

**Python大作业实验报告**

( 2019 -- 2020 年度第 2 学期)

名 称： python大作业

题 目： 爬取2020年python工程师薪资信息

院 系： 控制与计算机工程学院

班 级： 软件1801

学 号： 120181080403

学生姓名： 冯梓硕

日期：2020 年 6 月 22 日

**一、选题的目的与要求**

**选题目的**：运用python的爬虫技术，精准、全面、快速的获取特定网址的特定信息，并且用python设计特定的数据结构来存放这些数据并分析，最后得到具有现实意义的结论。

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

1. **作业设计正文**

**这次程序设计运用的最重要的模块是爬虫相关模块，包括requests、beautifulsoup、lxml等模块，还有re模块来实现正则表达式来筛选信息，threading模块来实现多线程，提高了整个程序的运行速度。**

**第一部分为函数get\_content用于获取html信息。要知道，在51job网站上的python开发工程师的信息不止一页，一共有37页，所以要摘取一部分网页的网址来分析每一页的地址有何不同才好组织url的形式。另外，有关爬取的部分遇到了一些阻力，在正常编写的情况下，系统会提示类似“远程主机强制关闭了连接”等错误，通过上网搜索知道了这一错误是基于远程主机识别出这是一个爬虫程序的访问，因此拒绝了访问，为此我加入了header来伪装成浏览器，这一问题才最终被解决。**

**第二部分是多线程部分。这一部分的主要功能为对第一部分获取到的信息进行筛选和分类。一共分为4个线程，一个线程负责10页。根据html信息当中的标签来将获取的信息分类，并且归档到五个专门的列表当中，需要注意的是，这五个列表属于全局变量，但是四个线程归档都是归到这五个列表当中，因此需要加入互斥锁让四个线程对这五个列表互斥访问才可以保证程序和数据不出错。**

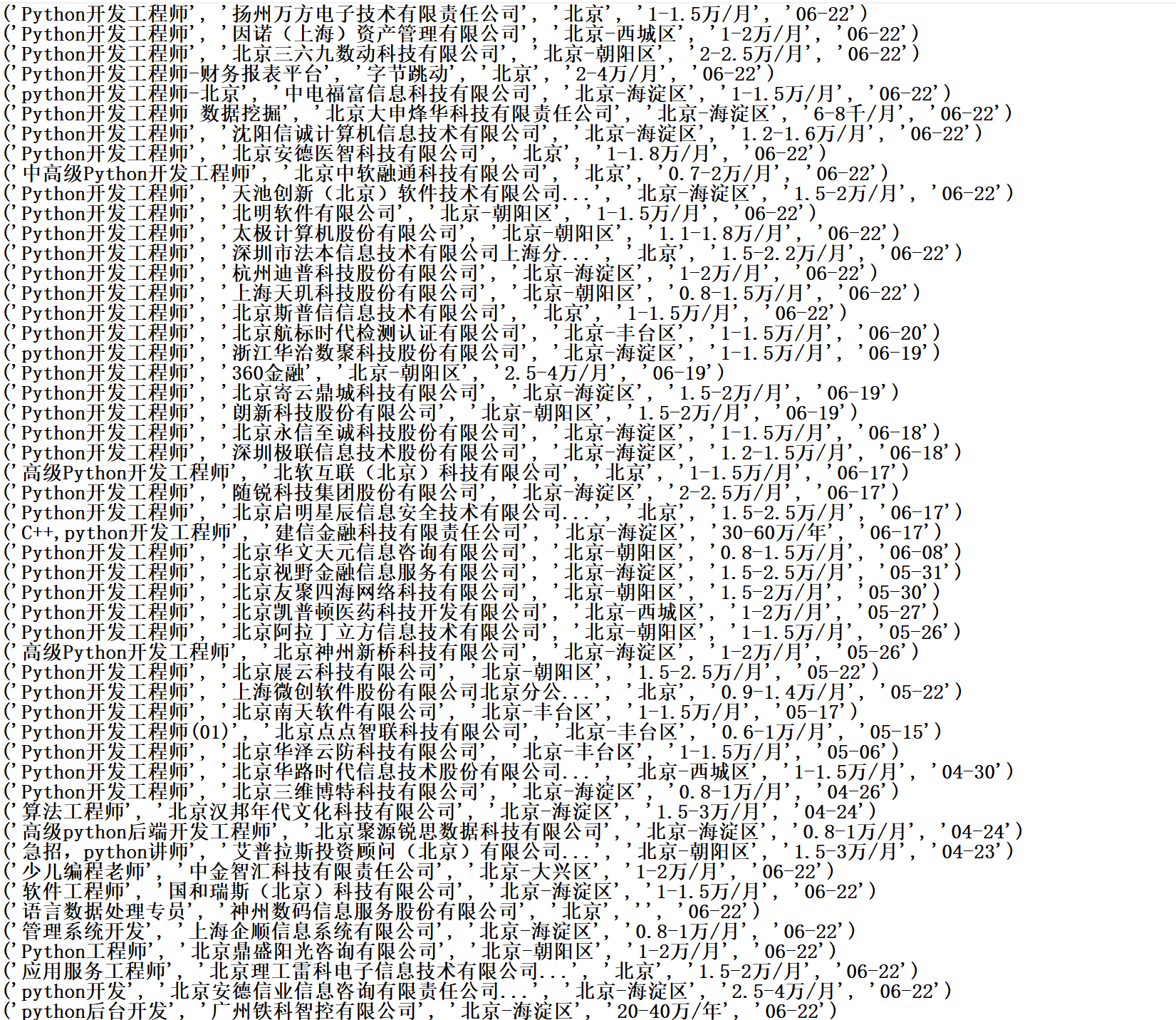
**第三部分是主程序部分。这一部分的原理较为简单，将上面归档的数据进行处理、计算并且展示结果。**

1. **总结或结论**

**总结来说，本次大作业算是对自己本学期所学内容的集大成。对于python的基础语法知识以及代码结构和格式是必不可少的。再往上层来看，多线程、互斥锁、爬虫、正则表达式、列表数据结构、函数方法的运用、python开发过程中的编码制度等等，可以说这一学期的所有知识都可以在这一次作业中得到体现甚至是提高。在本次作业当中，自己的短板也体现的很明显。对于正则表达式的规则以及运用还不熟练，无法做到随用随取。还有爬虫相关的知识也太匮乏，在这方面甚至是对于上网搜索的资料都是一知半解。还有最重要的是对于python的代码构成不太会组织，结构太过繁琐，也不知道怎么改进，这些错误都是以后将要改正、改进的。**

**附录**

**爬取的python工程师信息（部分）**

****

**北京地区薪资平均值（保留小数点后三位）**

**O6LF_Q22D~EX5D31CKCKTPO**