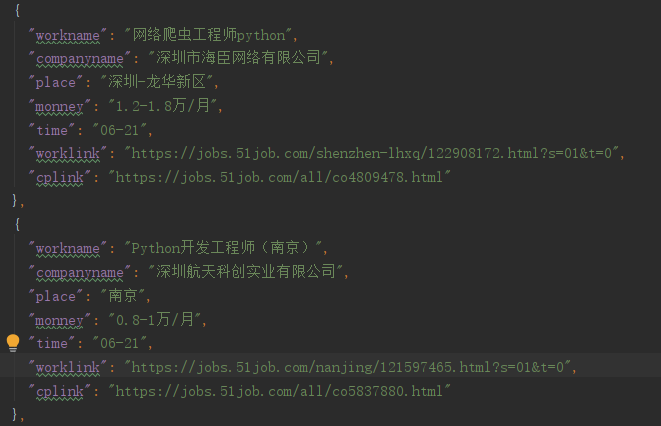
提取职位名、职位名称链接、公司名、公司链接、工作地点、薪资、发布时间

先解析页面，然后xpath提取





### 存储结果



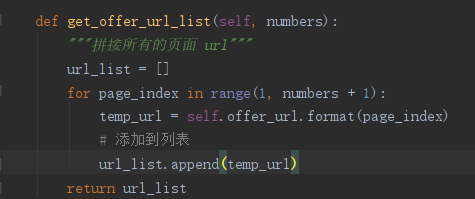
## 爬取所有岗位信息（五号黑体）

（1）分析所有页面的 url 网址，得出规律，处理所有页面网址； （2）遍历所有网址，一页一页数据爬取； （3）提取有效数据； （4）保存数据内容。

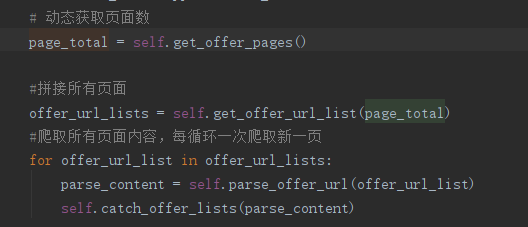
（1）遍历所有的 url 列表网址； （2）分页爬取页面源码内容； （3）提取页面源码中的岗位信息值； （4）保存所有岗位信息

### 分析所有页面url并拼接网址(五号黑体)

每一页的网站只有一个数字不同，每增加一页数字加一



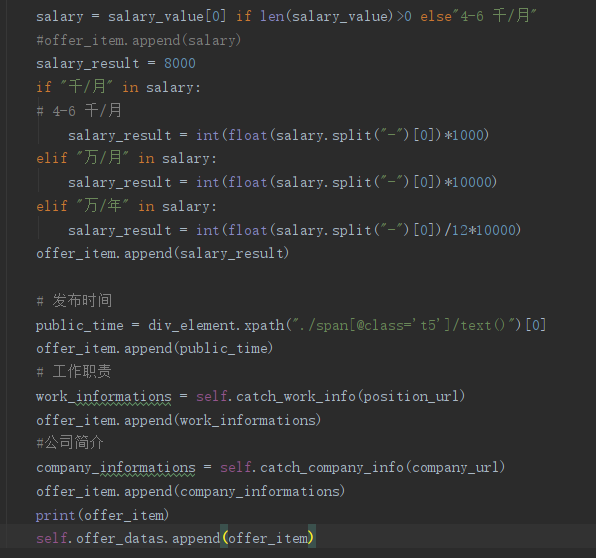
### 分页爬取数据信息



### 提取所有岗位信息

提取职位名、职位名称链接、公司名、公司链接、工作地点、薪资、发布时间 -->职位的工作职责/公司简介





### 保存数据信息



## 课程设计体会与总结（五号黑体）

### 课程设计体会

我是从python3开始做爬虫的，首先，python3的语法必须知道，不过python3并不难，语法也非常简洁。然后是学习python的各种库，为了打好基础，从基础的库开始学习会比较好，比如urlib，然后学习requests，这种是用来发请求的。返回的数据又是各种各样的类型，有html，js，json，字符串等，针对每种格式，当然都需要合适的解析器，当然，正则都能做。。。这里，解析html用到xpath。经历了困难重重，终于得到了想要的数据，学习到了很多技术，发现了爬虫非常好玩，还可以结合数据分析一起。

### 个人总结

通过这次实训使我学到了很多，同时也让我意识到了我需要学习的还有很多。通过实训让我更深进一步的对自己的专业，对自己的能力，对自己所学的有正确的认识，并且能在以后的学习工作中不断提高和完善自己。

## 参考文献（五号黑体）

韩克、薛迎霄主编．单片机应用技术[M]．北京：电子工业出版社，2013.03

彭伟主编．单片机C语言程序设计实训100例[M]．北京：电子工业出版社，2012.10

姚敏主编．检测系统数字化测试技术[M]．北京：机械工业出版社，2013.10

孙静.基于单片机的酒精浓度检测仪设计[J].通化师范学院学报.2016.06

郭霖编著．《第一行代码》．北京：人民邮电出版社，2013.05

李刚编著．《疯狂Android讲义》．北京：电子工业出版社，2014.08

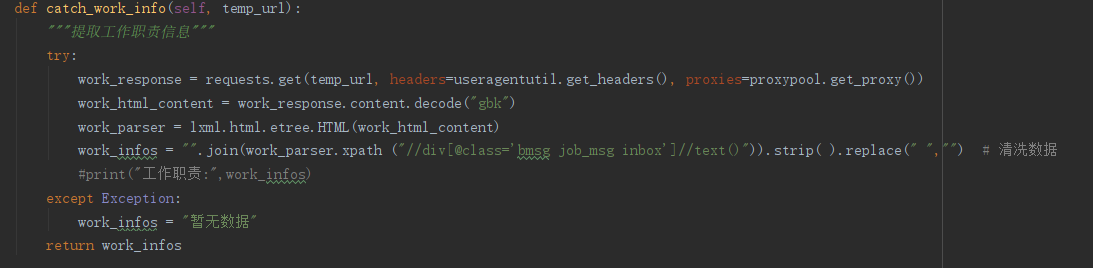
袁国忠译．《Python基础教程(第3版)》．北京：人民邮电出版社，2018.02

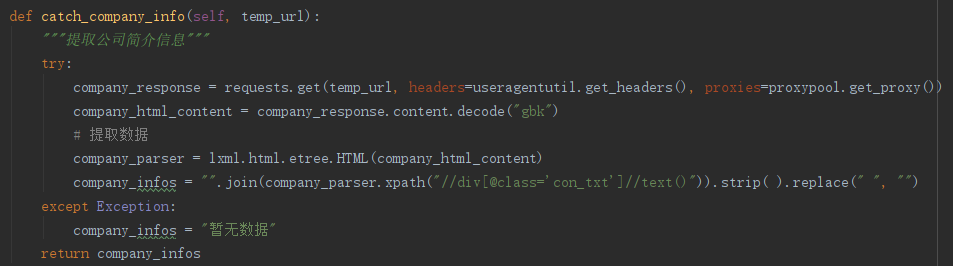
宋吉广译.《Python核心编程(第二版)》．北京：人民邮电出版社，2008.07

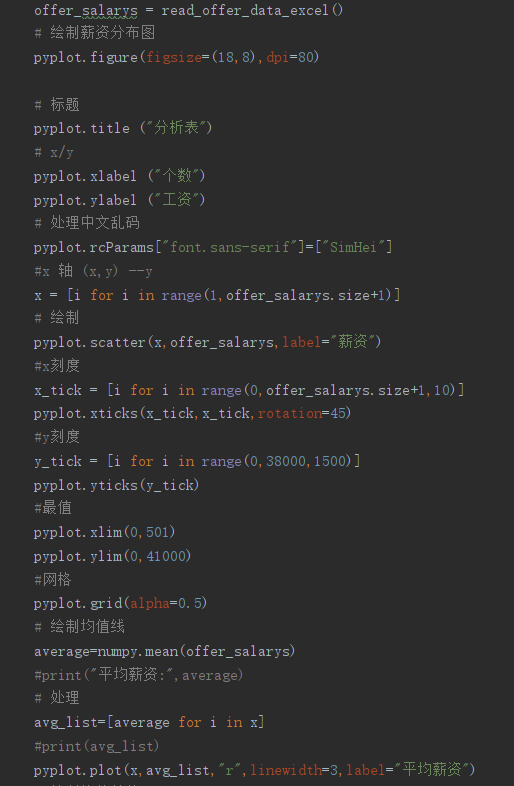
韦玮著．《精通Python网络爬虫》．北京：机械工业出版社，2018.04

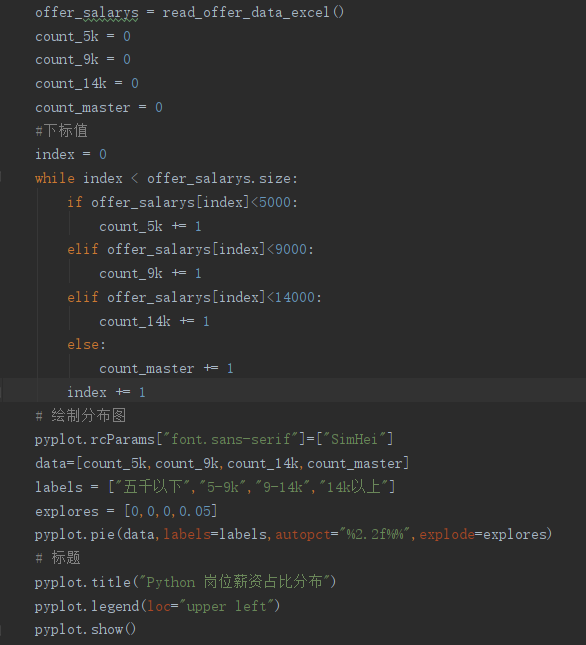
## 附录

部分重要核心代码，上面已经截了大部分









显示运行图片或结果图片

