Python 岗位信息的抓取与可视化分析

刘永伦

（大连大学 信息工程学院 ， 辽宁 大连 116622）

摘要：为了能更加直观地了解到国内 Python 有关的职业对学历、工作经验、地区分布与薪资分布等情况，采用 Python 的数据分析和处理功能，通过 Python 爬虫拉勾网在全国范围内的 Python 相关职位信息。根据实际需求删除有空值的信息、与 Python 无关的职业等方法对数据进行预处理，然后将清洗后的数据存入数据库，再利用 echarts 框架对数据进行可视化分析，用 Flask 框架开发 Web 应用程序，通过将数据直观地展示在网页，提高了用户查询信息的速度，方便求职者找到适合的职位。

关键词：爬虫；Python；可视化；Echarts；Flask

1. 引言

目前在各类求职网站中进行职位搜索时，都是需要求职者自己明确到具体的职位，地区等各类详细参数才能精准搜索到所需要的岗位信息，而对于 Python 岗位在哪个地区需求量高，薪资分布，学历分布与工作年限和薪资分布等概括类信息，求职网站并不能直观地展示给求职者。本文对求职网站的数据进行了有效的抓取和清理，最后将这些数据显示在可视化的 Web 界面上，从而实现对这些数据的多角度可视化分析，满足求职者的需求。

2. 总体设计

基于 Python+echart+Flask 的可视化设计需要实现数据抓取、数据清洗、数据入库、可视化等功能。通过实现对拉勾网 Python 相关职位的爬取与清洗后将数据存入 MySQL 数据库中，再通过 Flask 框架将数据库中的信息进行 JSON 序列化后提供到 api 接口中，最后使用 Echarts 框架在前端网页进行读取数据与可视化展示。功能结构图如图 1：

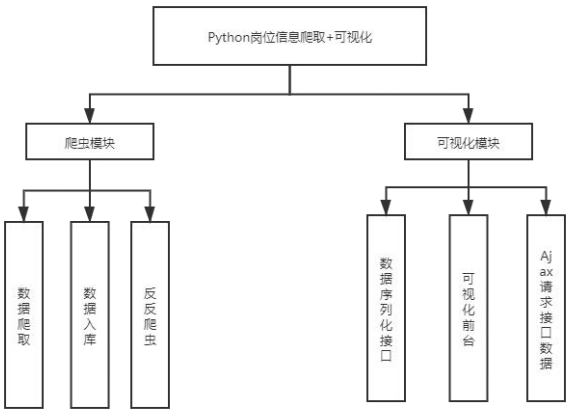

图 1. 功能结构图

3. 详细设计

3.1 数据获取

爬取拉勾网 Python 相关职位数据，因为网站禁止访问不是通过浏览器的正常访问， 所以我们需要手动在 header 里加上 UA 属性，伪装成浏览器进行访问[1]。构造 header 方法如下：

Self.header={

(1)“Host”:要访问的网站信息，

(2) 'User-Agent':浏览器的 UA 属性} [2]。

这里我们通常使用的 UA 是 PC 端的谷歌浏览器,即: Mozilla / 5.0(Windows NT 10.0;WOW64) AppleWebKit / 537.36(KHTML, likeGecko) Chrome / 75.0.3770.100Safari / 537.36

接收到 URL 地址的 JSON 数据页面后采用正则表达式进行匹配字符串利用双层循环来实现换页爬取与换行输出[3] [4]，获取数据，保存到 MySQL 数据库中。流程如图 2 所示：

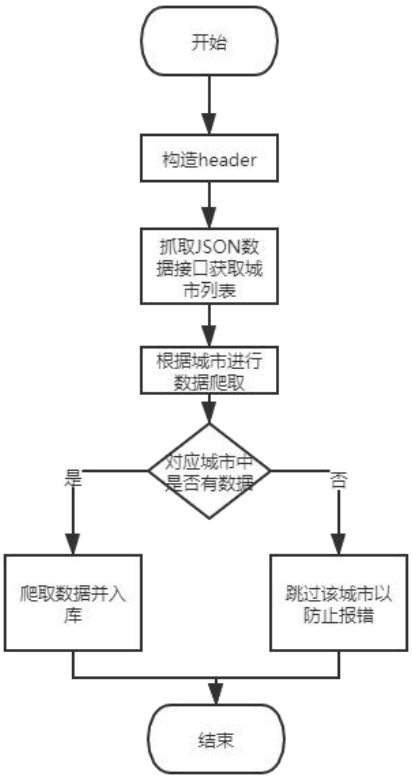

图 2. 数据获取流程图

3.2 反反爬虫与数据清洗

在抓取一个网站的数据时，不可避免的会因为请求太快而导致应用被屏蔽，甚至 IP 地址可能会被网站直接屏蔽，导致数据无法再获取。因此，在对数据进行爬行之前，我们应该设计出程序

的反爬虫方案。在本文中，我们使用慢爬虫技术完成此功能，因为慢爬虫技术不用像代理池技术那样需要很多钱来买或租代理服务器，这是一种人人都可以使用的技术。

爬行数据时，我们从数据接口取 10 个数据，依次放入仓库。在请求网站数据时收到频繁响应后，我们随机等待 5-20s，然后重新启动爬虫。

在数据入库时，我们采用清洗非空数据的方法，在数据插入之前设置一个标记位 flag，如果数据不完整或者出现其他问题，则改变 flag 的值来让数据不能入库。总体流程图如图 3 所示：

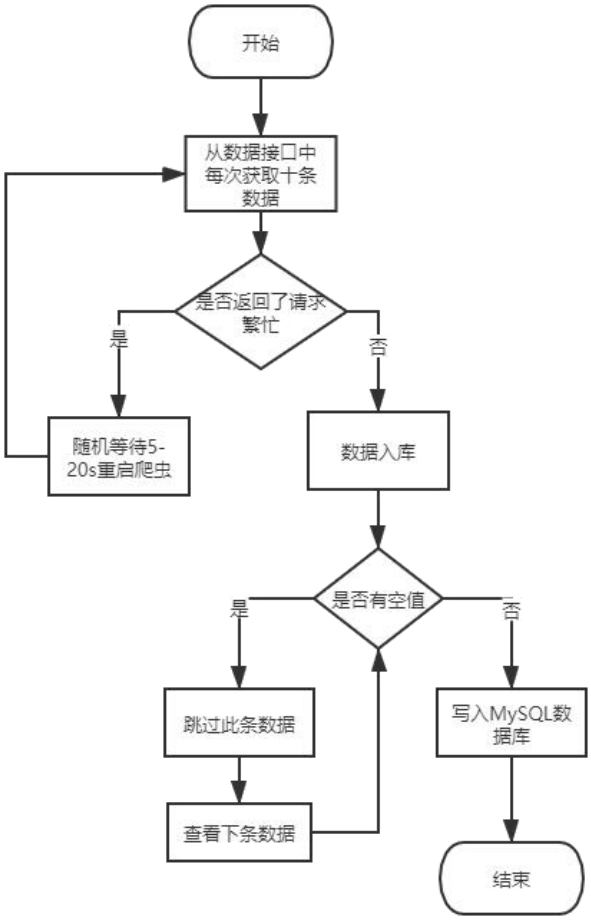

图 3. 数据清洗与入库流程

3.3 数据分析

经过清洗后的表格职位信息应全是 Python 相关的职业，表格信息还包括公司名称，公司地点，公司性质， 薪资，学历要求等信息。求职者想要找到合适的职位， 需要对各招聘公司的学历要求，工作经验要求，公司性质、所处地区进行可视化分析。了解哪些性质的公司在招聘人才，何种性质的公司对人才需求最多，分析公司招聘何种学历和工作经验的人才更为普遍， 也就是求职者应达到的最基本的要求。此外，对招聘公司所处地区经济是否发达进行分析，是对公司的 发展和经济实力的预测，方便求职者更准确的做出选择[5] [6]。

3.3.1 学历分布玫瑰图

可视化结果如图 4 所示。采用玫瑰图可以十分直观的看出不同学历的占比情况，其中 82%的企业对学历的要求都是本科学历，不超过 9% 的企业对学历基本要求是大专，在总数 2155 个职位中，，硕士学历要求有 144 个，博士学历要求有 42 个，只有 102 个职位是不限制学历的。

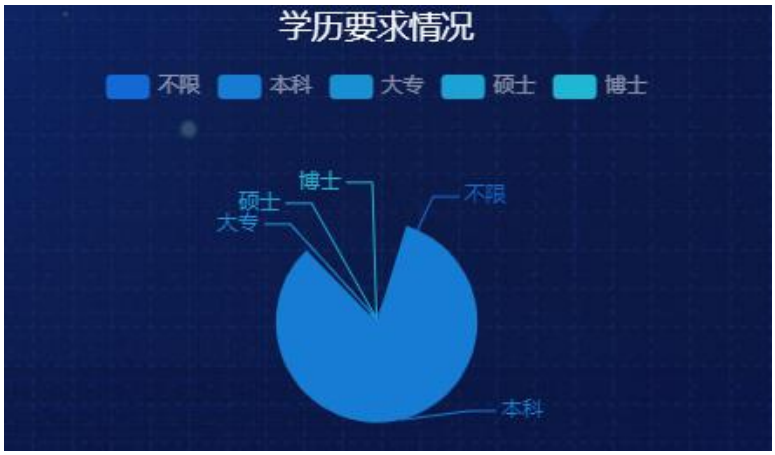

图 4. 学历分布玫瑰图

3.3.2 职位地理分布图

通过将城市地理分布信息可视化不难发现对 Python 岗位需求强烈的城市分布在沿海或者省会城市中，如图 5 所示：

图 5. 职位地理分布图

3.3.3 工作年限直方图和工资分布直方图

如图 6 所示，在已发布的 2155 个职位中，对应届生的需求岗位仅匹配了 202 个，对 Python 需求最高的岗位是需要工作 3-5 年和 1-3 年工作经验的人。由此可见，与 python 相关职位的求职者大多是跳槽，对应届毕业生的需求并没有想象的那么大。

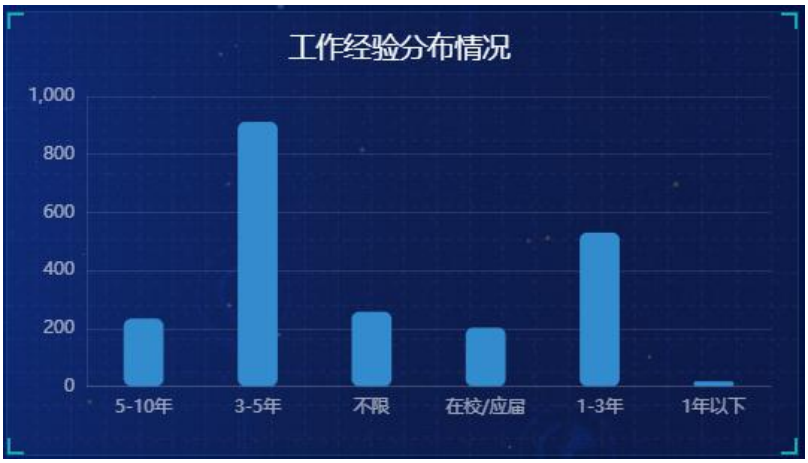

图 6. 工作经验分布柱状图

而对于薪资分布情况，如图 7 所示，最低为 8k 左右，最高在 40k 左右，但是数据大体平均分布在 15k 以上，说明 Python 类的相关岗位平均的薪资还不错，基本可以超越当地或全国的平均薪资水平。

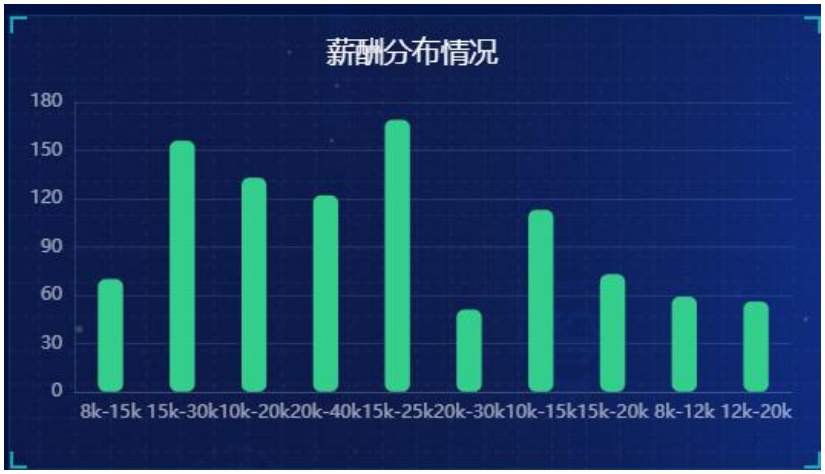

图 7. 薪资分布柱状图

4. 总结

使用 Python 和 Echarts 可视化框架实现了 Python 在拉勾网中的 Python 相关职位信息的抓取和可视化系统，实现了数据到图表的转换，使数据更具可读性和价值。再通过将图标变为可视化的 Web 应用直观地展示结果，使得用户可以更加方便地发掘隐藏的数据关联，从而筛选信息，做出决策。此外，通过多种不同可视化图标的相互组合，取长补短，也使得可视化技术可以更加便捷地投入到生产实践中去，让非专业从业人员也可以快速读懂数据。

参考文献：

[1] 黄岷昊, 丁浪, 张雪莲. 基于 Python 的网络爬虫及文本可视化[J]. 电脑编程技巧与维护, 2020(7): 24-25.
[2] 陈清. 基于 Python 的网站爬虫应用研究[J]. 通讯世界, 2020, 27(1): 202-203.
[3] 刘鑫. 网络爬虫在信息检索中的研究与应用[J]. 数字技术与应用, 2017(5): 95-97.
[4] 熊畅. 基于 Python 爬虫技术的网页数据抓取与分析研究[J]. 数字技术与应用, 2017(9): 35-36.
[5] 何佳, 惠建忠, 王曙东, 洪晓媛, 王阔音. Python 在 CINRAD 风暴数据可视化中的应用[J]. 气象科技, 2020(3): 374-379.
[6] 刘艳玲, 姚建盛. Python 在数据可视化中的应用[J]. 福建电脑, 2020(3): 30-31, 34

Job Information of Python Data Analysis and Visualization System Implementation

Liu Yong-lun[1]

(1. Dalian University, Dalian 116611，China)

Abstract: In order to have a more intuitive understanding of Python related occupations in China, such as education background, work experience, regional distribution and salary distribution, Python data analysis and processing functions are adopted, and Python-related job information nationwide is pulled through Python crawler. Free to delete value according to the actual demand of information, has nothing to do with the Python career methods for data preprocessing, and then the data in the database after cleaning, recycling Echarts framework for data visualization analysis, with a Flask framework for Web application development, shown visually through the data on the Web, improve the speed of the user query information, convenient for job seekers to find a suitable job.

Key words: Crawler; Python; Visualization; Echarts; Flask