

**课程设计(综合实验)报告**

( 2019 -- 2020 年度第 2 学期)

名 称： Python课程设计

题 目： 爬取2020年Python开发工程师职位薪资

院 系： 控制与计算机工程学院

班 级： 软件1801

学 号： 120181080429

学生姓名： 施星宇

设计周数： 1

成 绩：

日期：2020 年 6 月 27 日

**一、课程设计的目的与要求**

目的：运用本学期学习的Python爬虫知识，对互联网信息进行精确筛选，找出所需信息。

要求：在51job网站上，爬取2020年发布的Python开发工程师的职位的薪酬，计算北京地区该职位的平均薪酬，要求采用多进程或者多线程的方式完成编程

**二、设计（实验）正文**

此次关于爬虫的题目，运用了较多模块，如re，requests等，同时加入了线程池提高爬取速度。

代码整体分为六个函数块，包括main函数：

首先是get\_html()函数，是对访问页面原码的获取。通过对于51job网页上Python开发工程师的搜索，找到不同页码之间的相同点，再通过urllib库模拟发起网络请求，得到源码。

第二个get\_info()函数，顾名思义是通过扒下来的网页源码，得到我们所需要的特定信息，如题目中所需要的工作职位名，公司名，工作地点，薪资等。

第三部分crawler()函数与run\_thread\_pool()函数，则是运用了线程池提高了爬取页面的速度，同时爬取下来的信息我这里将其放在了文件中。

第四部分cal()则是对爬取下来的北京地区Python开发工程师薪资进行计算，最后得出平均薪资。

**三、课程设计总结**

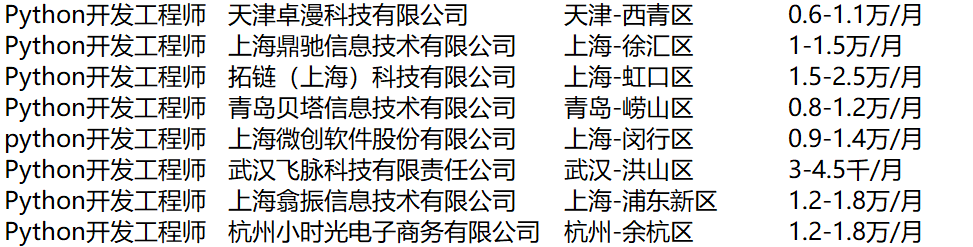
在没有系统学习爬虫知识的情况下，对于爬虫相关的知识还是有些难以理解的。但是通过阅读老师分享的资料以及自主学习，还是摸到了一点入门。

在爬取网页源码的过程中，对于阅读html语言还是有些困难，通过对于类似题目的搜索，筛选出了所需信息。

正则表达式、多线程与爬虫知识相结合，综合运用才能写出一段程序。

**附录（设计流程图、程序、表格、数据等）**

爬取结果（文件一部分展示）：



北京地区平均薪资：

