

基于网络爬虫的就业岗位信息分析

杨海迎

(云南机电职业技术学院, 云南 昆明 650203)

摘要: 随着人工智能技术逐步的深入, 大数据时代的来临, 大数据也开始在社会中发挥巨大作用。本文以拉勾网为例, 在 Python3.8 和 MySQL 的基础上利用网络爬虫技术对当前就业大数据进行分析处理。以“Python”相关的职业岗位为实例, 设计研发了一套针对于招聘就业的专用爬虫, 对就业起到了积极作用。

关键词: 网络爬虫; 就业; 职业岗位

为了帮助高校毕业生在择业时能够快速获取特定的岗位需求信息, 并且通过快速数据分析得到自身择业的准确定位, 从而做出更好的选择, 为此本文以拉勾网为基础, 以“Python”相关的职业岗位为实例, 设计研发了一套针对于招聘就业的专用爬虫。

一、数据获取

(一) 数据需求分析

我们在就业的时候首先考虑的是自己能做什么领域的工作, 自己掌握了一些什么技能。那我们就需要了解, 一个岗位的技术需求, 学历需求, 是否需要工作经验。当我们满足了这个岗位的所有需求以后就要考虑, 工作地点(是否是北上广), 融资阶段(是否是上市公司), 公司规模, 薪资, 工作时间, 福利待遇(是否有带薪休假)。总结一下就是: 工作地点, 工作经验, 学历要求, 融资阶段, 公司规模, 行业领域, 薪资。下面就围绕这些关键字展开。

(二) 爬虫设计的整体思路

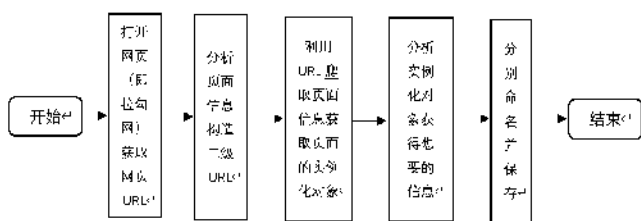


图1 爬虫设计的整体思路

(三) 具体代码

1. 引入库

利用 requests 库完成网页强求和下载, 利用 bs4 库对网页进行解析, 利用 time 库控制频率防止被反爬检测, 如图 2 所示。

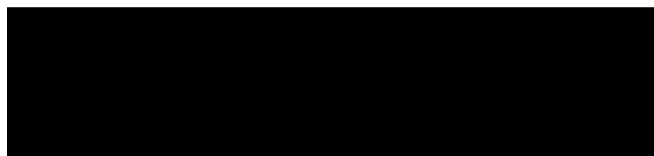


图2 引入库

2. 获取页面总页数

利用获得的 URL 进行初步的页面爬取, 获得总页数, 为 URL 的二级构造做准备, 如图 3 所示。

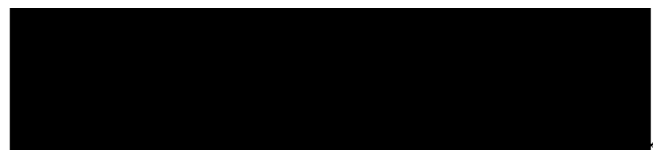


图3 获取页面数

3. URL 的二级构造

利用获取的总页数对 URL 进行二级构造, 利用 requests 库下载网页, 获取实例化对象。

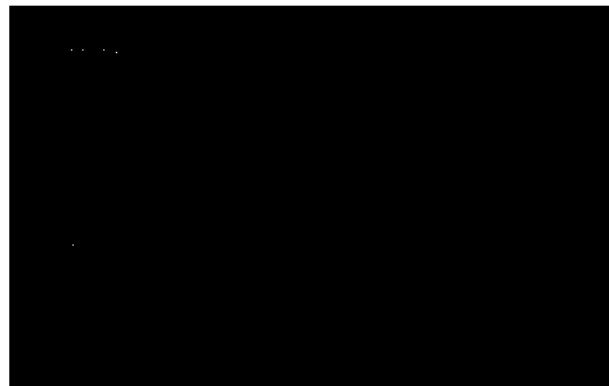


图4 URL 的二级构造

4. 解析实例化对象

获取信息并保存利用 bs4 库解析对网页进行解析获得正确的信息, 分别赋值给变量。job, address, company, salary, experience, education, industry, financing, scale, welfare 改变变量的格式利于保存, 将正确的信息写入文件, 每爬取一页停 20 秒, 防止被反爬机制检测出来, 如图 5 所示。

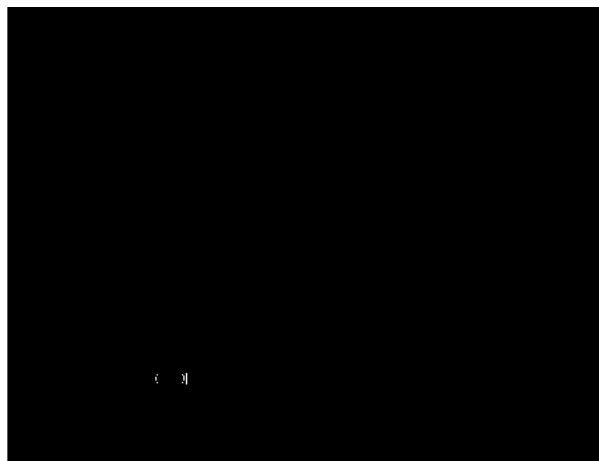


图5 解析实例化对象

5. 定义主函数

接收用户输入并赋值给变量，利用其定义初始 URL，获取网页头文件 user-agent 键值对，对代码进行伪装，防止被反爬机制检测，用 open 打开文件定义好文件保存的位置，利用 for 循环重复调用函数，爬取所有的页面信息，最后关闭文件，如图 6 所示。

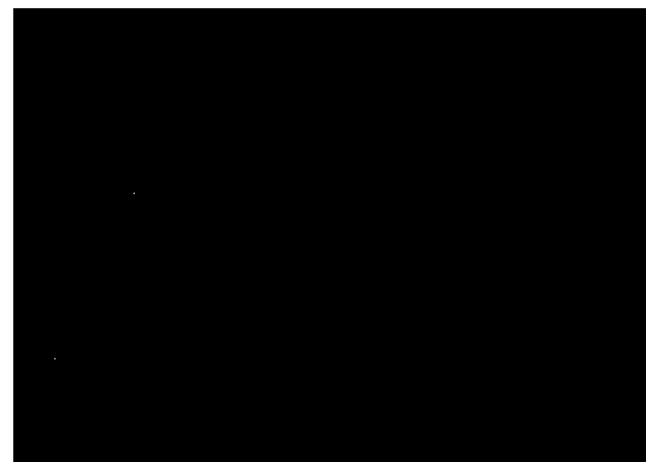


图 6 定义主函数

6. 数据展示

通过这样的形式能够实现数据的有效抓取，并且将研究过程中抓取到的相关信息输入到电脑中，为接下来的就业数据分析工作奠定基础。部分数据如下图所示：

counts	job	address	company	salary	experience	education	industry	financing	scale	welfare
1	Python	北京	贝壳	20k-40k	经验3-5年	本科	房产家居	上市公司	2000人以上	"五险三薪"差
2	Python	北京	猎户星空	18k-35k	经验3-5年	本科	移动互联网/B端		500-2000	"高绩效发展"
3	python	北京	字节跳动	20k-35k	经验1-3年	本科	文娱 内容	C轮	2000人以上	"六险一金"
4	Python	青岛	什么值得买	8k-15k	经验3-5年	大专	电子商务	上市公司	500-2000	"五险一金"
5	Python	上海	字节跳动	25k-40k	经验3-5年	本科	文娱 内容	C轮	2000人以上	"六险一金"
6	python	青岛	什么值得买	8k-15k	经验3-5年	不限	电子商务	上市公司	500-2000	"五险一金"
7	python	杭州	蚂蚁科技	12k-24k	经验1-3年	本科	移动互联网/A轮		50-150人	"六险一金"
8	python	北京	Gridsum 国双	10k-20k	经验1-3年	本科	数据服务/企业	上市公司	500-2000	"上市大数据"
9	python	深圳	新纳康	12k-24k	经验1-3年	本科	电商/数据服务/不需要融资		15-50人	"五险一金"成长
10	Python	杭州	圆通	15k-30k	经验3-5年	本科	物流 运输	上市公司	2000人以上	"上市公司福利"
11	Python	北京	GamesVerse	15k-30k	经验1-3年	本科	游戏/软件/开发/不需要融资		50-150人	"下午茶"swll
12	python	北京	北京新纳康科技	15k-30k	经验3-5年	本科	移动互联网/上市公司		500-2000	"六险一金"福利"
13	python	深圳	金证股份	10k-20k	经验1-3年	本科	金融	上市公司	2000人以上	"五险一金"福利"
14	python	北京	金证股份	17k-34k	经验不限	硕士	移动互联网/天理轮		15-50人	"集团子公司"
15	python	北京	聚肉	14k-20k	经验不限	本科	移动互联网/C轮		150-500	"七险一金"
16	python	北京	一淘互通	13k-16k	经验1-3年	本科	广告营销/移动互联网		50-150人	"海外社交媒"
17	python	北京	圆通科技	18k-28k	经验3-5年	本科	移动互联网/A轮		15-50人	"发展前景高"
18	python	广州	医维久联	10k-18k	经验3-5年	本科	医疗 健康	A轮	50-150人	"团队氛围好"
19	Python	北京	字节跳动	20k-40k	经验1-3年	本科	文娱 内容	C轮	2000人以上	"优厚下午茶"
20	python	北京	字节跳动	30k-45k	经验3-5年	本科	文娱 内容	C轮	2000人以上	"六险一金"
21	python	深圳	金证股份	8k-13k	经验在校/应届	本科	金融	上市公司	2000人以上	"五险一金"福利"
22	python	广州	小满游戏	4k-6k	经验在校/应届	本科	游戏	不需要融资	150-500	"双休双薪"大"
23	Python	惠州	豆蔻科技	8k-12k	经验1-3年	大专	社交/广告营销/不需要融资		少于15人	"五险午餐"
24	python	杭州	蚂蚁科技	10k-15k	经验3-5年	本科	移动互联网/不需要融资		15-50人	"多劳多得"基

图 7 部分数据

二、数据保存

利用爬虫爬出的数据首先保存为 txt 文件然后创建数据库和数据表，利用 python 连接 MySQL 数据库把数据利用 for 循环逐行遍历存入数据库。

(一) 创建数据库

连接本地的数据库用 create database 创建数据库。采用 utf8mb4 字符集。

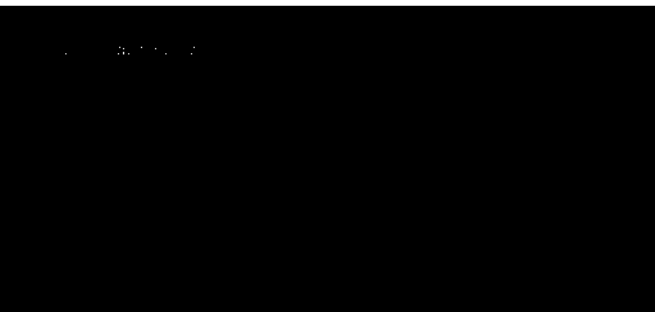


图 8 创建数据库

(二) 创建数据表

连接本地的数据库用 create table 创建数据库，将 counts 设为主键，如图 9 所示。

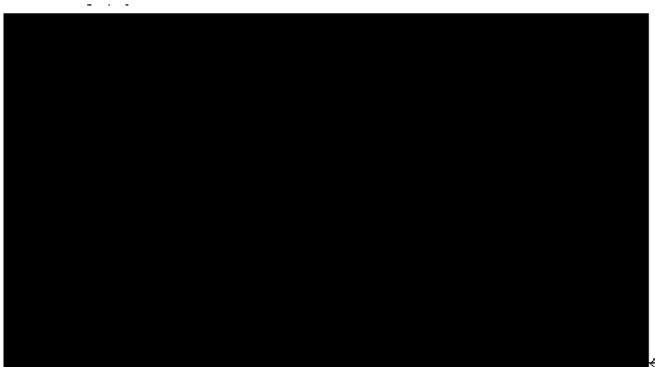


图 9 建数据表

(三) 写入数据

读取保存好的 TXT 文件，利用 for 循环逐行遍历将每一行转换成 SQL 语句格式写入数据库，如图 10 所示。

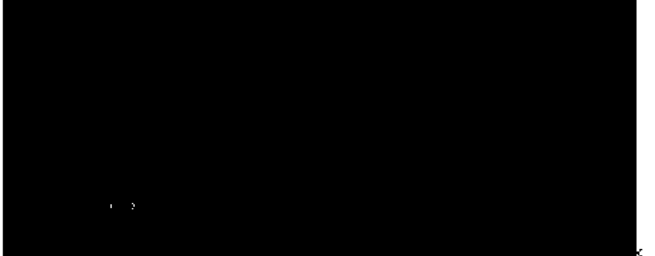


图 10 写入数据

三、数据分析

(一) 数据处理

在实际情况下，我们可以应用已经设计好的网络爬虫系统，从相应的拉勾网上抓取关于“Python”的岗位招聘信息，对工作的薪资、要求的学习和相应的工作经验需求等几个方向对岗位招聘数据进行分析，初步指导从业人员的岗位选择活动。此外，需要应用 NavicatPremium15 将 MySQL 数据库的相关信息以 csv 的后缀导出来，并且在 Python 中应用 pandas 库进行数据处理活动——

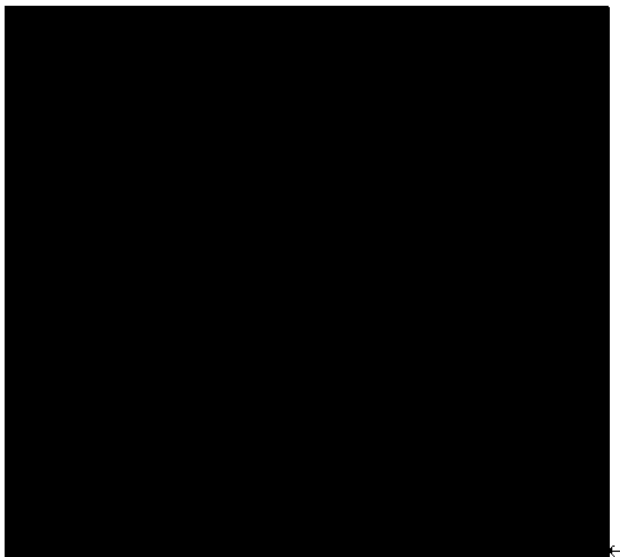


图 11 从数据库中提取数据

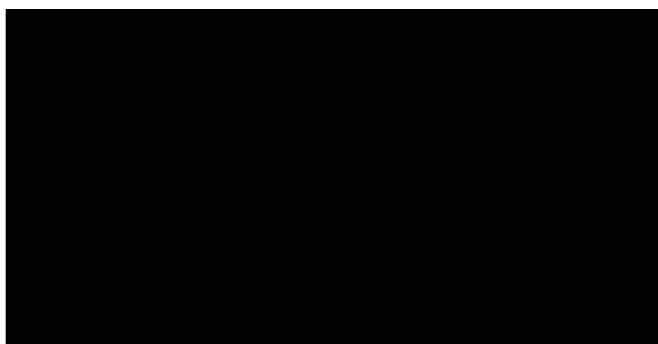


图 12 求 avgsalary

把每一列的数据提取出来构造成 address、salary、experience、education 四个列表。对于相就业数据分析工作来说，薪资分析活

动是很重要的——需要将提取出来的 salary 列表，对每个元素采用正则表达式，对前两个数字进行匹配，针对相应工作月薪的上下限，得出平均值，从而在整个列表中对薪水分布进行统计。通过系统定位和包含字符段的方式，将不同地区的相应工作以及薪资信息提取出来，并根据之前对上海不同地区的工作岗位统计方式，分析伤害地区的平均薪资水平。这样，我们也能够根据行业的岗位数量以及平均薪资水平对行业的薪资水平进行正确的计算。

(二) 数据分析结果

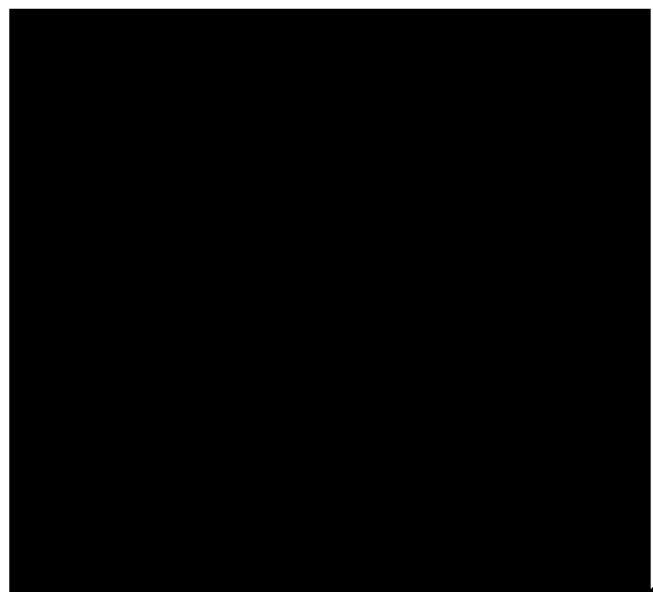


图 13 各统计图

在我国现代教育不断普及的背景下，社会就业压力的增加使得从业者们在激烈市场竞争中的选择越来越重要，如何作出最优选择也成为当前研究的重要热点问题。

1. 薪资分布图

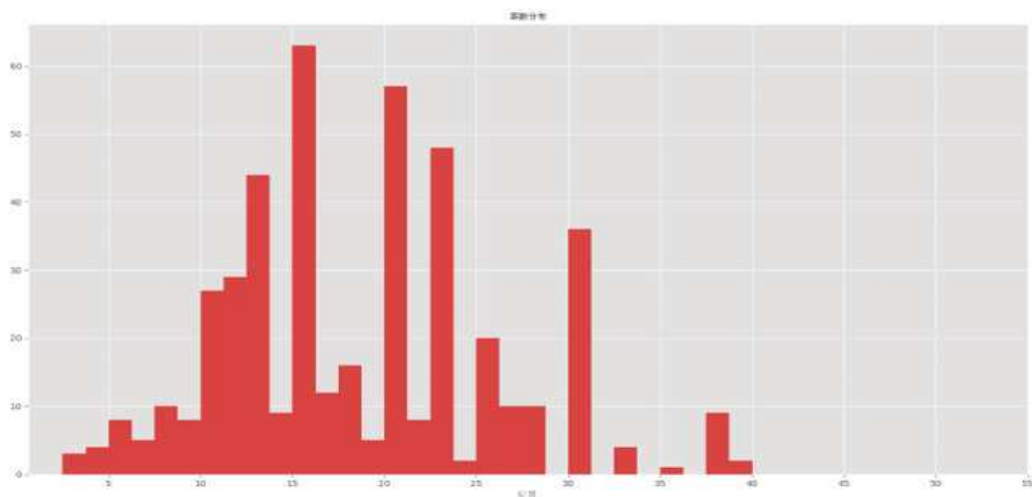


图 14 薪资分布图

从整体的薪资水平方面来看，薪资分布比较令人满意，大部分月薪集中于每月 15—25K 之间，月薪在 20K 以上和 20K 以下的

分别占据了 50% 左右，整体的收入水平较高，最高月薪为 40K。
2. 不同城市的薪酬分布情况图

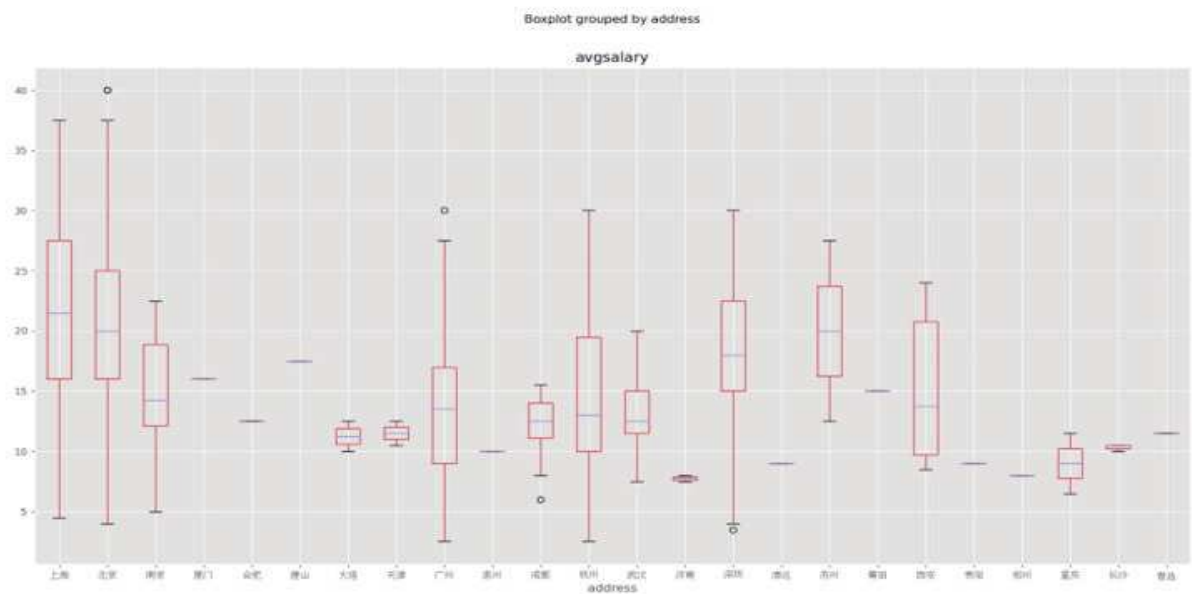


图 15 不同城市的薪酬分布情况图

从上图可以看出，在 python 行业发展过程汇总，上海市的薪酬分布的中位数大约为 22K 左右，是行业内薪酬中位数最高的地区，其次是北京、苏州中位数大约为 20k 左右，而济南中位数只大约为 8k。由此可以看出 python 的发展主要集中在北上深苏等地

区，如果想要从事 python 相关工作的去到北京和上海无疑不是一个很好的选择。下文我们将围绕北京和上海两座城市进行展开分析。
3. 北京上海工作经验不同薪酬分布情况图

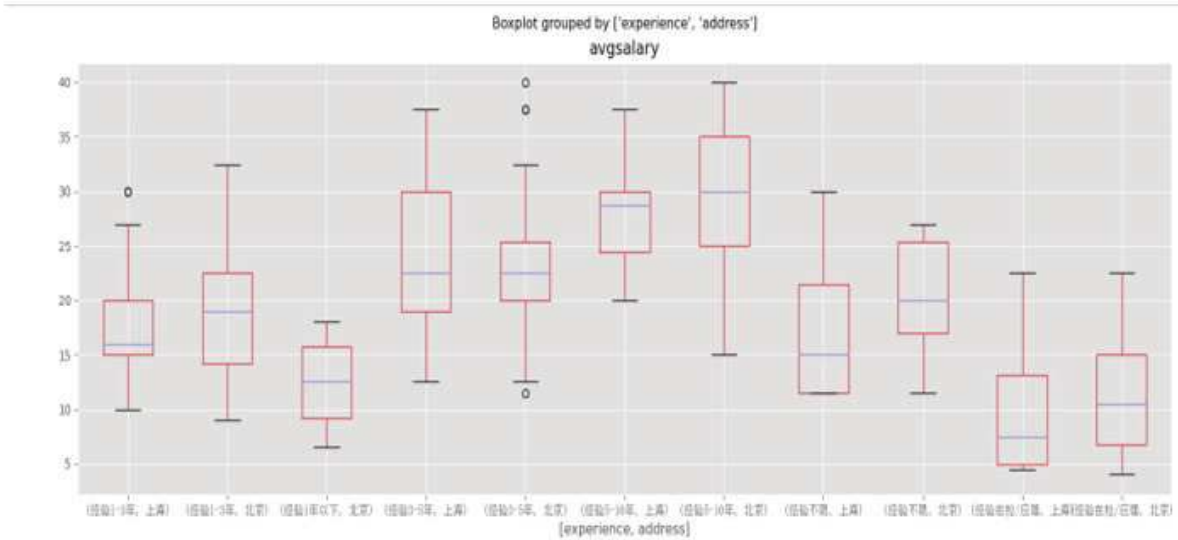


图 16 北京上海工作经验不同薪酬分布情况图

对于工作一年以下的，北京要高于上海，而工作经验丰富的北京和上海都能够得到较高的薪水。对于工作 1-3 年的人，北京

工资的中位数明显高于上海。
4. 不同学历的薪酬分布情况图

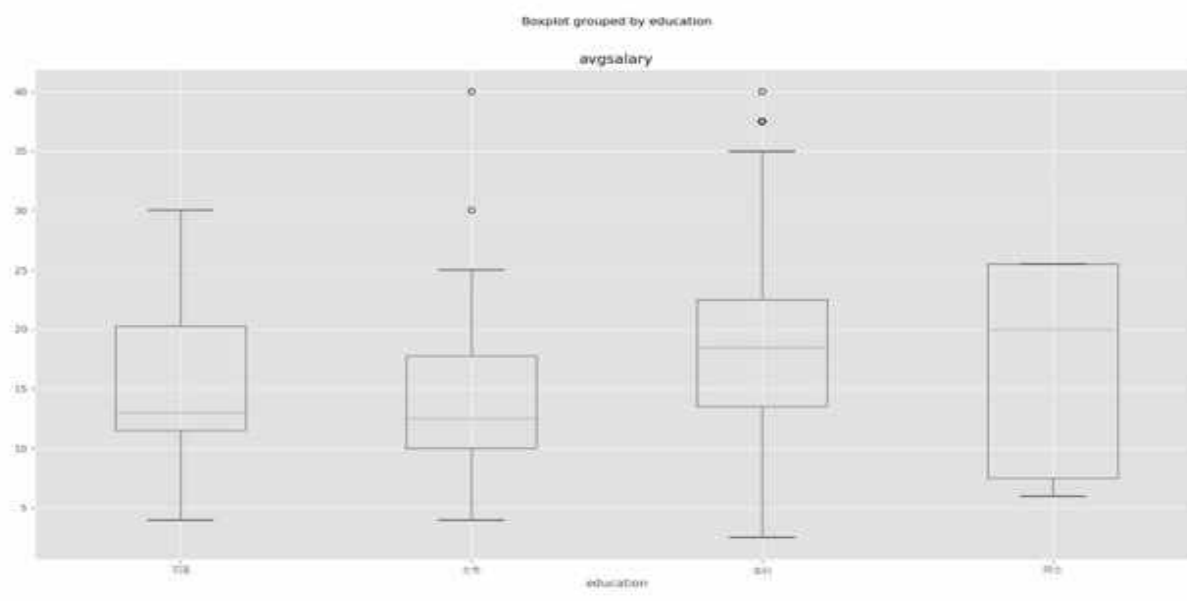


图 17 不同学历的薪酬分布情况图

不难看出 python 岗位作为一个技术性的岗位对学历要求并不是很高, 只要你技术够硬就能拿到一个不错的收入, 但是当下依然很缺少尖端人才, 只要你学历够高依旧能在开始工作的时候就拿到一个不错的工资。

5. 词云图



图 18 词云图

从图中可以看出本科学历是 python 行业的敲门砖, 当然你有丰富的工作经验和强大的技术能力也是可以的, 现在的 python 很缺少开发类的岗位。主要的领域为移动互联网。

四、结语

本文通过 Python 网络爬虫技术, 主要在拉勾网的基础上构建了一个用于数据收集和分析的网络系统, 用户可以应用该系统登录到拉勾网上, 并获取相关的页面信息, 对页面上的 URL 进行有效分析, 同时对 URL 进行数据筛选, 对用户所获取的数据进行存储, 并对这些数据进行深层次的挖掘, 通过数据分析的方式获取关于 python 岗位的相关信息, 从而为相关就业人员提供全面的信息参考。

参考文献:

- [1] 项博良, 唐淳淳等. 基于网络爬虫的就业数据分析 [J]. 计算机应用, 2017 (12).
- [2] 曹素娥. 基于爬虫技术的就业信息管理平台设计 [J]. 微机计算机信息, 2018 (4).