

**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定向爬取

院 系： 控制与计算机工程学院

班 级： 软件1801

学 号： 120181080409

学生姓名： 赵嘉鸣

指导教师： 熊建国

设计周数： 1

成 绩：

日期：2020 年 6 月 25 日

**一、项目开发要求**

1.定向爬取51job网站

2.获取Python开发工程师的薪酬

3.计算北京地区该职位的平均薪酬

4.将爬取内容保存到数据库

5.采用多线程或多进程完成爬取

1. **项目设计正文**

程序结构：

-----data\_clean.py (数据清洗)

-----calculate.py (计算薪资)

-----craw\_3.py (主程序)

-----config.py (数据库配置文件)

1.进入网站，分析网站结构，解析页面

采用python内置模块完成页面请求解析



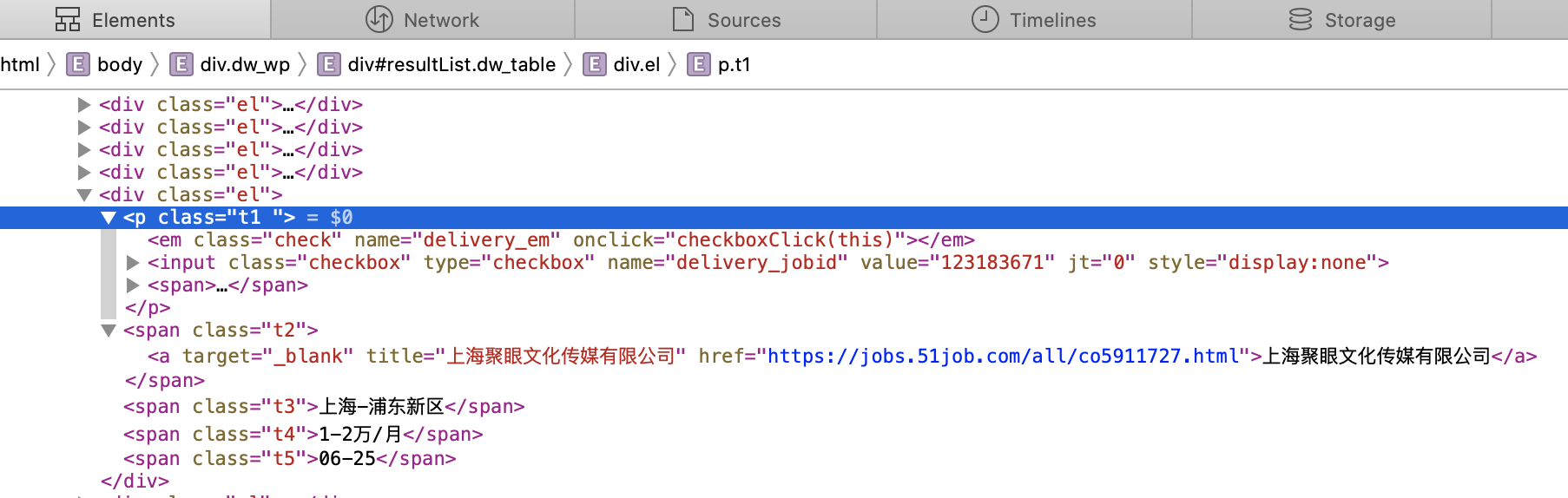
职位相关信息包括：职位名称，公司名称，工作地点，薪酬，发布时间

由此设计数据库表结构

class data\_table(Base):  
 \_\_tablename\_\_ = 'Jobs'  
 id = Column(Integer, primary\_key=True)  
 POSITION = Column(String(100))  
 COMPANY = Column(String(100))  
 ADDRESS = Column(String(100))  
 SALARY = Column(String(100))  
 DATE = Column(String(100))

采用sqlalchemy的orm，pymysql操作链接数据库

2.分析网页，采用beatifulsoup提取标签信息，



# 获取职位信息  
titles = soup.select("p[class='t1'] a")  
# 获取工作地点  
di = soup.select("span[class='t3']")  
# 获取公司  
company = soup.select("span[class='t2']")  
# 获取薪水信息  
salaries = soup.select("span[class='t4']") # CSS 选择器  
# 获取发布时间  
time = soup.select("span[class='t5']")

3.对薪资进行处理统一转换为万/月，并筛选北京Python爬虫工程师

数据清洗在data\_clean.py模块进行

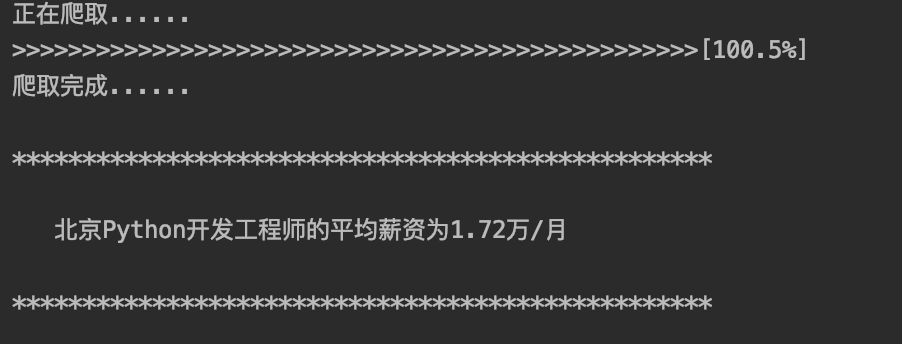
对薪资进行处理统一转换为万/月

4．分析网页可知勾选Python共有684个页面，每个页面有50条记录，共30000多条职位信息，

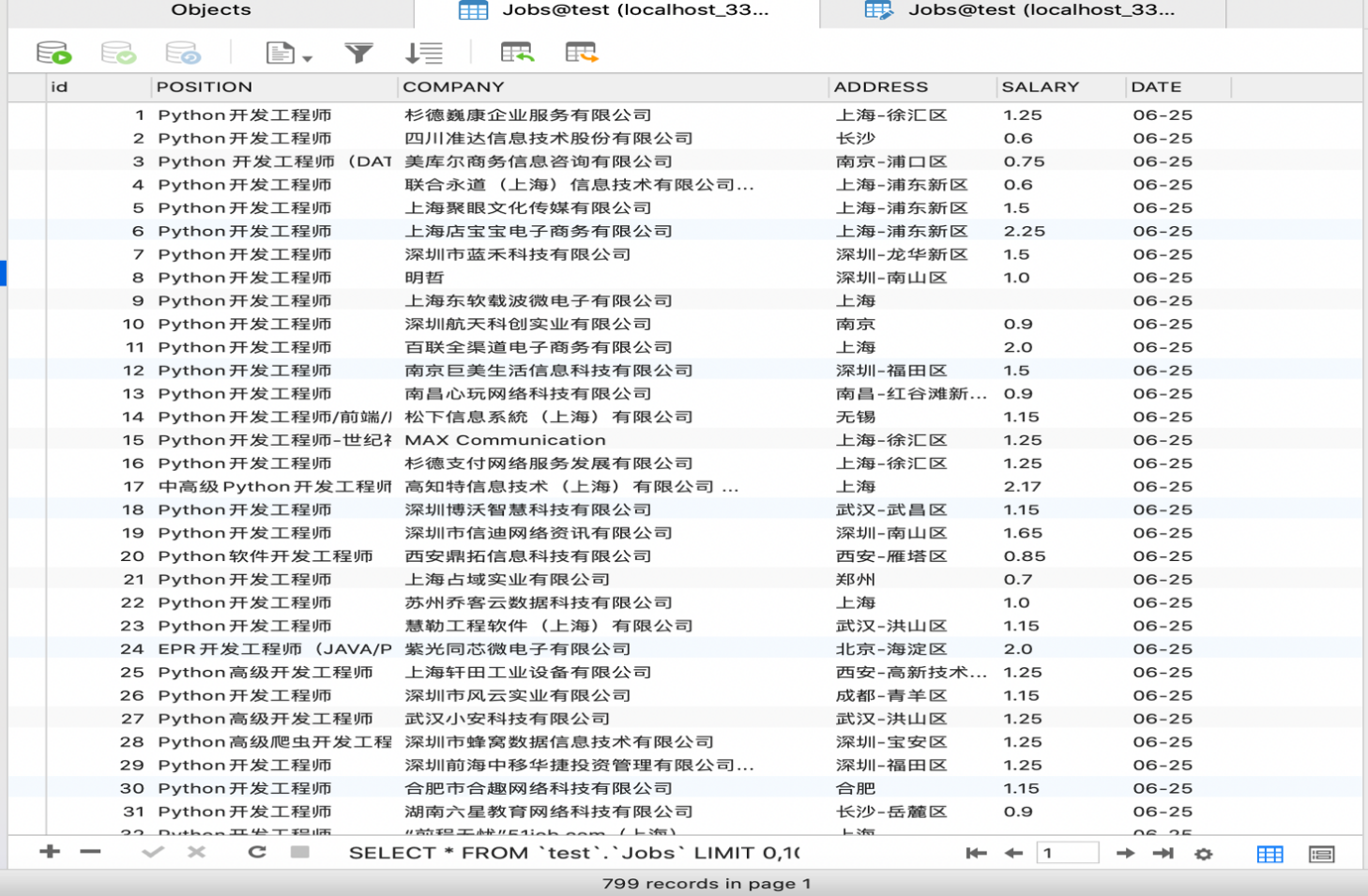
但经分析后面的页面无Python开发工程师职位，这里爬取前400页，共20000多条信息，

运用threading筛选出约800条Python开发工程师的职位信息，保存到mysql数据库中

5.计算北京地区python开发工程师平均薪资



**三、课程设计（综合实验）总结或结论**

数据采集

**总结**：设置header爬虫程序伪装成浏览器访问51job网站，防止网站反扒机制，特殊情况下。 可使用代理ip进行爬取。在爬取20000多条数据中，多线程优势明显，程序在2min内运行完成，数据清洗过程筛除了大量数据，最后仅有大概800条数据符合要求，多线程并发将数据写入mysql，并完成相关异常处理。在程序设计初期，本尝试用面向对象，采用进程池，线程队列对数据进行爬取，但爬取速度不尽人意，进一步优化得到该版本程序。

**参考文献**

1.sqlalchemy https://docs.sqlalchemy.org © [Copyright](https://docs.sqlalchemy.org/en/13/copyright.html) 2007-2020, the SQLAlchemy authors and contributors. Created using [Sphinx](http://sphinx.pocoo.org/) 3.0.4

2. beautifulsoup https://beautifulsoup.readthedocs.io/zh\_CN/v4.4.0/ © 版权所有 2013, delong.wang. Revision 355a041b.

Built with [Sphinx](http://sphinx-doc.org/) using a [theme](https://github.com/snide/sphinx_rtd_theme) provided by [Read the Docs](https://readthedocs.org).

3.网络爬虫课件 <http://note.youdao.com/s/S8G5aSjI>

具体格式参考：《信息与文献参考文献著录规则》GB／T 7714-2015.