# **软件设计报告书**

**项目内容：**用爬虫统计51job上北京地区python开发工程师的平均薪资

班级：软件1802

姓名：李义龙

学号：120181080611

日期：2020/6/23

1. **引言**

整个项目完成了以下功能：

以多线程爬虫爬取51job上指定的信息；

用数据库对链接和爬取到的信息进行存储；

对信息进行整合处理，得到“平均薪资”。

文件构成如下图所示

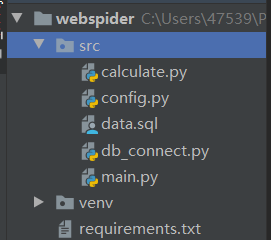


图1

1. **功能设计**

如图1所示，各文件的功能分别如下：

main.py 主文件，负责爬虫的绝大部分功能，同时引用、链接其他模块

db\_connect.py 内含数据库链接及调用的函数，负责与数据库相关的所有功能

config.py 数据库的配置文件，进行整体的参数修改

calculate.py 对得到的数据进行统计，算出结果

data.sql 存储数据库信息

整个功能的实现流程如下：首先，main通过db\_connect从数据库中读取需要爬取的网页。然后main通过自身的爬虫功能对各个网页进行多线程爬取。在爬取的过程中，爬取完成的html文件会传入解析html的函数，并将提取的信息通过db\_connect传入数据库。最后，所有网页的爬取结束后，main会调用calculate对数据库存储的信息进行读取、整合、统计，最后计算出平均薪资。

1. **数据库设计**

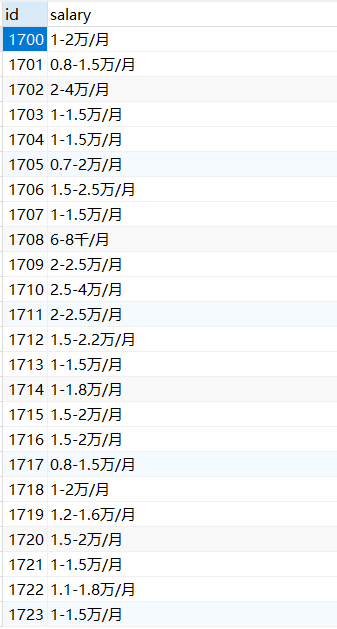
数据库分为 urls 和 salaries 两个表单

urls 表单用来存储需要爬取的网页



salaries表用来存储爬取到的信息(eg.1.3-2万/月)

QQ截图20200625232554



1. **经验和教训**

对于爬虫来说，效率和稳定性是关键。在这次的项目中，目标网站链接稳定、模板单一，爬虫可以简单而高效。但面对其他项目时，网络的链接效果、响应时间、信息的提取都有很大的随机性，需要更有经验的处理。

所以在这次项目中，即使目标网站相应速度快，我也仍然选择了使用retrying包来对响应时间进行处理，希望在以后的项目中能有所帮助。