

**Python课程设计报告**

( 2019 -- 2020 年度第 2 学期)

名 称： python课程设计

题 目： [51job职位薪资爬虫](https://github.com/TwistF/homework/blob/master/webspider/%E3%80%8A51job%E8%81%8C%E4%BD%8D%E8%96%AA%E8%B5%84%E7%88%AC%E8%99%AB%E3%80%8B%E8%AF%B4%E6%98%8E%E6%96%87%E6%A1%A3.docx" \o "《51job职位薪资爬虫》说明文档.docx)

院 系： 控制与计算机工程学院

班 级： 软件1802

学 号： 120181080706

学生姓名： 钟诗伟

指导教师： 熊建国

设计周数： 1

成 绩：

日期：2020 年 6 月 27 日

1. **课程设计(综合实验)的目的与要求**

**1.**目的与要求：熟练掌握python的网络爬取数据、HTML文本分析以及多进程、多线程、数据库的灵活运用

1. **设计（实验）正文**

1.题目：

爬虫程序：

在51job网站上，爬取2020年发布的Python开发工程师的职位的薪酬，计算北京地区改职位的平均薪酬；

1. 难点：python库函数的灵活运用、HTML文本的分析、多进程的进程数量的限制等
2. 思路：
3. 爬取网页：使用requests爬取网页内容，正确返回HTML文档，出错返回404
4. 分割HTML：使用BeautifulSoup查找内容，将找到依次内容加入列表，方便后续多进

程爬取

1. 寻找内容：使用正则法则对找到需要的内容
2. 将内容存入数据库
3. 多进程读取数据库中关于薪资的数据，求出平均值
4. 关键代码：

1）*#将网页数据分成小份***def** get\_part(url):  
 h\_text = getHtmlText(url)  
 **if** h\_text == **"404"**:  
 **return "404"  
 else**:  
 list = []  
 soup = BeautifulSoup(h\_text,**'html.parser'**)  
 **for** i **in** soup.find\_all(**'div'**,class\_=**"el"**):  
 i = str(i)  
 **if '<span class="t2">' in** i **and '<div class="el title">' not in** i: *#过滤掉一些不合格的检索结果* list.append(i)  
 **return** list

2）*#保存信息到数据库***def** save(text):  
 **if** get\_data(text) == **'None'**:  
 **pass  
 else**:  
 list = get\_data(text)  
 SQL.Add(list)

1. **课程设计（综合实验）总结或结论**

总结：总共课设做下来，自己的感受就是写代码一定要有框架，有注释，否则第二天就会自己也不知道在写啥了、写到哪了

**附录**

附上Navicat数据库的完成图：

