目　录

基于 Python 的招聘数据爬取与可视化分析

摘要：为了便于求职者和招聘单位更精准地把握当前就业市场的变化与需求，提出了基于 Python 语言的招聘信息采集分析系统，利用 Python 爬虫获取招聘信息，将数据清理并格式化后将招聘信息中的各地招聘需求情况、经验要求、工资情况、学历要求、公司规模等进行了可视化展示和分析。所得出的数据模型可以帮助求职者有效的评估出合适岗位和薪资，让求职者的求职效率得以提升,同时招聘单位也能通过此数据可视化分析模型对招聘市场的趋势进行总体把握,提高企业的竞争力。

关键词： 可视化分析；网络爬虫；数据分析；就业分析；IT 行业前景

Python-based Recruitment Data Crawling and Data Visualization Analysis

Abstract: In order to better job seekers and recruitment units to more accurately grasp the changes and needs of the current job market, this paper proposes a Python language-based recruitment information collection and analysis system, using Python crawlers to obtain recruitment information, after cleaning and formatting the data, the recruitment information The recruitment requirements, experience requirements, salary conditions, academic requirements, company size, etc. of various regions of the country were visualized and analyzed. The generated data model can help job seekers to effectively evaluate suitable positions and salaries, thereby improving job seekers' job search efficiency. At the same time, recruiters can also use this model to grasp the overall trend of the recruitment market and improve the competitiveness of enterprises.

Key words： Visualization analysis; Crawling; Data analysis; Employment analysis；IT industry prospects

1 引言

1.1 论文研究背景

随着科技的飞速发展，计算机科学专业及其他相关专业成为了全国热门专业，同时伴随着大量创新的计算机职业岗位的出现。岗位的出现必然对人才提出了充分的需求。现在大数据时代，如何在海量数据中分析并且能够从中获取到对自身有价值的数据无疑成为了一个重要的话题和研究对象。对于有些专业技能硬和综合素质高的应聘者，由于不了解就业行情以及企业招聘需求、缺乏应聘经验和策略，屡次与自己理想的企业擦肩而过，而企业也不易招聘到所需的可靠型人才。提供一个能够对其进行数据转化与可视化分析的方法，针对这一日益突出的就业问题，在这一背景下我们将应用 Python 的大数据分析与应用技术，对计算机行业招聘信息做一些比较详细的分析。

1.2 论文研究意义

将信息处理并通过可视化的方式，面向用户设计并构建了一个数据爬取、数据分析和数据可视化的系统。通过可视化分析帮助求职者更多地了解企业招聘的动向、职场信息的变化、当今社会最紧缺、最热门的技术等等。希望通过数据分析，可以帮助学生或者准备跻身于 IT 行业的学习者们明确学习动向、确立更清晰的学习目标与方向；让求职者们在职场展现风貌，提升就业竞争力；协助企业招聘者了解国内各知名企业招聘的大体趋势，以便做好招聘方向的调整[1]。

2 国内外可视化研究现状

2.1 国内的可视化研究现状

近十年以来，中国数据可视化的市场的确正处于快速增长之中。越来越多的企业认识到数据可视化在日常管理中的重要性：它可以将复杂、漫长的数据消化过程，简化到看深浅，比长短，辨趋势。国内做数据可视化的公司比起以往来说变得多起来了，本来早起都是国家和研究所，或者比较大一点的公司在对可视化分析工具注入大量成本研究，现如今广泛的研究也有了可观的成果，帆软——通用类应用工具，报表软件 finereport、商业智能 finebi、大屏可视化数字冰雹 ——大屏可视化、工业、航天等可视化，三维展现方面百度——开源图表控件 Echarts 阿里——蚂蚁金服可视化控件 AntV、数据可视化大屏 DataV 网易——数据分析平台（BI）网易有数。

2.2 国外的可视化研究现状

在 2008 年迈克尔提出数据可视化主要可以分为主题图与统计图形这两个部分。到目前，数据可视化技术发展迅速，影响与应用范围扩大了很多。现在政府和各大企业都对可视化进行了研究，关于数据可视化的国际会议也开展了很多。Microsoft、SAS 和 IBM 等知名企业在数据可视化做出了许多成就，旗下有许多优秀的可视化产品和工具。比如著名的 directX 与 OPENGL 还有 SASR Visual BI；目前制作大型三维数字场景的软件 TerraBuilder ,TerraExplorer,TerraGate，SAS 公司的商务智能软件。Skyline 软件。工具 Gephi，还有 BI 分析工具 BO、BIEE、Tableau,style intelligence 等可视化效果出众。

2 相关技术简介

2.1 Python 及相关可视化模块介绍

（1） Python

Python 是一种面向对象、解释型的脚本语言,是跨平台脚本语言[4]。Python 支持现有的各种主流操作系统。源代码免费开放、功能强大灵活，语法简洁优美。

（2）Matplotlib

Matplotlib 是 Python 的绘图库。 能够使用折线图、饼状图、柱状图、散点图等等图形可视化，它能和 PyQt 和 wxPython 等图形工具包同时使用。

（3）Pandas

Pandas 是基于 NumPy 的一种工具， Pandas 中包含的函数和方法能够处理数据。是 Python 强大而高效的数据分析模块。

（4）Pyecharts 和 Echarts

Pyecharts 是 Python 与 Echarts 相关联的可视化工具。

Echarts 是百度商业可视化团队基于 html5 Canvas 创建的最热门的可视化工具之一。提供上百个可视化图表。其中图表的表现也做的相当优秀和可观。是数据可视化分析的有力工具。

2.2 数据库 SQL Sever 相关概述

（1）SQL Sever

SQL 是结构化查询语言。SQL 语言可以用来执行数据增删查改数据库等操作[2]。SQL Sever 是关系数据库管理系统（RDBMS）。

（2）Pymssql

Pymssql 是一个简单的 Python 数据库接口，它建立在 FreeTDS 之上，为微软 SQL Server 提供接口。让 python 能够操作 SQL Sever,特点是友好的 Unicode、对 Python 3 友好、在最流行的操作系统上工作.

3 需求分析和概要设计

3.1 可行性分析

3.1.1 社会可行性分析

大数据时代的到来，让移动互联网和物联网产生了非常可观的数据，数据的收集、存储、计算、分析的问题被大数据计算技术完美解决。大数据时代让人类社会学会利用数据价值。对于网络中众多繁杂多量的招聘数据，需要做大量的简历筛选工作，需要一个能够可视化分析的系统。

3.1.2 系统可行性分析

运行环境为 windows 10 操作系统，Pycharm，使用 csv 格式的数据库，用 SQL Sever 2018,采用 pysmsql 与后台数据库相连接，完成数据存储的功能[5]，硬件方面，科技飞速发展的今天，硬件更新速度快，可靠性非常强，低廉的价格，完全能够符合运行需求。

程序分成三大模块——获取数据和处理、可视化分析，代码模块化较强，可分块使用，数据公用文本或者表格文件，数据库存取也是模块代码[6]，可分别单独使用，程序规模小可以快速运行。 本系统的爬取代码量比较简洁，分块明显，在爬取过程中能够做到速度较快，同时数据库等连接已经提前做好，所以在运行这三个模块的代码时能够很流畅的达到效果。

3.1.3 难点和解决方案

（1）爬取数据

招聘网站一般都会对爬取信息做一个防范措施，会防止信息泄漏，所以一般网站都会有反爬虫机制，所以在做爬虫的时候，怎么解决这些问题就变得很重要了。

常见的爬虫获取信息难点和方法如下：

1．无请求头，被当成是机器爬虫而拒绝访问。正常模拟 url 请求即可避免被认为是机器请求。

2．需要登录状态。Cookie 在加载页面时要正常调用和发送，避免现权限不够问题。

3．IP 被封。在一定时间中一直使用相同的 ip 访问而被禁止访问或者改 ip 被发现是异常请求所以被封。解决方法就是可以通过代理的方式模拟和设置随机间隔爬取来解决 IP 被封的问题。

（2）格式存在乱码

数据存放形式为二进制的比特流，在输出显示的时候，就要转化格式，变成人们能够识别的格式。二进制编码转化为 utf-8、gbk 等格式。

在本程序处理数据时，保存为格式文件或者数据库交互时等待都有格式转换编码的过程，以下列出了乱码错误发生可能性最大的几个地方和解决方法：

1．从网站下载源代码时。网页可能会拥有不同的格式，一般而言在网页头信息都会有所表明格式，这个时候需要在代码中指定格式去下载源代码，这样就会正常下载编码格式的源代码。

2．保存数据时。需要搞清楚代码和数据库各自的编码格式，通过转换成统一的格式解决问题。

3．通过表格文件或者或者数据库提取数据到程序中时，也是要换成一样的格式来进行读取数据就不会出现乱码的情况[9]。

3.2 需求分析

3.2.1 功能性需求分析

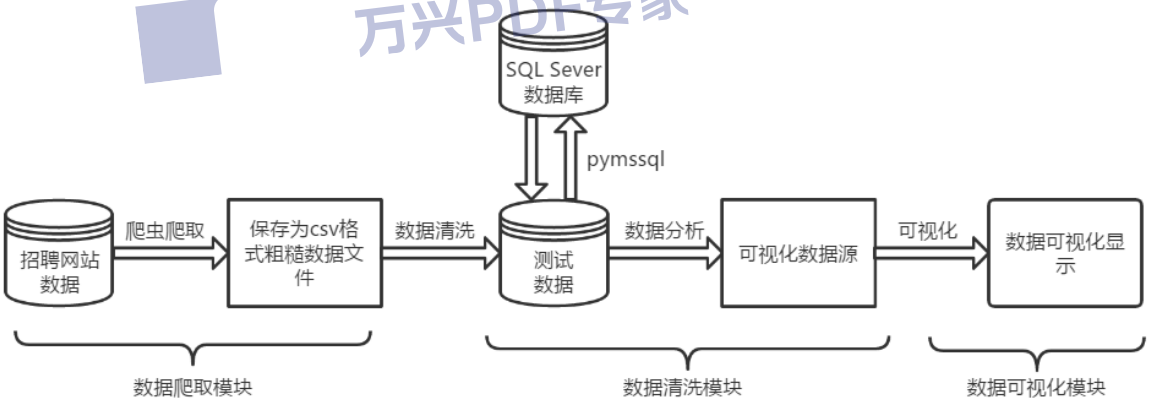

图 3-1 处理过程

因为信息技术的发展，各方面数字化程度显著提高，日常生活中的数据爆炸增长，数据对于我们来说，有着非常重要的意义，因为我们能从中获取到海量的价值信息[13]。可视化分析这一方法就是对获取有用价值信息的一种捷径。可视化筛选出有用的价值信息的过程根据图 3-1 所示其主要分为如下几个部分：

（1）数据爬取：通过 python 来对网页的源码进行解析和下载，在对源代码进行解析，提取出其中需要的信息并保存。

（2）数据清洗：在爬取下来的招聘信息中可能存在一些姓名和岗位薪水等信息错误的情况，或者在表格中发生错位、有些关键信息不全等问题、处理这一问题就是这一模块的目的。

（3）数据分析及可视化：通过用户的需求改写程序，生成分析数据，使用柱状图、折线图、云词图清晰的将数据分析后的结果以图片的方式展示。

3.3 概要设计

3.3.1 数据爬取

数据爬取与存取模块进行更详细划分为发送请求、获取响应内容、解析内容、下载数据。

模块执行流程如图 3-2：

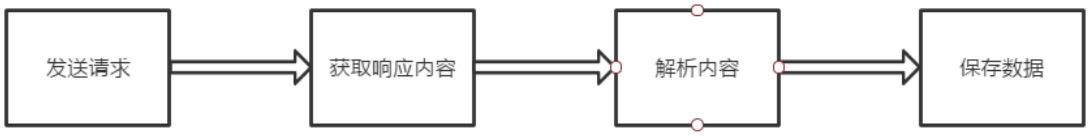

图 3-2 数据爬取流程图

（一）建立简单搜索页面

URL 通过在开发者工具中选择搜索页面，在 Headers 中可以看到 RequestURL 地址，通过对地址分析定义一个简单搜索页面 URL。

（二）翻页

翻页后 ，分析可发现 JS 中多了请求 ，通过多次切换页面找出地址拼接规律。

（三）爬取

模拟请求发送。获取链接然后取得原始数据分析后存在表格数据中。

3.3.2 数据清洗

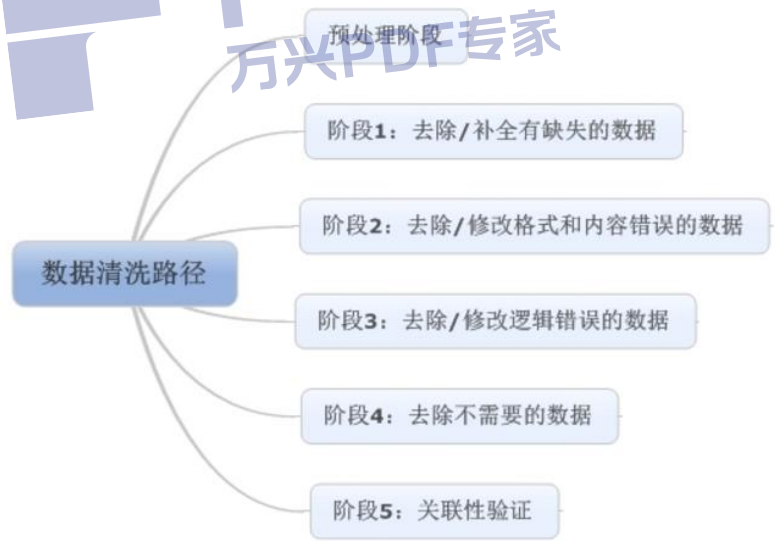

图 3-3 数据清洗过程

数据清洗大致分为图 3-3 所示的六个部分，首先是预处理阶段预处理阶段主要做两件事情：

（1）将数据源导入到 Python 数据清洗程序。

（2）人工先观察数据的构成，了解其组成部分。

第一步发现缺失的信息。

（1）找出缺失值范围：分析缺失比例和字段重要性。

（2）去除不需要的字段：这一步用代码筛选出不需要的数据行，直接删掉。

（3）填充缺失内容：通过数学分析、业务知识或经验补全缺失值。

（4）重新取数：如果缺失的数据非常重要，就需要想办法查询取得。

第二步是格式内容清洗。

（1）时间、日期、数值、全半角不同等问题，转换为需要的统一格式

（2）内容中有不该存在的字符

某些内容可能只包括一部分字符，比如数据中出现意外的空格，或者不符合当前条目的格式等。只能半自动校验半人工方式消除。

（3）内容与该字段应有内容不符

人工或者机器在填写数据发生填写错位，或者顺序填反的问题。

第三步为逻辑错误清洗，发现明显不符合逻辑的数据并清除。

（1）去重

匹配重复数据并删除

（2）去除不合理值

不合理的值通过代码筛选删除

第四步是非需求数据清洗，这一步用一段代码，输入主观需要去除的数据。

第五步是关联性验证，如果数据有多个来源，验证匹不匹配原来的数据。

3.3.3 数据分析与可视化

数据在数据分析和数据可视化两个模块之间的处理流程如下图 3-4：

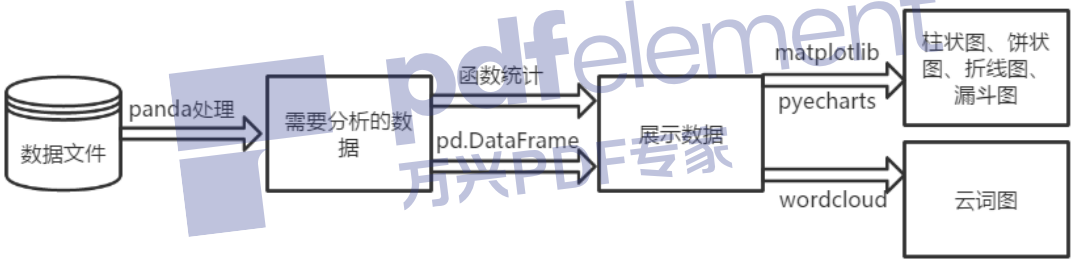

图 3-4 数据处理与可视化流程图

pandas 模块处理数据，使用 matplotlib 和 pyecharts 实现可视化分析。

（1）数据分析

Pandas 筛选出要求的数据，将数据从数据库中取出后，根据程序能够处理的格式要求等进行对原始数据的截取和筛选，最后将其中要展示的数据单独拿出来进行统计和使用。

（2）数据可视化

利用 matplotilib 和 pyechaets 库来对数据进行可视化分析[17]，其中分析了职位最多的几个城市、学历需求饼状图、几个城市的平均工资比较、工作经验要求的统计、薪水与工作经验的关系、薪水与学历的关系、薪水与公司融资情况的关系、薪水与公司规模的关系、最后，使用 wordcloud 来对指定的数据进行云词分析，其中，出现频率越高的词汇，在云词图中越居中、且字体越大，更直观的显示词频热点，图 3-5 展示了可视化模块的目的。

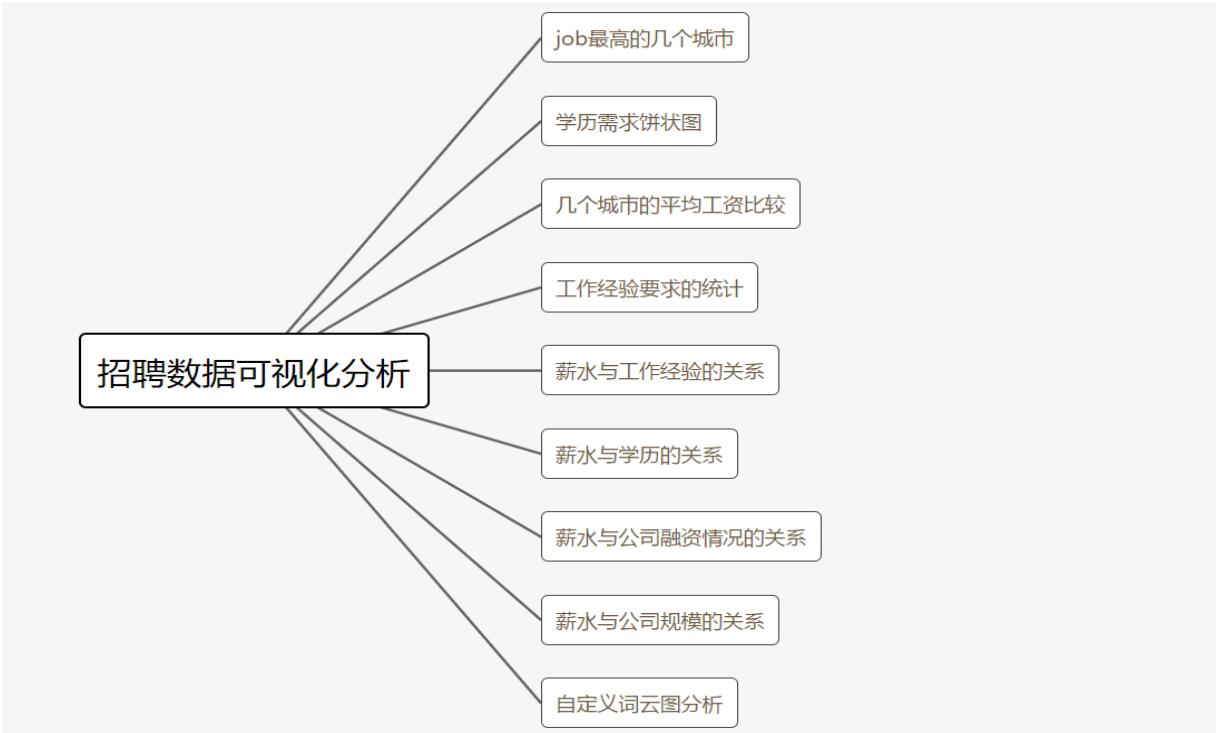

图 3-5 可视化分析目的

4 系统详细设计与实现

4.1 数据爬取模块

数据爬取主要使用库 urllib 库来获得信息通过 re 库，正则表达式来处理爬下来的信息，最后通过 xlwt 来将信息保存到 excel 文件中,算法伪代码如表 4-1 所示。

表 4-1 数据爬取模块伪代码

getdata.py

输入：需要爬取的网页信息和搜索的职位名字
输出：招聘数据表格文件
#首先模拟浏览器，使用 header 头信息指定访问信息
header={}
#翻页与网址拼接函数，page 是页数，item 是输入的字符串
def getfront(page,item):
#获取网页信息函数
def getInformation(html):
# compile 函数用于编译正则表达式 	
re.compile()
#新建表格空间
excel1 = xlwt.Workbook()

```
#新建一个 sheet,设置单元格格式
sheet1 = excel1.add_sheet()
#保存公司信息
company = re.findall(re.compile(……))
#保存招聘条件信息
job_need = re.findall(re.compile(……))
#保存福利信息
welfare = re.findall(re.compile(……))
#write 函数用于在表格中创建相应的列，根据 51job 网爬取的数据相应的的设置的序号、职位、公司名称、公司地点、公司性质、薪资、学历要求、工作经验、公司规模、公司类型、公司福利、发布时间列
sheet1.write(……)
#在爬取的过程中设置休息间隔，避免爬取海量数据时被误判为攻击，IP 遭到封禁
time.sleep(0.3)
#在程序完成指定的页数循环爬取后完成数据爬取与保存，保存为 job.xls 文件
excel1.save("job.xls")
```

运行后保存的数据如图 4-1 所示，保存在表格文件中供后续操作使用。

	A	B	C	D	E	F	G	H	I	J	K	L
1	序号	职位	公司名称	公司地点	公司性质	薪资	学历要求	工作经验	公司规模	公司类型	公司福利	发布时间
2	1	前端工程师	广东优德科	广州-海珠	民营公司	0.6-1.2万/	大专	1年经验	少于50人	互联网/电	五险一金	05-12
3	2	前端开发工	保利地产控	广州	国企	1.5-2万/月	本科	2年经验	5000-1000	房地产	五险一金	05-11
4	3	DevOps工	Ericsson (	南京-江宁	外资（欧	1.4-2.8万/	本科	2年经验	5000-1000	通信/电信/	五险一金	05-11
5	4	Web前端开	奇虎360科	上海-普陀	上市公司	2.5-3万/月	本科	3-4年经验	5000-1000	互联网/电子商务		05-11
6	5	Web前端开	星辉游戏	广州-天河	上市公司	0.9-1.2万/	大专	1年经验	500-1000人	网络游戏,互	五险一金	05-11
7	6	Web前端开	中科讯飞互	北京-海淀	民营公司	1.5-2万/月	本科	3-4年经验	5000-1000	计算机软件	五险一金	05-11
8	7	前端开发工	湖北航天信	武汉-东西	国企	0.8-1.4万/	本科	5-7年经验	500-1000人	计算机服务	五险一金	05-11
9	8	WEB前端	上海文思海	异地招聘	合资	1.5-2万/月	本科	5-7年经验	10000人以	计算机软件		05-11
10	9	高级前端开	深圳市云中	深圳-南山	民营公司	2-2.5万/月	本科	5-7年经验	50-150人	互联网/电	周末双休	05-12
11	10	web前端实	深圳市马丁	深圳-南山	民营公司	3-7千/月	大专	在校生/应届生		计算机软件		05-11
12	11	月薪15K前	迪昆集团	广州-黄埔	民营公司	1-1.5万/月	本科	2年经验	150-500人	专业服务(	五险一金	05-11
13	12	Web前端工	宁波捷创控	宁波	上市公司	0.8-1万/月	大专	1年经验	150-500人	多元化业务	五险一金	05-11
14	13	高级前端开	（CCE GF	上海	民营公司	1-1.5万/月	本科	3-4年经验	150-500人	广告,公关/	五险一金	05-11
15	14	Web前端开	深圳市红与	深圳-龙华	民营公司	1-1.5万/月	大专	1年经验	150-500人	计算机软件		05-11
16	15	Java前端开	绍兴缔华丝	绍兴	民营公司	0.8-1万/月	大专	3-4年经验	150-500人	服装/纺织/	专业培训	05-11
17	16	Web前端开	上海澄码信	上海-浦东	民营公司	1-2万/月	本科	2年经验	50-150人	计算机服务	餐饮补贴	05-11
18	17	Web前端开	北京腾信软	上海-闵行	上市公司	1-1.5万/月	本科	3-4年经验	500-1000人	计算机软件	定期体检	05-11
19	18	前端工程师	武汉市迅易	武汉-江夏	民营公司	0.8-1万/月	大专	1年经验	少于50人	计算机软件,通信/电信		05-11
20	19	Web前端开	哈尔滨上元	哈尔滨-南	民营公司	0.8-1万/月	招5人	5-7年经验	50-150人	计算机软件		05-11

图 4-1 爬取的粗糙数据保存在 excel 文件中

4.2 数据清洗模块

数据清洗模块主要运用 pandas 与正则表达式来清洗数据。使用正则表达式来统一数据格式。以下表 4-2 是核心代码：

表 4-2 数据清洗模块伪代码

Datacleaning.py
输入：爬取的粗糙招聘表格数据

输出：清洗后的招聘表格数据

```
# 读取数据
data = pd.read_excel(r'job.xls',sheet_name='Job')
#将数据总表模板 加载为 pandas 二维表格式，以供处理
result = pd.DataFrame(data)
#统一薪资单位为万/月，用正则表达式来匹配并修改
x = re.findall(r'\d*\.?\d+',li3[i])
#转换成浮点型并保留两位小数
format(float(x[0])/12,'.2f')
#清除指定条件的脏数据函数，
函数 def clean(a,b,c):
# 删除整行操作
a = a.drop(i, axis=0)
#保存清洗后的数据为表格文件
a.to_excel('job2.xls')
```


4.3 数据库存取

连接好数据库后，使用本代码将数据保存到 SQL Sever 中。核心代码如表 4-3 所示。

表 4-3 数据库存取伪代码

PYSQL.py

输入：清洗后的招聘表格数据

输出：数据库文件

```
#连接数据库
conn=pymssql.connect(server='RESCUER-K',user='sa',password='765429192',data
base='andata')
#打开要保存的 excel 文件的表格
chart = xlrd.open_workbook("job2.xls")
```

```
#使用 sql 语句将数值存入数据库中创建好的相应的列
sqlNonQuery = "insert into job values（……）"
#关闭数据库
conn.close()
```

数据库中部分数据展示如图 4-2：

图 4-2 数据库结果图

4.4 数据分析与可视化

4.4.1 Pyecharts

为了让数据以一种更加直观，清楚的形式展现出来，这里运用 python 自带的绘图库 Pyecharts 对每个网站的 每个职业绘图。Pyecharts 生成 Echarts 图表。系统中主要调用了 pyecharts 库中的 Geo 模块来绘出职业对应工作地点的需求，能直观地看出哪些城市/区域的职位更多，就业机会更大。本次分析设计使用了饼状图、地理分布图、漏斗图。饼状图设计伪代码如表 4-4 所示。

表 4-4 饼状图设计伪代码

饼状图设计
输入：筛选的学历数据
输出：学历需求饼状图
#统计岗位中各学历要求的个数
def get_edu(list):
#将岗位中各学历要求的个数绘制成饼图
pie.add(parameter)
#输出饼状图
pie.render(' ')

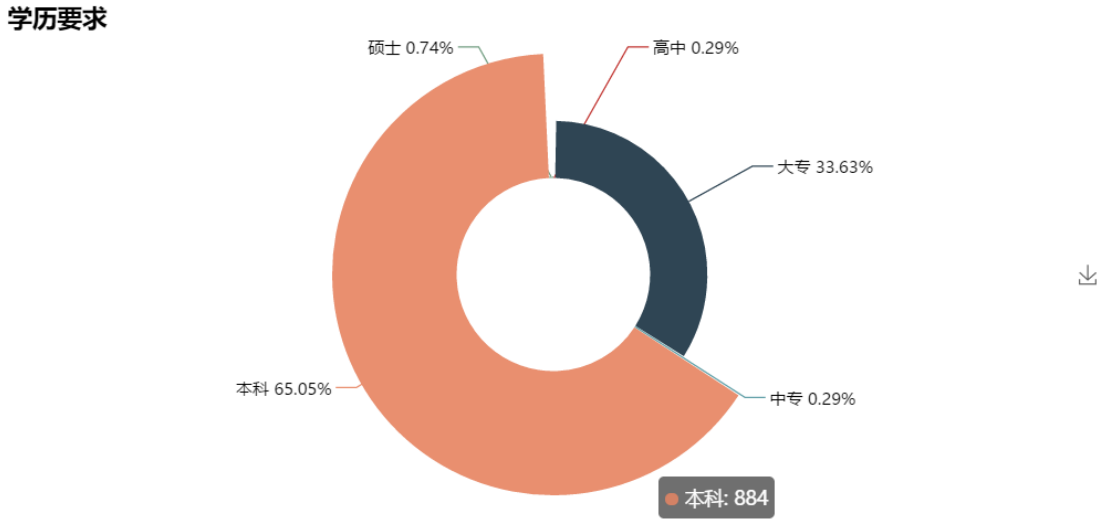

图 4-3 学历要求饼状图

输出图片 4-3 对学历统计，然后是地理分布图的设计，设计代码如表 4-5 所示。

表 4-5 地理分布图设计伪代码

地理分布图设计

输入：筛选的地区招聘需求数据
输出：人才需求分布图
#统计各地区岗位个数
def get_address(list)
#将各地区岗位个数绘制成地理坐标图
geo=Geo()
#添加数据进图
geo.add()
#输出人才需求分布图
geo.render('人才需求分布图.html')

图 4-4 岗位需求地理分布

运行代码后输出图片 4-4 岗位需求地理分布，最后是漏斗图的设计，设计代码如表 4-6 所示。

表 4-6 漏斗图设计伪代码

漏斗图设计

输入：筛选的学历需求数据
输出：工作经验漏斗图
#统计各学历要求岗位个数
def get_experience(list):
#创建漏斗图
funnel = Funnel()
#设置漏斗图参数
funnel.add()
#工作经验要求漏斗图
unnel.render()

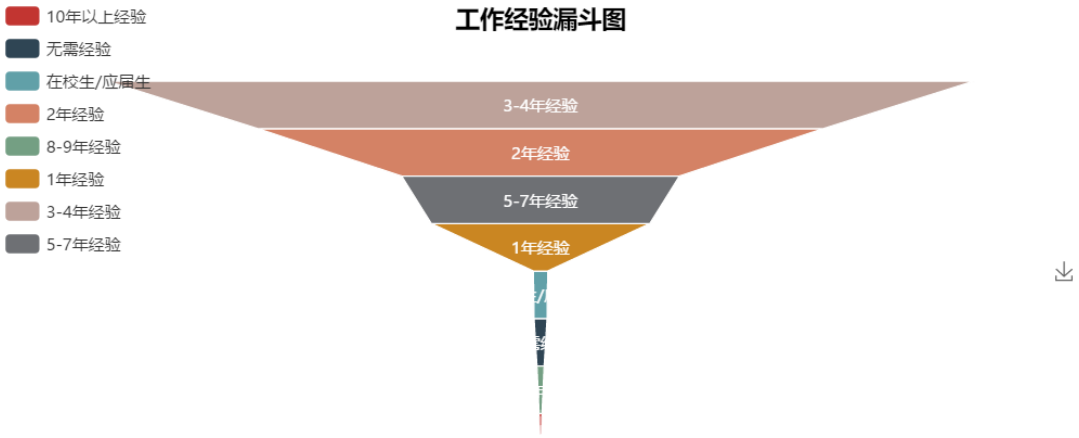

图 4-5 工作经验漏斗图

输出的经验漏斗图如图 4-3 所示。

4.4.2 Matplotlib

这个模块分析用到的是 pandas 模块，实现数据可视化用到的工具是 matplotlib。柱状图的设计算法如表 4-7 所示。

表 4-7 柱状图设计伪代码

柱状图设计

输入：筛选的处理数据
输出：可视化柱状图
#获取每个元素出现的次数
def all_np(arr):
#设置中文显示
plt.rcParams['font.sans-serif'] = 'SimHei'
#横坐标数值输入
x = []
#纵坐标数值输入
y = []
#设置柱体颜色
plt.bar()
#设置标题
plt.title()
#设置横坐标
plt.xlabel()
#设置纵坐标
plt.ylabel()
#输出柱状图
plt.show()

饼状图设计算法如表 4-8 所示。

表 4-8 饼状图设计伪代码

饼状图设计

输入：筛选的处理数据
输出：可视化饼状图
#设置中文显示
plt.rcParams['font.sans-serif']='SimHei'
#将画布设定为正方形，则绘制的饼图是正圆
plt.figure(figsize=(8,8))
#定义饼图的标签，标签是列表
label=[']
#设定各项距离圆心 n 个半径
explode=[]
#各个扇区的值
values=[]
#绘制饼图
plt.pie()
#绘制标题
plt.title()
#输出饼状图图
plt.show()

5 可视化分析结果

第一步，我们首先做了对各个城市职位需求量的统计，得到了存放相关数据的表格，matplotlib 进行展示， 需求量越大的城市机遇多且经济发达，能够明确自身发展方向。得到的数据可视化图片如图 5-1 所示。

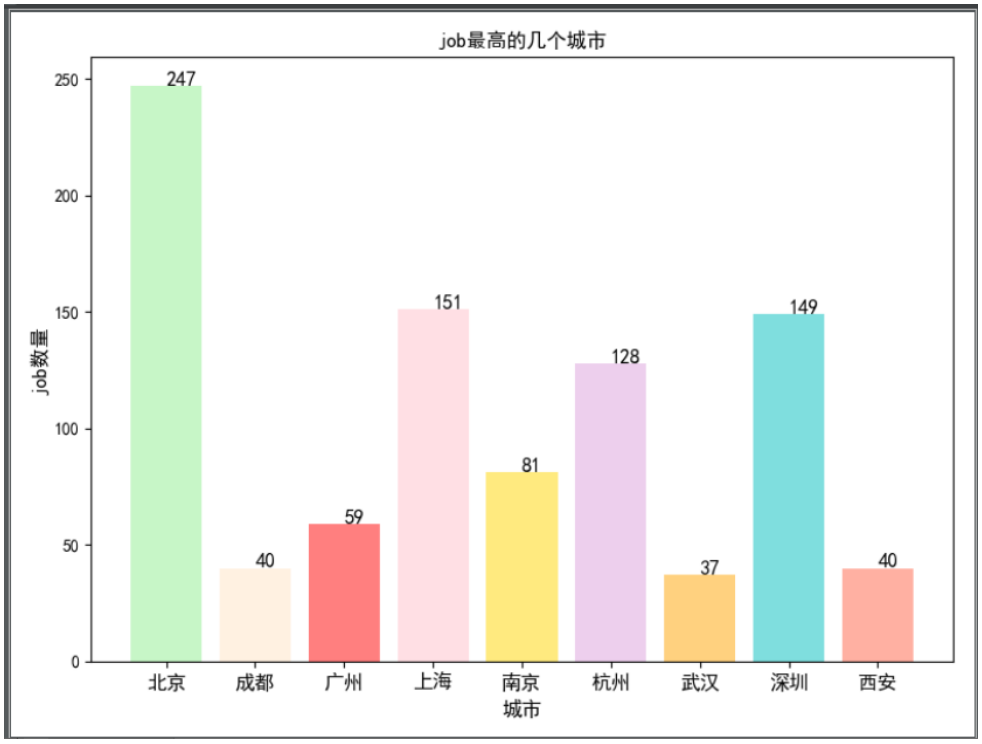

图 5-1 城市对岗位的需求量

观察图表可以分析，北京、上海、深圳招聘岗位比其他都要多，需求能够更多的接纳 IT 求职人员。杭州的招聘量需求也非常多这也说明了发展潜力大，在杭州发展前景相较于其他城市而言更好。其余几个城市是中等需求量，虽然没有北上深城市那么热门，但是能够提供非常好的起步。一开始适合在自己熟悉的城市起步，不一定要去北京、深圳等城市发展，日后发展起步之后在转战北京深圳上海等大城市。之前分析的图 4-2 岗位需求地理位置的分布分析，大多数 IT 职业岗位都分部在西南部。尤其广州、上海、深圳等科技发达的沿海城市，对计算机类人才的需求还是很多的。

再来看看招聘中学历的要求如图 5-2 所示。

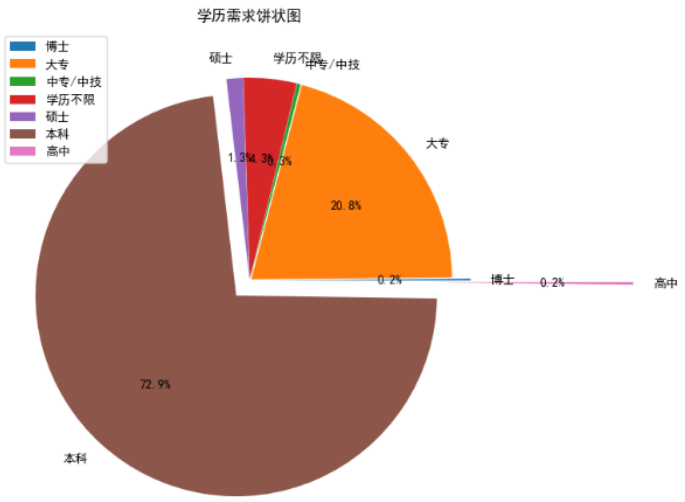

图 5-2 学历需求饼状图

本科和大专占绝大多数。学历是招聘的门槛看得出一般而言都需要大专以上学历。硕士、博士的需求量不大，对其需求量可能是一些特殊要求的职位或者高要求的职位。本科需求量最大，本科生相较于其他而言能力并不算弱，薪资方面需求比硕士或者博士低，用人单位考虑实际情况可能更愿意招聘，当然大专、不限在这里也可以看到需求量也是很大的，说明有能力对于求职来说更重要。

对于工作经验，这点特别的重要，如果有了工作经验，那对于你寻找工作、跳槽都是一个最有力的条件。本次对岗位对工作经验的需求做了一个统计，柱状图如图 5-3 工作经验要求的统计所示。

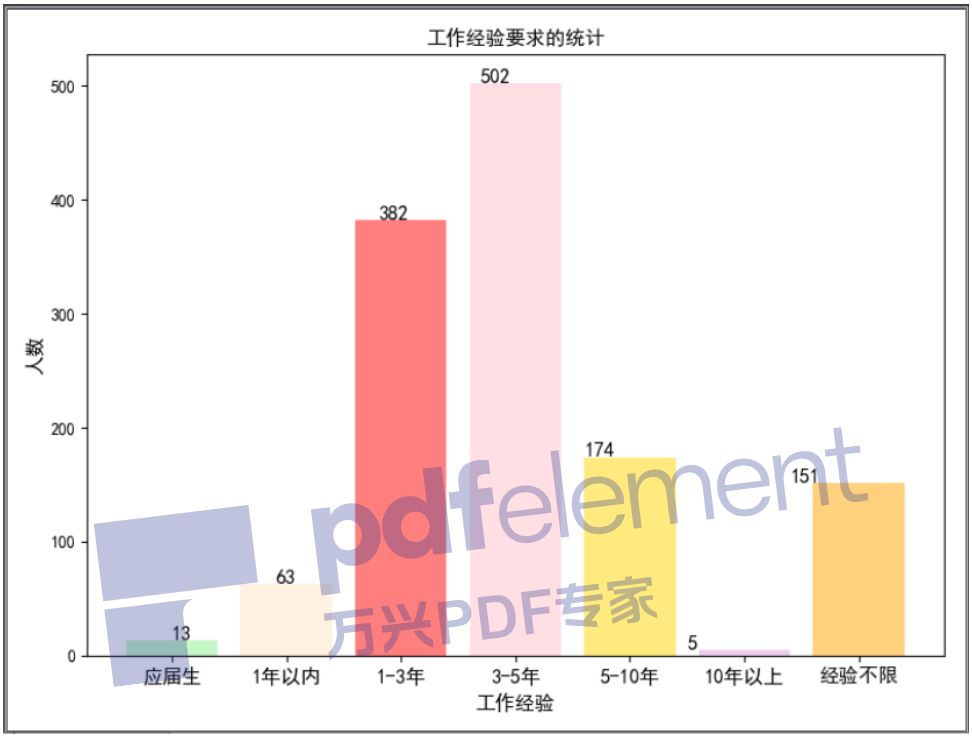

图 5-3 工作经验要求的统计

分析表中数据可以看出 1 到 5 年岗位需求占比非常多，岗位更看重经验。因为大多的招聘都是面向社会的，在经验折中的情况下企业企业更愿意招揽培养成本少又有一定经验的人群。公司希望招聘到的人马上就能开始运作经验过少的求职者培养需要更大的投入成本，而且一般而言公司希望招聘到的人马上就能开始运作，5-10 年的经验招聘成本比较大所以需求也不多。应届毕业生相比占比比较少。

薪水是重要岗位吸引指标。通过分析各城市的薪资分布了解各热门城市的平均薪资，可以选择未来工作的地点，通过分析也可以得知自己的薪水在同行中同行中处于哪一个层面。 这里统计了一部分比较热门的城市。柱状图如图 5-4，成都的平均薪资也是在 10k 左右处于最低，最高的是北京，大约在 31k 左右。

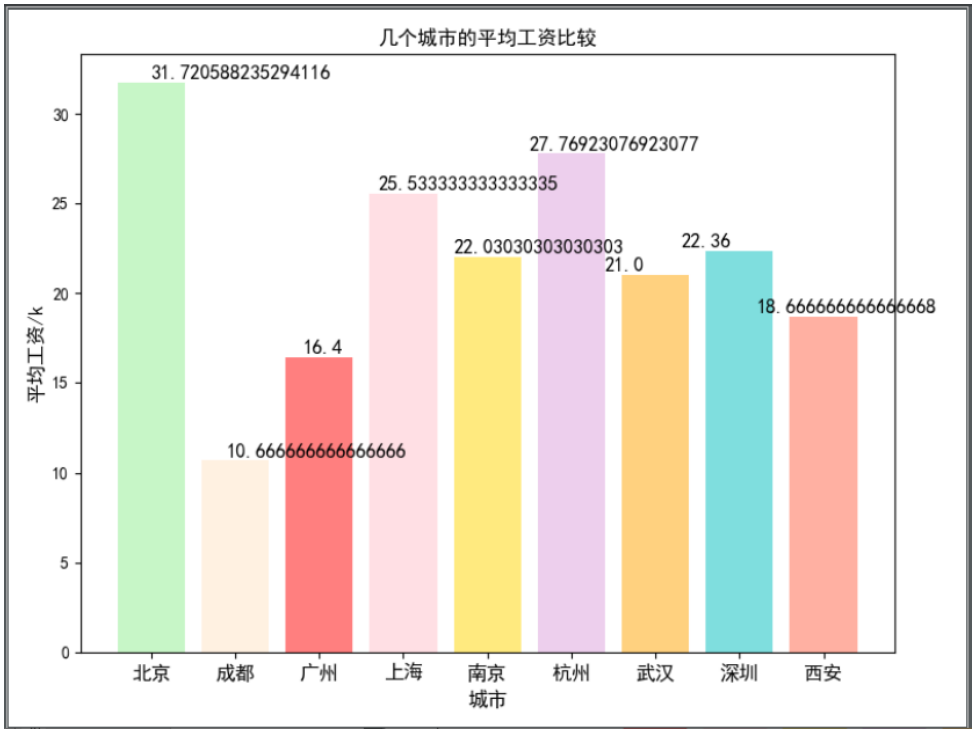

图 5-4 城市平均工资柱状图

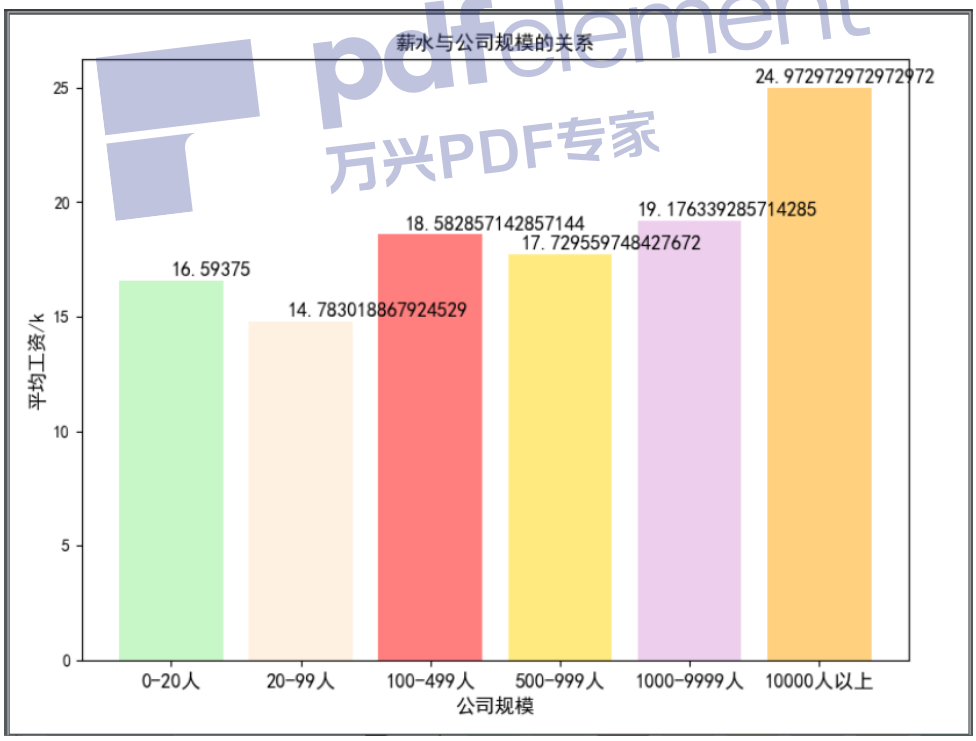

图 5-5 薪水与公司规模的关系

公司规模与薪资分析如图 5-5 所示，理论上工资与公司规模应该是成正比。显示了薪水与公司规模的关系柱状图在 100-9999 人规模的公司平均工资在 18k-20k 左右，0-20 人的公司大约在 16k，确实规模越和公司待遇成正比，相对的能力要求也成正比。下面我们在来分析一下工作经验能带来多少经济效益。

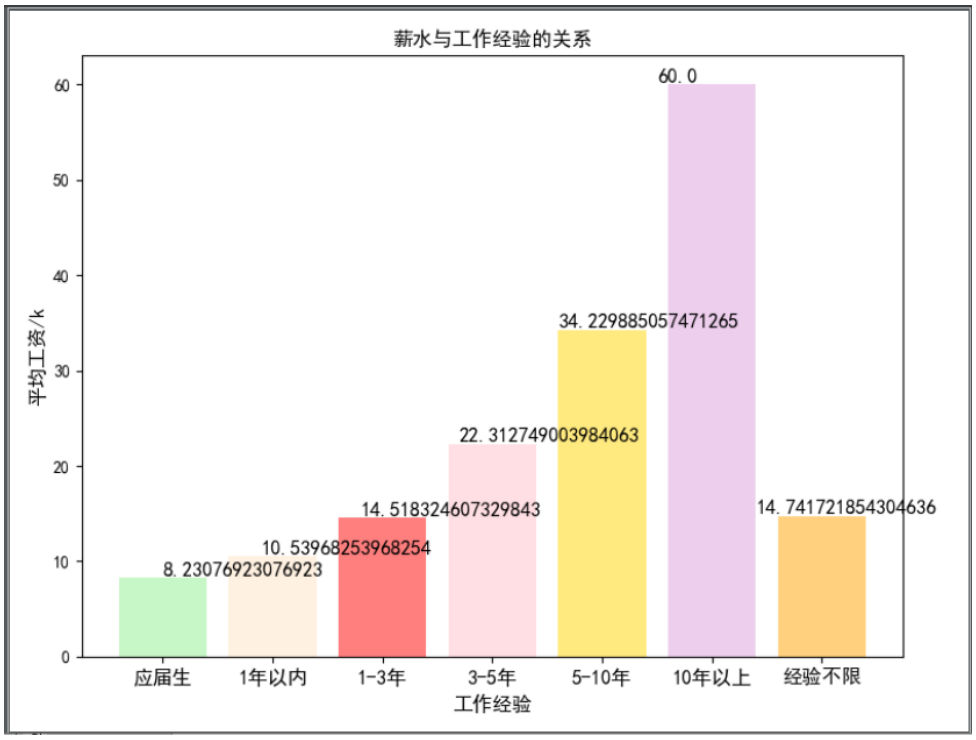

图 5-6 薪水与工作经验的关系

明显的，工作经验越长，那么薪资也越多，这也是取决于工作经验能给公司带来的收益，说明拥有工作经验是非常重要的。

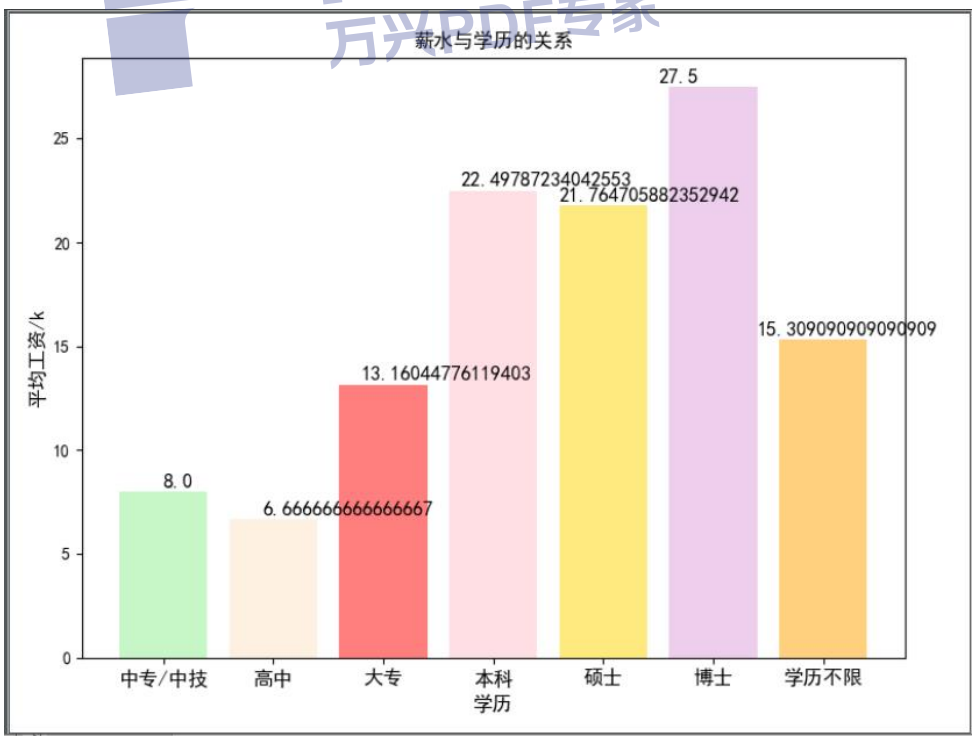

图 5-7 薪水与学历的关系

通过对图 5-7 薪水与学历的关系分析不难发现，学历越高，平均薪水也就越高，而前面对学

历需求发现，本科学历需求最大，而对学历不限的人招聘的一般最多的也是本科学历，所以学历不限接近于本科学历的平均薪资。

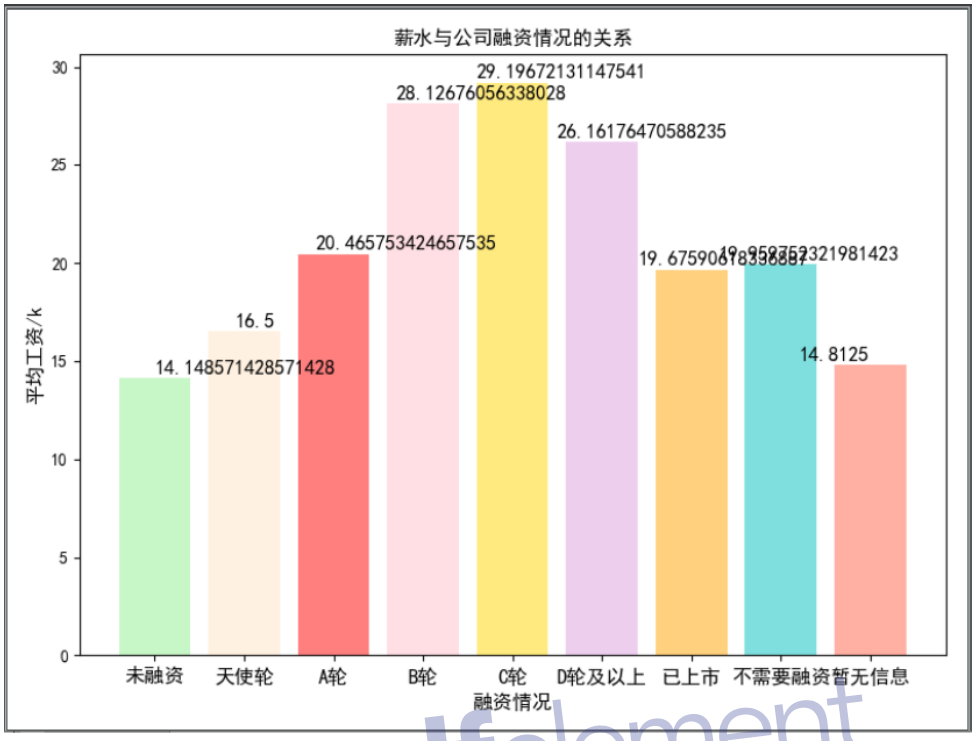

图 5-8 薪水与公司融资关系

由图 5-8 薪水与公司的融资关系可以看出，不同融资阶段公司平均薪资略有不同，D 轮、C 轮、B 轮平均工资相对较大一些。

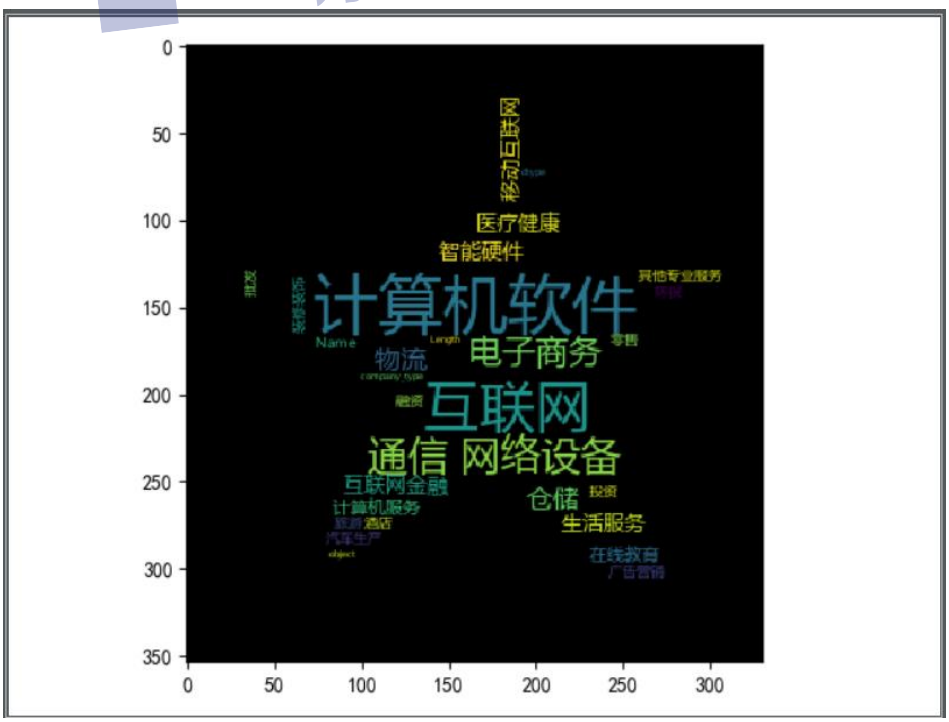

图 5-9 职位领域词云图

通过上图 5-9 职位领域词云图分析 IT 行业时下热门就业方向：

（1）计算机软件。软件工程师主要进行软件前期的项目需求的分析，软件的开发和和测试。

（2）互联网，移动端的发展迅速和 5G 时代来临，是一个移动互联网的又一高峰，其相关行业也处于快速发展阶段。互联网领域绝对是经济发展的重要组成。

（3）通信、网络设备等，在硬件方面各公司也都有着一定的需求，网络的维护，硬件的安装和维护，都提供了很多就业岗位。

随着 IT 业的发展，行业的竞争也逐渐加剧。 弄清热门行业与走势，对就业和未来也有着重要的影响。

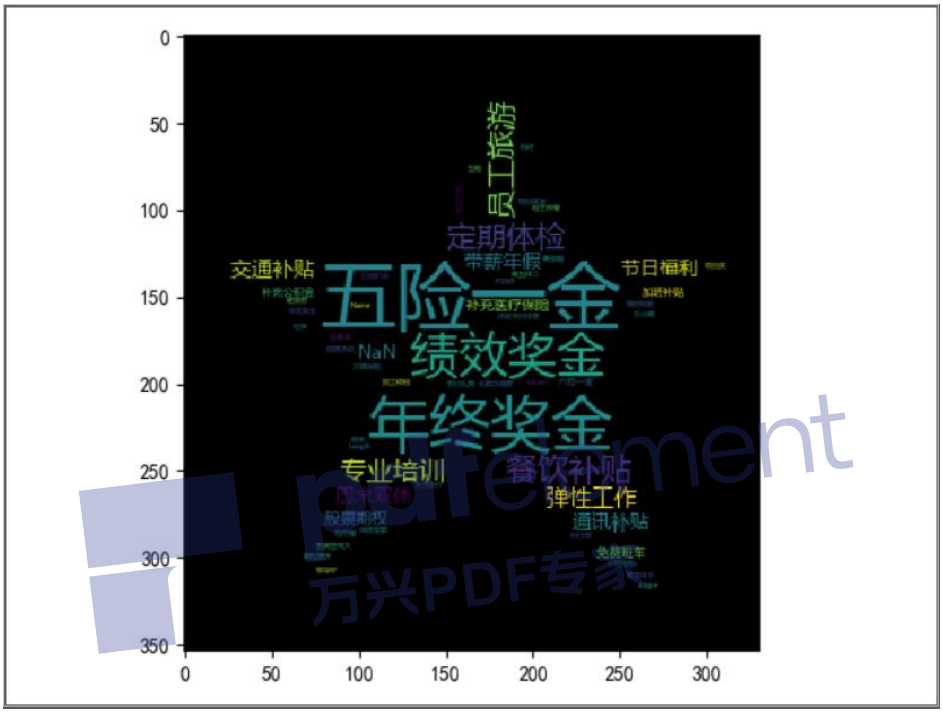

图 5-10 公司福利词云图

通过对图 5-10 公司福利词云图的分析得知大部分企业都会拿“五险一金”、“ 年终奖金”等福利待遇来吸引求职者，而这些福利恰好也是大部分求职者渴望得到的。

总而言之，通过采集国内计算机相关招聘信息，并对这些数据并进行一系列分析，从各方面把我国当前计算机行业就业情况做了一定的了解和汇报，而得到的结果总体而言符合实际虽然我们进行数据分析的方案和做法还存在很多不足，但这一过程中我们获益匪浅，进步了很多，我们仍会继续努力，力求统计更具规模性、多样性、及时性的数据，采用更加有效的分析方式，研究更高性能、更深层次的数据挖掘算法，这样才能总结、展示出更加真实、有效的分析成果[19]。

6 结论

基于 Python 的招聘信息爬取与可视化分析是互联网领域的应用典型。在信息技术发展飞速的今天，所有行业都在推进数字化，海量数据超越了以往任何时代。可视化技术能够非常有力的协助人们对数据的理解和分析，让数据中的规律和联系直观的体现出来。目前开源可视化工具如 D3.js, Echart, Processing 等虽然提供了可视化图表的展示功能，但是需要用户自己来完成数据的获取、数据清洗、数据处理分析、应用于可视化分析等[20];同时，面对大数据，处理需要有一定的条件和能力，此时通过可视化技术协助用户探索大数据中有价值的信息。通过爬虫技术从

各大招聘网站获取海量信息，分析招聘数据，最后数据清理并格式化后 进行可视化展示。该系统可以有效的展示各地区热门岗位统计情况，各求职岗位的工作地点分布情况，各地区岗位薪资数据对比情况分析等等，根据各数据模型求职者和求职单位可以根据自己不同的需求获取相应信息并对求职市场进行总体把握，从而做出正确决策。