

**本科毕业设计说明书**

基于selenium爬虫的招聘信息数据分析与多维度可视化应用

Recruitment information data analysis and multi-dimensional visualization application based on selenium crawler

学院（部）： 数学与大数据学院

专业班级： 大数据19-1

学生姓名： 王珲

指导教师： 袁旭东 讲师

2023年5月25日

基于selenium爬虫的招聘信息数据分析与多维度可视化应用

# 摘要

随着我国企业信息化发展的大力推动,大数据分析时代到来,形成了对招聘网站数据信息的巨大需求，各大招聘网站所发布的岗位需求纷繁杂乱，招聘专业和要求等更是层出不穷，传统的浏览页面方式不能简单直观的给予求职者所需要的信息。高速、安全、健壮、分布式的爬虫过程已经呼之欲出，对于由爬虫程序所获得数据信息进行处理与分析，可以更迅速地得到所需求的信息。这篇文章就是讨论招聘信息的数据处理、可视化及多元回归预测。

在文章里，我们通过对招聘网站全国范围内的信息进行爬取，搜集了大量与数据岗位有关的招聘信息，进行数据处理后，对招聘信息中的岗位、城市、薪资、学历要求、经验等字段多维度展开比较并进行可视化呈现，以及建立数据多元回归建模预测薪资。最后为学校大数据人才如何培养与社会人才需求匹配，从而可以为学生提供良好职业预期发展。

**关键词：爬虫，数据分析，数据可视化，多元回归分析**

Recruitment information data analysis and multi-dimensional visualization application based on selenium crawler

# **ABSTRACT**

With the strong promotion of enterprise informatization development in China and the arrival of the era of big data analysis, there is a huge demand for recruitment website data information. The job requirements posted by major recruitment websites are complex and chaotic, and recruitment majors and requirements are constantly emerging. Traditional page browsing methods cannot simply and intuitively provide job seekers with the information they need. A high-speed, secure, robust, and distributed crawling process is already emerging, and processing and analyzing the data information obtained by the crawling program can quickly obtain the required information. This article discusses the data processing, visualization, and multiple regression prediction of recruitment information.

In the article, we collected a large amount of recruitment information related to data positions by crawling information from recruitment websites nationwide. After data processing, we compared and visualized the multi-dimensional fields such as position, city, salary, educational requirements, experience, etc. in the recruitment information, and established data multiple regression modeling to predict salary. Finally, how to cultivate big data talents in schools that match the needs of social talents can provide students with good career prospects for development.

**KEYWORDS：crawler, data analysis, data visualization, multiple regression analysis**

目录

[摘要（中文） i](#_Toc26522)

[摘要（外文） ii](#_Toc8846)

[1网络爬虫技术 2](#_Toc25596)

[1.1 网络爬虫概念 2](#_Toc4162)

[1.2 selenium 2](#_Toc9076)

[1.3 关键技术模块 3](#_Toc9409)

[2信息爬取与解析 5](#_Toc9834)

[2.1 爬取目标 5](#_Toc1355)

[2.2 爬虫流程分析 5](#_Toc17040)

[2.3 爬虫解析分析 7](#_Toc29497)

[3数据预处理 8](#_Toc2752)

[3.1数据介绍 8](#_Toc4194)

[3.2重复值处理 8](#_Toc8807)

[3.3岗位名处理 9](#_Toc5133)

[3.4工资数据处理 9](#_Toc22191)

[3.5工作地点处理 10](#_Toc23548)

[3.6发布日期处理 11](#_Toc1562)

[3.7招聘专业处理 11](#_Toc32161)

[3.8公司规模字段的处理 12](#_Toc22535)

[3.9构造新数据 12](#_Toc20834)

[4数据分析与可视化 13](#_Toc17534)

[4.1热门招聘岗位 13](#_Toc8141)

[4.2热门招聘城市 14](#_Toc14694)

[4.3招聘岗位的薪酬分布 14](#_Toc19901)

[4.4招聘岗位对学历的要求 15](#_Toc582)

[4.5招聘岗位对工作经验的要求 16](#_Toc4399)

[4.6招聘公司所属行业的分布情况 17](#_Toc7000)

[4.7公司类型占比统计 17](#_Toc25316)

[4.8不同公司规模的公司数 18](#_Toc31258)

[4.9热门城市工资比较 19](#_Toc19226)

[4.10热门招聘公司地理位置 19](#_Toc13222)

[4.11学历和工资的关系 20](#_Toc18700)

[5数据多元回归建模预测薪资 21](#_Toc13995)

[5.1多元线性回归 21](#_Toc21384)

[5.2提取相关字段 22](#_Toc15924)

[5.3字符串函数处理 22](#_Toc8984)

[5.4建立模型 23](#_Toc2978)

[结论 24](#_Toc14994)

[参考文献 25](#_Toc10631)

[致谢 27](#_Toc2317)

1网络爬虫技术

1.1 网络爬虫概念

随着大数据时代的到临，网络爬虫在互联网中的地位将愈加重要。互联网中的数据是海量的，如何自动高效地获得我们感兴趣的信息并加以利用，是一个重要的问题，而爬虫技术便是为了解决这些问题而产生的[1]。

针对不同的需求，我们所感兴趣的信息也会有所差异。如果我们仅仅是使用搜索引擎，我们会需要尽可能多高质量的网页[2]；但如果我们需要获取某一特定领域的数据或者有明确的检索需求，我们就需要针对性地搜索。此时，我们所感兴趣的信息是基于我们具体的检索和需求而定位的信息，但是我们需要过滤掉一些无用的信息。一种为前者是通用网络爬虫，另一种为后者则是聚焦网络爬虫[2]。

网络爬虫技术是一项在网页数据爬取和分析方面的高级技术。它广泛应用于多种领域，例如搜索引擎、科学研究、市场调查和产品研发等方面[3]。在现实生活中，普通用户通常使用搜索引擎如360、百度和Google等，以获取所需的网络数据信息。对于不同科学领域的专业研究而言，大数据信息挖掘、机器学习以及数字图像处理等领域都需要网络爬虫技术获取数据。

此外，在网络购物和产品研发过程中，网络爬虫技术也可以帮助开发者采集和比对不同商品的市场价格，提供优质且高性价比的产品给用户[4]。最后，在网站数据传输和用户访问舆情监管方面，网络爬虫技术可以通过抓取和分析网站数据源和内容来检测和识别网络中存在的浏览器和服务器攻击，并判断某个用户是否在进行非法访问[5]。总之，Python编程语言实现的网络爬虫技术可以被广泛运用于多种网站和网页数据的搜集及分析，在避免外来攻击和保证数据安全方面发挥着至关重要的作用[6]。

## 1.2 selenium

本质上，Selenium是一项通过驱动浏览器来完全模拟浏览器操作的技术。通过模拟人类操作，例如跳转、输入、点击、下拉等行为，可以获取经过网页渲染后的结果。借助预设好的逻辑代码，也可以自动执行这些操作。

实际上，Selenium就像一个可以自动化使用浏览器访问目标站点并操作的工具。这也意味着，我们可以利用Selenium技术来实现爬虫任务。在Selenium中，我们可以通过操作webdriver等工具，以便进入浏览器设置和启动Chrome浏览器。

1. 优点

Selenium具有明显优势，因为它可以绕过一系列复杂的通信流程，如requests模块。使用requests模块模拟请求时必须先分析所有的通信流程才能发出请求并返回响应。但如果目标站点有许多复杂的通信流程，如登录时的滑动验证等，使用requests模块时就会特别麻烦。此外，requests请求库无法执行js，因此如果网站需要发送ajax请求，并异步获取数据以渲染页面，则需要使用js发送请求。与此不同的是，浏览器直接访问目标站点并获取数据，然后将其渲染到页面上，因此，Selenium的抗反爬措施非常强大。

1. 缺点

本质上，使用Selenium是通过驱动浏览器向目标站点发送请求，但浏览器必须等待静态资源加载完成（如HTML、CSS、JS等），这会拖慢速度。因此，Selenium的缺点是效率较低。但由于本文所爬取的招聘网站反爬能力较强，可能需要进行滑块验证等操作，因此，Selenium的反爬能力非常强大，适合用于该网站的爬取。

## 1.3 关键技术模块

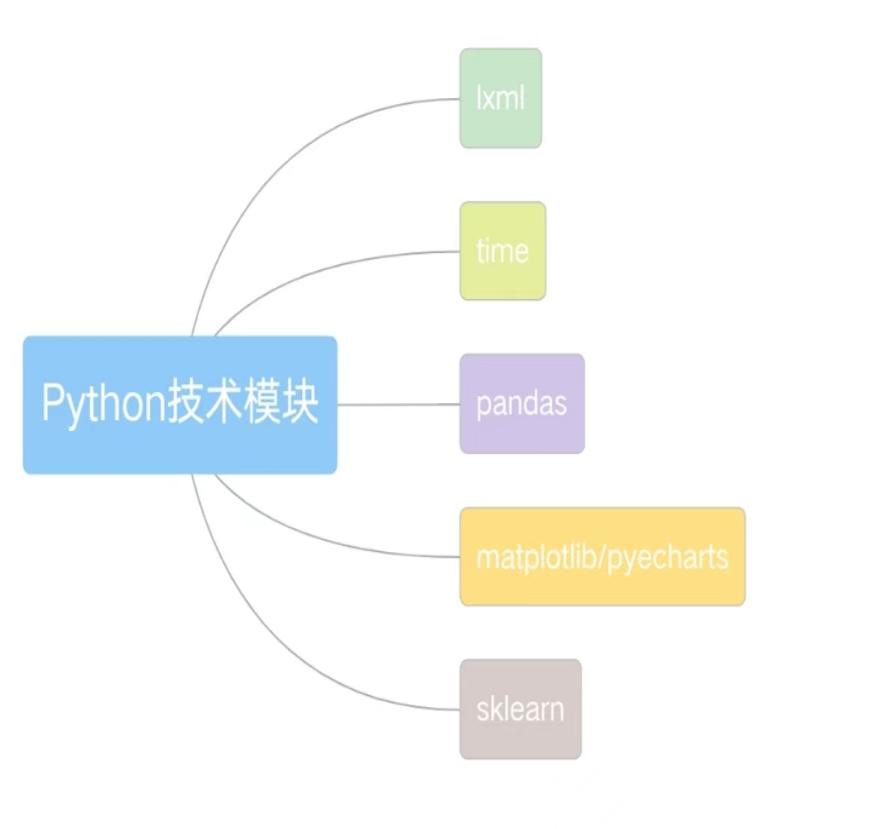


图1-1 技术模块

1. lxml

lxml 是一种使用 Python 编写的库，可以迅速、灵活地处理 XML 和 HTML，还可以用于web爬取，我们通过里面的etree.使用HTML()函数可以将输入的字符串格式的HTML文档对象转化成\_Element对象，从而更便捷地进行getparent()、remove()、xpath()等方法操作。这一函数为我们解析HTML文档提供了便利，不仅可以将其他格式的HTML文档转化为\_Element对象，还可以使我们在\_Element对象上进行更为高效的文本操作。

这种解析技术在网络爬虫技术中被广泛应用，能够大幅提高爬虫任务的效率和精度。利用lxml库当中的etree对得到的网页源码进行处理，在利用Xpath对得到的处理过的网页源代码进行定位，查找我们所需要的字段信息的位置，写入到Xpath中。

1. time

time库计算机时间的表达，我们利用sleep(5 \* random.random())方法让程序每爬取完一个界面随机休息再去爬取另外一个界面，模拟真人下载数据过程。

1. pandas

pandas是一种基于numpy的数据分析工具，拥有大量的数据模型和库，提供了高效的大数据集操作工具[7]。pandas还提供了大量方便快捷的数据处理函数和方法，使数据的处理变得更简单和高效[8]。pandas提供了大量能使我们快速便捷地处理数据的函数和方式。本文将爬取到的字段信息“岗位名”、“公司名”、“招聘专业”、“发布时间”、“公司类型”、“公司规模”、“工资”、“工作地点”、“工作经验”、“学历”利用Pandas当中的数据框DadaFrame，将招聘数据存储到本地，便于后续数据的预处理以及可视化操作。

1. matplotlib/pyecharts

[matplotlib](https://so.csdn.net/so/search?q=Matplotlib)是一款Python 绘图库，能在各种平台上以各种硬拷贝格式和交互式环境中生成出拥有出版品质的图形[9]。

它旨在使简单的事情更简单，让不可能的事情变成可能。只需要几行代码，便可生成各种图表，如绘图、直方图、功率谱、条形图、错误图和散点图等[10]。

pyecharts 是一种用于生成 Echarts 图表的类库[11]。Echarts 是百度开源的一款数据可视化JS库。利用Echarts生成的图表效果非常棒，为了便于在Python中直接利用数据生成图表，pyecharts应运而生[12]。  
  pyecharts可以生成动态图表，网页展示效果美观，同时展示数据时也非常便捷，鼠标悬停在图表上即可显示数值和标签等信息。

1. sklearn

Sklearn 是一个基于 Python 语言的机器学习工具[13]。它深扎于NumPy、SciPy、Pandas和Matplotlib等开源计算库之上，API设计非常优秀，每个对象的接口都异常简单。Sklearn为我们提供了便捷的数据挖掘、数据分析及机器学习算法实现等服务，充分利用了Python语言的灵活性、易用性和高效性，为科学计算和数据分析领域带来了许多福音。本文主要使用其PLabelEncoder标签编码器和LinearRegression线性回归进行分析操作。

# 2信息爬取与解析

## 2.1 爬取目标

这次我们要提取前程无忧51job招聘官网相关招聘信息[14]，搜索关键字为:数据

爬取任务需求：爬取与数据、大数据、数据分析、机器学习、人工智能等相关的岗位。

爬取提取字段：公司名、岗位名、工作地点、工资、发布时间、公司类型、公司规模、员工人数、所属行业。

提取站点的URL为：

https://we.51job.com/pc/search?keyword=%E6%95%B0%E6%8D%AE&searchType=2&sortType=0&metro=

打开后看到如图2-1所示是前程无忧51job网上有关工作岗位的基本信息：



图2-1 前程无忧51job网有关数据工作岗位的基本信息

## 2.2 爬虫流程分析

爬虫相关运行程序如下所示：

1. 创建selenium

如图2-2创建selenium并通过一个名为 excludeSwitches 选项，然后排除 enable-automation 开关实现自动化，通过初始webdriver等操作启动Chrome浏览器和进入浏览器设置。

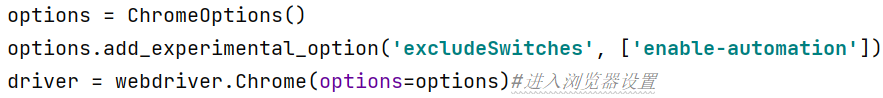


图2-2 创建selenium

1. 载入反反爬js

stealth.min.js是puppeteer中用于抹去自动化程序特征。当它被单独提取出来后就可以在selenium中加载并使用，使得可以抹掉selenium中的自动化特征，从而绕过一些网站或者验证程序的机器人检测。如图2-3先载入js，并调用函数在页面加载前执行脚本。



图2-3 载入反反爬js

1. 通过driver.get(url)获取网页

如图2-4利用driver.get()获取网页数据，并利用sleep(5)方法让程序每爬取完一个界面停顿5秒再去爬取另外一个界面，模拟真人下载数据过程。

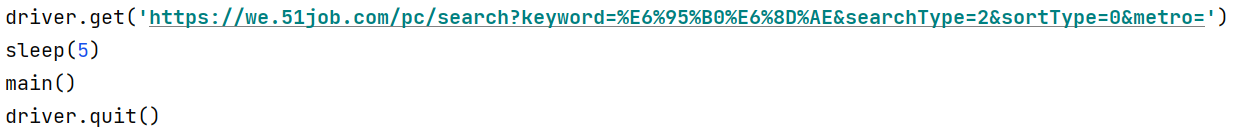


图2-4 获取网页数据

1. for循环

如图2-5进入页面后循环200次，每次利用sleep(random.random()/5)随机休息，调用js(execute\_script)，利用scrollBy() 使元素每隔一秒从当前的滚动条位置向下滚动50px，拿到网页后进行保存。

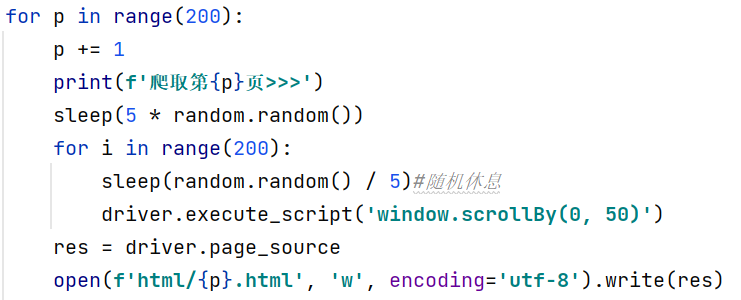


图2-5 for循环

1. 输入指定页面进行页面跳转翻页

如图2-6利用输入指定页面进行页面跳转翻页，判断如果不是等于最后一页即200页，清空然后输入想要的页数，休息后进行跳转。

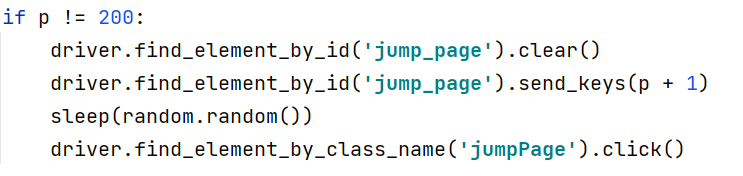


图2-6 跳转翻页

## 2.3 爬虫解析分析

1. 循环读取

如图2-7循环200次，每次读取一个html。

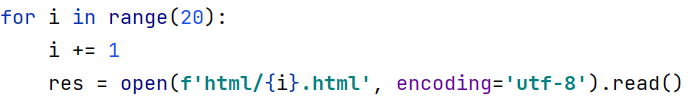


图2-7 循环读取html

1. 利用xpath进行解析

如图2-8通过lxml库，调用etree.xpath()对取所需数据进行解析，其中包含name，time，sala，addr，exp，edu，comp，kind，num，ind等类型，其中一些类型如addr，exp，edu取不到数据进行扩长列表，字段应为5个或没有，手动添加后不会报错，kind和num字段利用split进行分割字段，其中num字段可能没有第二个就强行变为两个进行取值。



图2-8 循环读取html

1. 保存

如图2-9保存进字典并命名，然后用pandas库中的DataFrame保存获得的所有数据，每页爬取完以后将数据框以追加（mode='a+'）的方式存入csv文件，使用utf-8编码方式存入。

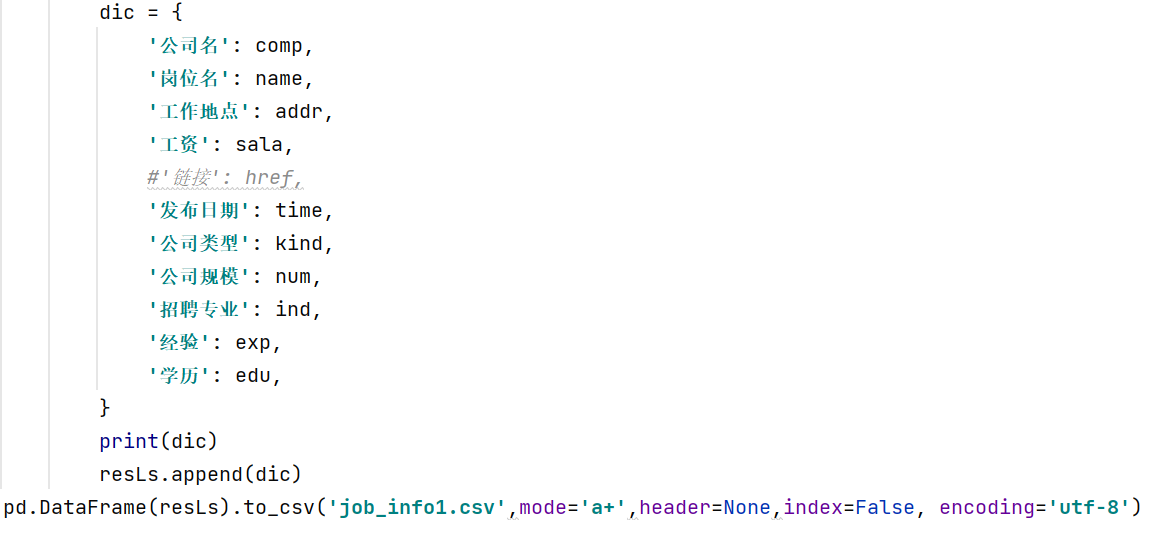


图2-9 循环读取html

# 3数据预处理

## 3.1数据介绍

如图3-1所示的数据集是通过上述爬取过程从前程无忧51job网站上爬取的数据集合。该数据集共有89125条数据记录，为其数据设置新的行，列索引，每条记录都包含公司名,岗位名,工作地点,工资,发布日期,公司类型,公司规模,招聘专业,学历,经验共10个字段的信息。



图3-1 原始数据

## 3.2重复值处理

先用data.shape获得去重之前的记录数，去除重复的公司名和岗位名两列行数据，再次利用data.shape获取去重之后的记录数，得到如下结果：

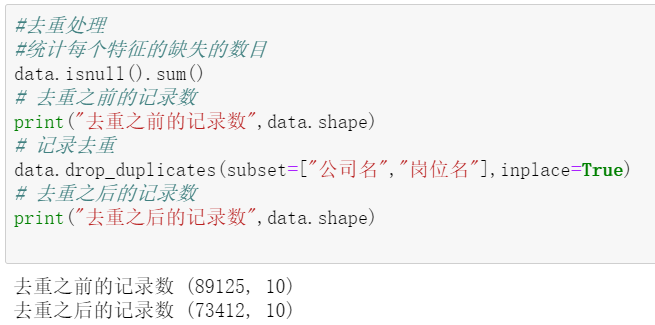


图3-2 去重记录数

## 3.3岗位名处理

先对岗位字段名进行探索，再构造想要分析的目标岗位，做一个数据筛选，将岗位名标准化, 设置rename函数，如果存在目标岗位，则index>0,对岗位名应用此函数进行数据筛选。得到相关数据为运营：7516，开发工程师：4740，数据分析：1513，软件工程：903，大数据：752等。如图3-3所示。



图3-3 岗位名处理

## 3.4工资数据处理

由于从前程无忧51job网爬取下来的工资数据不统一，单位存在万/月，千/月，万/年等，并且涉及范围有最低工资和最高工资，在这里我们使用python编写的程序对工资这一列数据进行专门的数据处理，全都转换成以元每月为单位的数据。为了得出便于后期的可视化工作，我们需要将范围化的数字转变为具体的数值，我们采取的策略为均值法。例如：0.8-1万/月，我们统一按照月薪进行处理，取8000元为最低工资、10000元为最高工资，最低工资和最高工资的中间值9000元作为该数据项的工资水平，对于单位匹配问题，我们采用正则表达式进行匹配和查找，统一数据及单位。

如图3-4所示利用get\_money\_max\_minwhile函数把多个单位的数据转换成了以元每月为单位的数据集，并利用mean方法求均值。

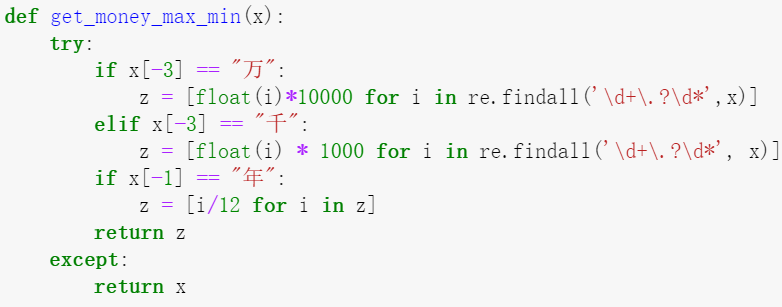


图3-4 转换单位

如图3-5所示，处理工资后的最终结果包含最低工资，最高工资和平均工资(工资)。



图3-5 工资处理结果

## 3.5工作地点处理

观察数据我们发现，工作地区有的为市级、有的为市级下面的区级，有的为一线城市、有的为三、四线城市，具体种类繁杂且不统一。我们对地区的处理方法为，选择一线二线的城市进行统计分析，并将工作地区的字段统一为一线、二线城市，清除掉区级，只具体到市级，便于后期的可视化。

我们具体以'北京', '上海', '广州', '深圳', '杭州', '苏州', '长沙','武汉', '天津', '成都', '西安', '东莞', '合肥', '佛山','宁波', '南京', '重庆', '长春', '郑州', '常州', '福州','沈阳', '济南', '宁波', '厦门', '贵州', '珠海', '青岛','中山', '大连'，这30个城市为主要城市，将工作地区字段当中包含有这些城市的数据项统一替换为以上30个城市。处理结果如图3-6所示。



图3-6 工作地点处理

## 3.6发布日期处理

为方便后续数据可视化，利用提取指定字符串内容的方式简化发布日期：job\_info["发布日期"] = job\_info["发布日期"].str[:-2]

## 3.7招聘专业处理

爬取下来的有关招聘专业信息，包含逗号，斜杠等分隔符，为整理出单独的招聘专业，利用切片划分方式获取。处理结果如图3-6所示。

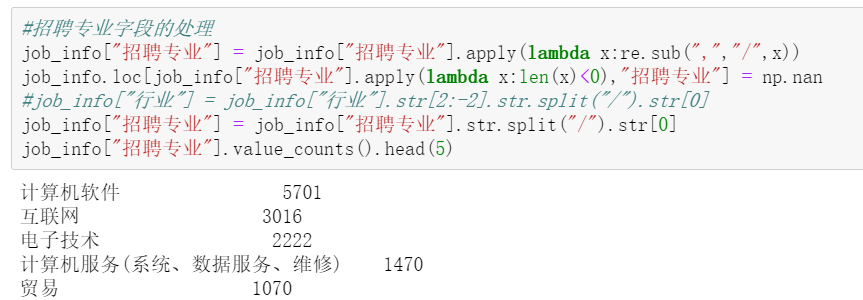


图3-7 招聘专业处理

## 3.8公司规模字段的处理

由于从前程无忧51job网爬取下来的公司规模数据不统一，例如少于50人，50-150人，10000人以上等，为方便后续可视化方便，设置func函数，进行公司规模字段的处理，统一化成数字形式<50，50-150，>10000等。



图3-8 公司规模范围及数量

## 3.9构造新数据

为处理好的数据设置新的列名，以.to\_csv的方式存入csv文件，使用GBK编码方式存入，以.groupby('岗位名').agg({'岗位名':'count'}).sort\_values('岗位名')方式分组，得到数据部分如图4-9。



图3-9 处理好的数据

# 4数据分析与可视化

数据可视化是将抽象数据变成图片或者图表的一种手段，通过视觉化的方式，直观地展现和呈现数据[15]。正是通过这种可视化的方式，我们可以更好地理解数据，准确快速地传达信息，并且从数据中发现潜在的规律和特征。

数据可视化是一种十分重要的数据交流方式，不仅可以为人们提供关于数据的重要属性和特征的确定，同时也可以帮助我们从海量数据中挖掘信息并做出明智的商业决策。从这个角度来说，数据可视化在现代的商业和科学研究中具有举足轻重的地位[16]。

## 4.1热门招聘岗位

对于以后学生将从事哪些岗位的工作有大致了解，我们对岗位名进行柱状图Bar()分析，由于爬取的招聘岗位种类繁多，使用value\_counts()[:10]方法抽取前10的热门的招聘岗位可视化进行分析，其中opts.MarkLineItem(type\_，name）标注线，用于标注岗位数最大值、平均值和最小值。

如图4-1所示，可以看出运营岗位招收需求最多，开发岗位次之，软件工程岗位相比较更为少些，算法、数据分析、java等岗位数量均在1000以下。

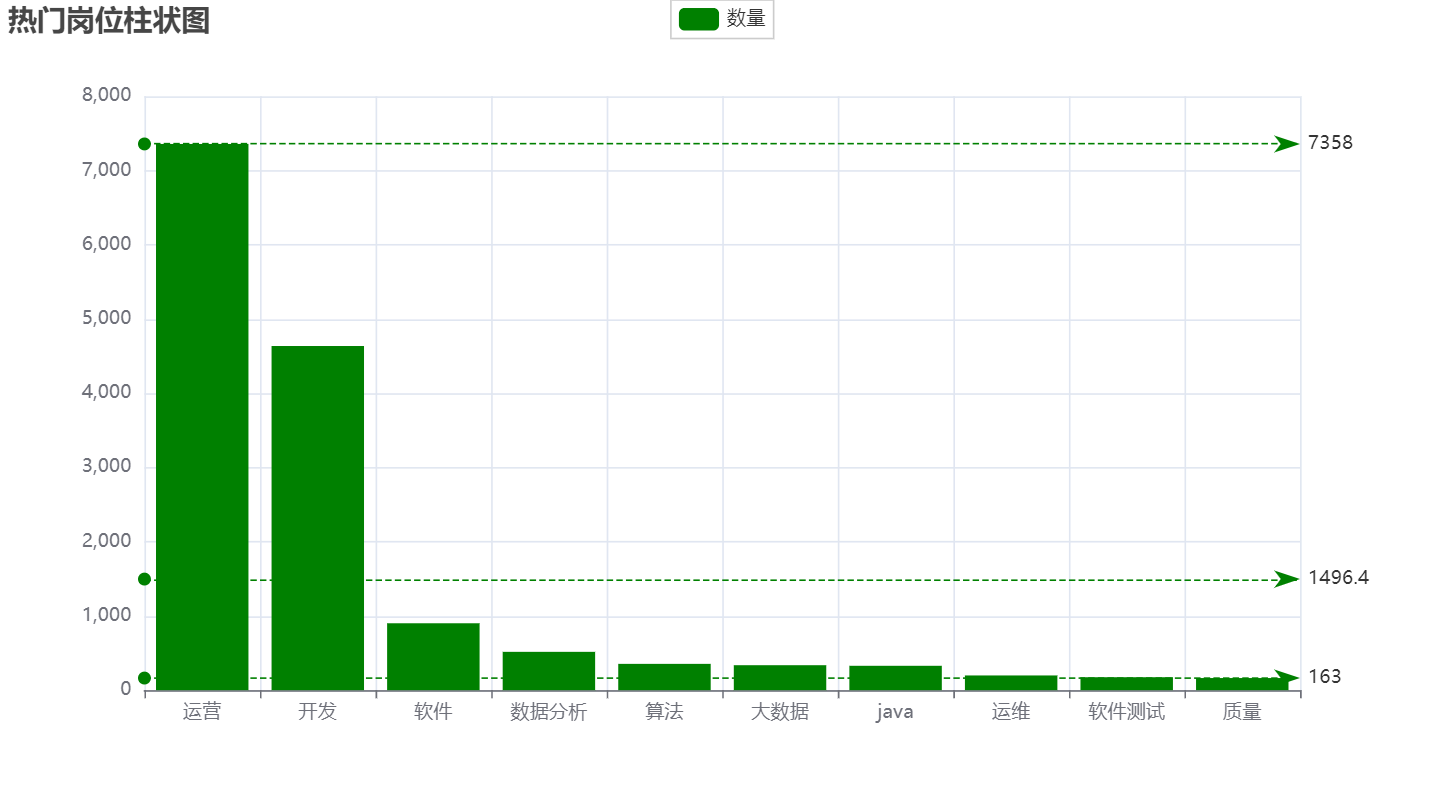


图4-1 热门招聘岗位Top10

## 4.2热门招聘城市

不论是毕业之后就业还是继续深造，工作地点的选择或者学校的地理位置的选择都是我们关心的内容，一个好的地理环境，或许可以带给我们更多的机遇和挑战。针对招聘数据，我们选取最热门的10个城市作为我们的研究对象，观察大多数招聘人才的市场主要集中在南方还是北方，主要集中在京津沪，还是其他二三线城市。

结果使用漏斗图Funnel()直观的来表达哪几个城市最受欢迎，使用value\_counts()[:10]方法抽取前10个城市，[list(z) for z in zip(a,d)]生成需要的列表数据，具体结果如图4-2所示。可以看出上海、深圳、广州的岗位数量较多，其中上海的岗位数量最多，杭州、武汉、北京岗位数量差不多，苏州、南京、成都岗位数量相差不大，东莞最少。

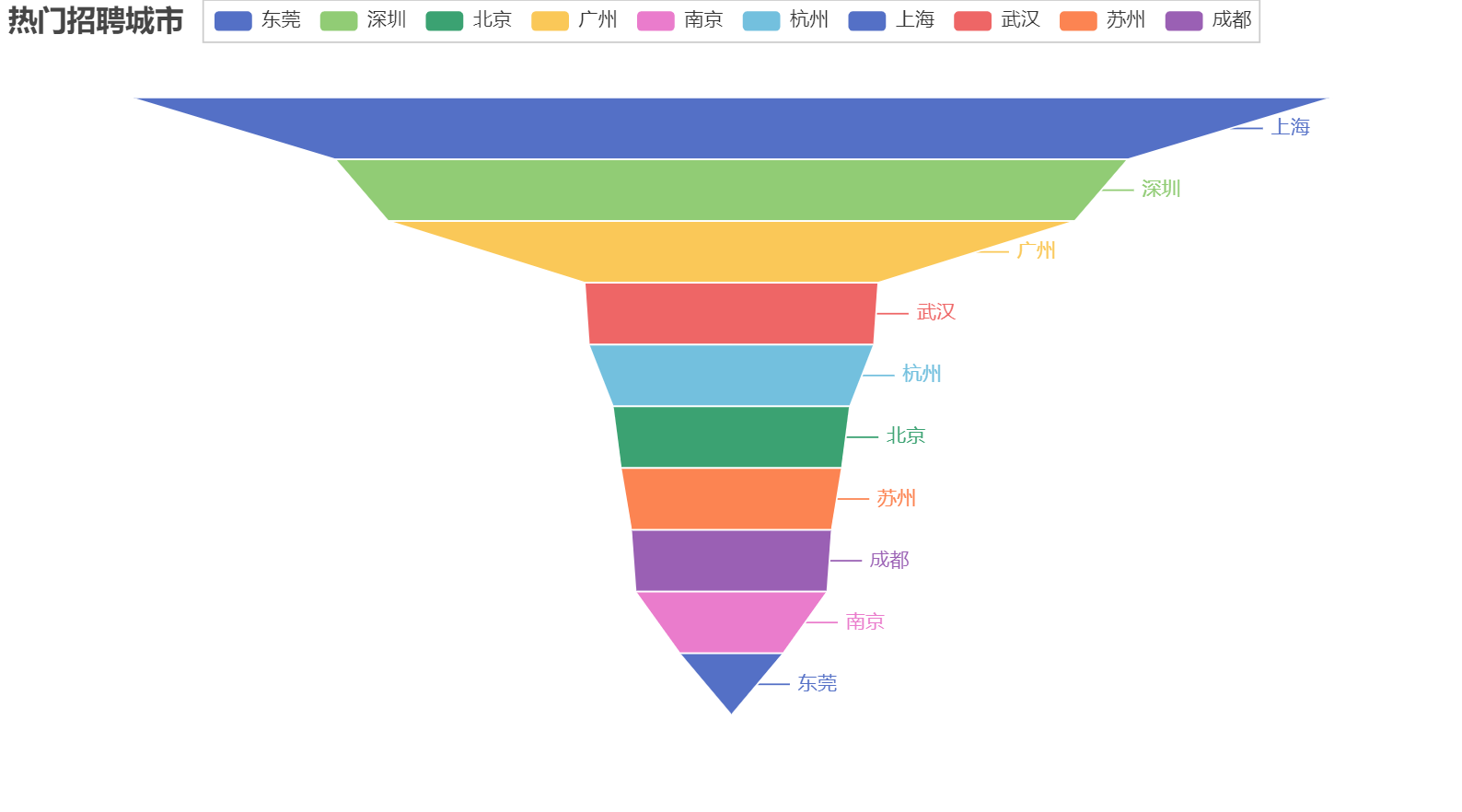


图4-2 热门城市 Top10

## 4.3招聘岗位的薪酬分布

为使学生了解自己毕业后就职岗位的薪资水平，特利用饼图展示，由于爬取信息数据工资水平种类繁多，设置for循环，利用饼图将其标准化为8个水平，分别是5K-10K, 10K-15K, 15K-20K, 20K-25K, 25K-30K, 30K-35K, 35-50K, 50K/月以上，在设置中文显示为 'SimHei'，axes(aspect='equal')可以保证饼图是个正圆，legend()设置图例，调节图例位置。

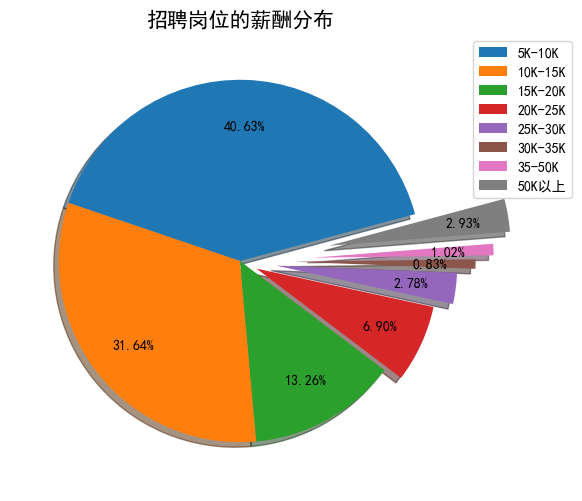


图4-3 招聘岗位的薪酬分布

饼图展示出招聘岗位的薪酬分布如图4-3所示，可以看出，招聘数据相关岗位的薪酬大多分布在每月5K-10K之间，占比40.63%，其次是10K-15K，占比31.64%，约占三分之一，15K-20K次之，占比13.26%。

## 4.4招聘岗位对学历的要求

从提取出的特定学历指标，利用水平柱形图barh()展示招聘岗位对学历的要求，figure(figsize=(), dpi)设置大小，像素。

如图4-4所示，招聘岗位对学历要求为本科的占大多数，在数据筛选后存在12616个岗位数，其次是大专占12270个岗位数，相比较而言对硕士要求较少，仅存在703个岗位数，对博士和初中及以下的学历要求就更少了，分别是29和9个，在柱状图中显示并不明显。所以可以推测本科就业形式良好。

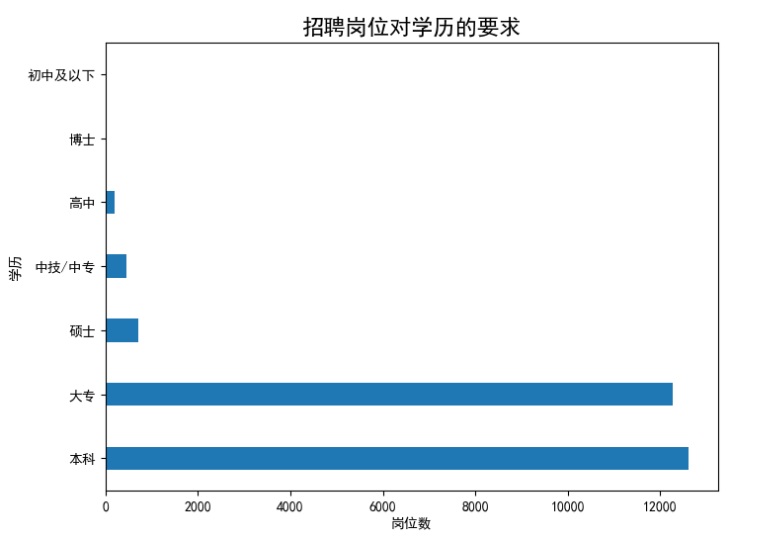


图4-4 招聘岗位对学历的要求

## 4.5招聘岗位对工作经验的要求

利用柱形图bar()展示招聘岗位对工作经验的要求，如图4-5所示，招聘岗位对有3-4年经验占大多数，1年经验和2年经验数据相似，其余对5-7年经验和无需经验相对要求较少，8-9年和10年以上经验最少。可以看出岗位招聘比较看重工作经验。

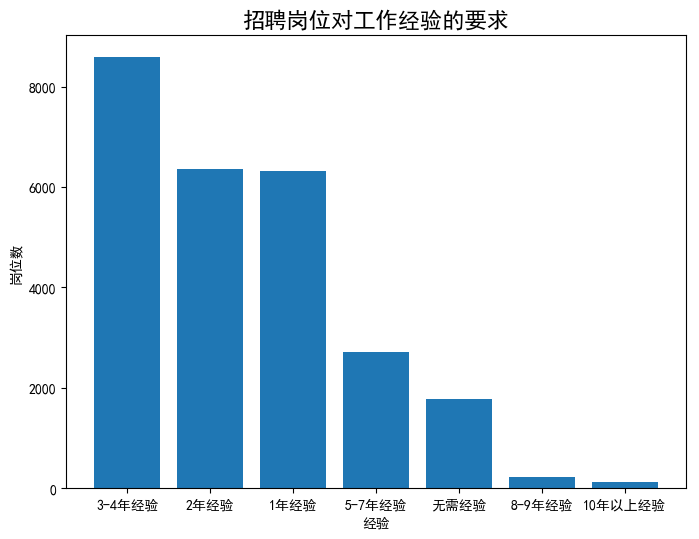


图4-5 招聘岗位对工作经验的要求

## 4.6招聘公司所属行业的分布情况

对于以后学生将从事哪方面的工作，我们对公司所在的行业进行词云统计，观察大多数毕业生将从事哪方面的行业，我们观察数据发现，每个公司所属的行业不唯一。我们采用split()方法将字段分离开，利用jieba库对中文字符进行处理，将一些特殊字符和停顿字符在我们的字符串中删除，得到完整数据，最后利用WordCloud进行图云绘制并保存。结果如图4-6所示。得出相对于热门的招聘专业所属行业是计算机软件，互联网，电子技术，电子商务，计算机服务，贸易，通信等[17]。学生能够提前了解有关该行业的相关知识以及背景，对未来向着那个行业发展做好规划和统筹。

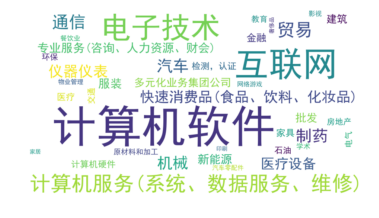


图4-6 招聘公司所属行业的分布情况

## 4.7公司类型占比统计

为了研究大数据专业的学生毕业之后，对就业、择业所选择的公司性质有一个很好地认识，我们探究关于选择数据专业的学生的公司是民营企业、还是国企、还是外资等，观察那种类型的公司所占比例最高，根据整理的公司类型字段的信息数据。

利用饼图Pie()的绘制，[list(z) for z in zip(types, number)生成所需要的类型和数量的列表，set\_global\_opts()设置全局选项，比如标题和图例，set\_series\_opts()设置系列选项，比如饼图的半径和样式。由图4-7所示，招聘数据类型岗位的公司中民营公司就业职位占比最高，其次是上市公司，其中政府机关和外企代表处最少。

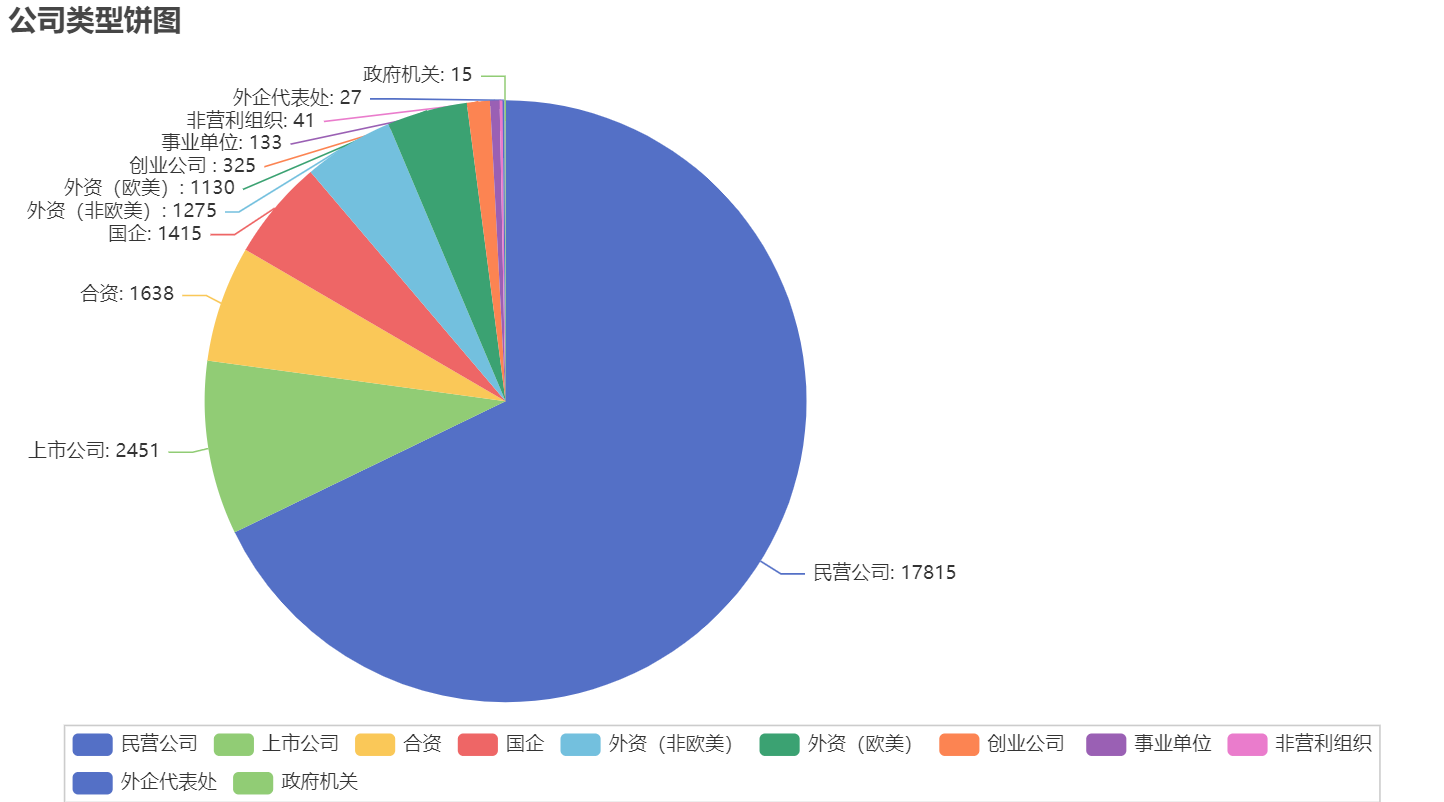


图4-7 公司类型占比数量统计

## 4.8不同公司规模的公司数

为了研究招聘公司的规模大小，我们采取柱状图Bar()的形式，来研究一般情况下所要招聘的公司的大致规模，来观察招聘数据相关专业的学生的公司大多为什么规模类型的公司，是大公司招聘的多些，还是小公司招聘的多些。opts.MarkLineItem(type\_，name）标注线，用于标注公司数最大值、平均值和最小值。

具体结果如图4-2所示。根据整理的公司规模字段的信息数据，由图5-9所示，公司规模数量在50-150和150-500人数的最多且属于小规模公司，都处在6000-7000之间，5000-10000规模人数的大规模公司最少，换句话说，小公司需求大，大公司需求小。

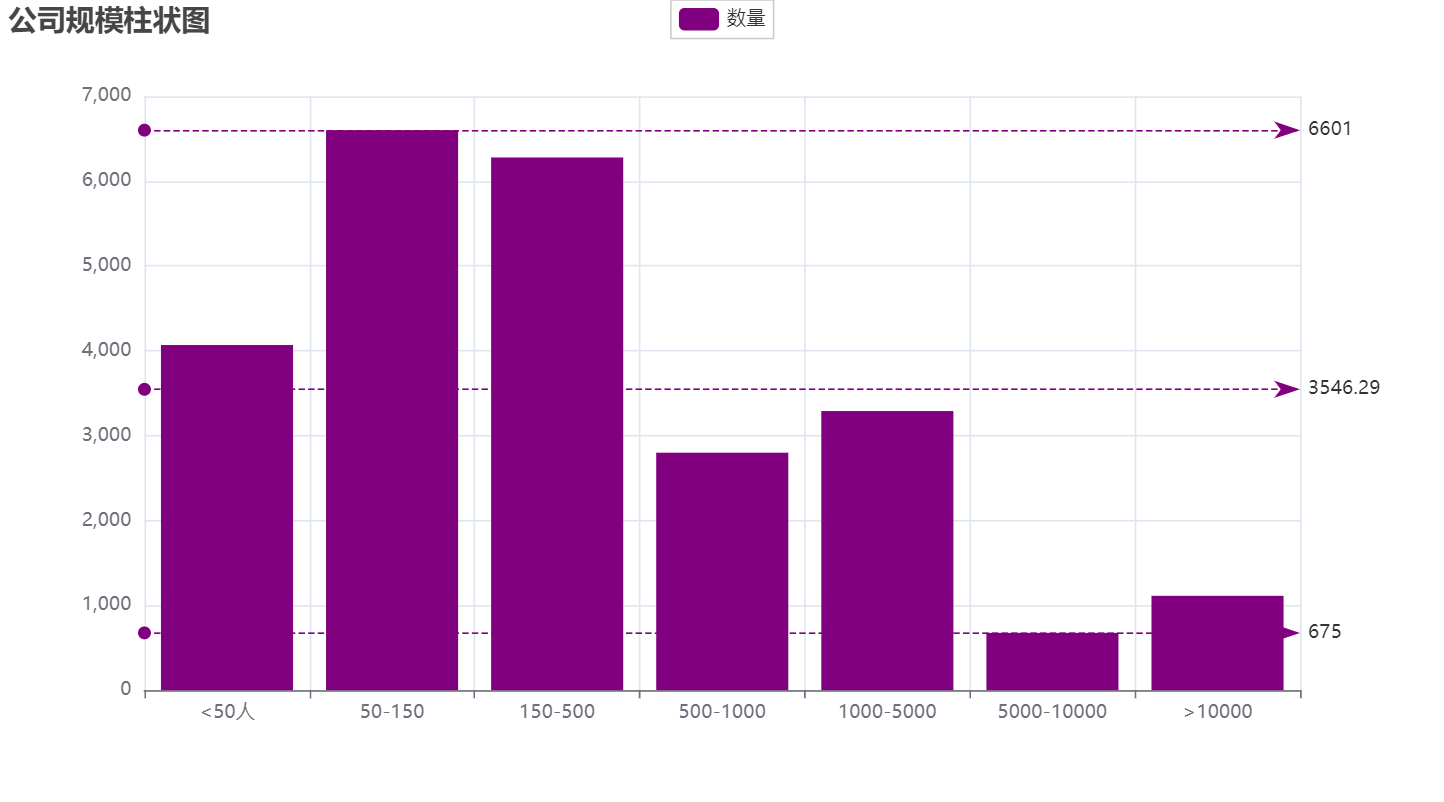


图4-8 不同公司规模的公司数

## 4.9热门城市工资比较

为了研究学生以后就业热门城市的工资水平，我们采用箱线图boxplot（）的形式观察，sort\_values( ascending=False)[:5]提取排名前五的热门城市，df.groupby(['工作地点'])['工资']排名前五热门城市工资进行分组，如图4-9所示，观察发现上海和杭州薪资较高，深圳，广州。武汉次之，但都相差不大。

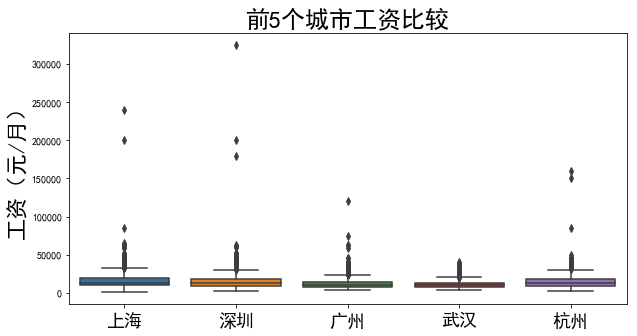


图4-9 招聘公司所属行业的分布情况

## 4.10热门招聘公司地理位置

通过对热门城市的漏斗图我们可以明显的看出，最热门的前三座城市为上海、深圳、广州。为方便研究热门招聘城市的地理位置，使用pyecharts绘制地图，map()构造函数，先构建坐标系列名称为城市，地图类型maptype: str = "china-cities"，有世界，国家名，省份名，市名四个等级，然后使用.add()设置具体内容,最后生成本地html文件，如下图4-10所示，从地理位置上不难看出，这三座城市都是沿海城市，且地处南方，对外贸易频繁，确实符合实际认知。也不难看出，这样的城市用人需求也很大，招聘人数也远高于首都北京，说明大多数人才流动更偏向于南方。

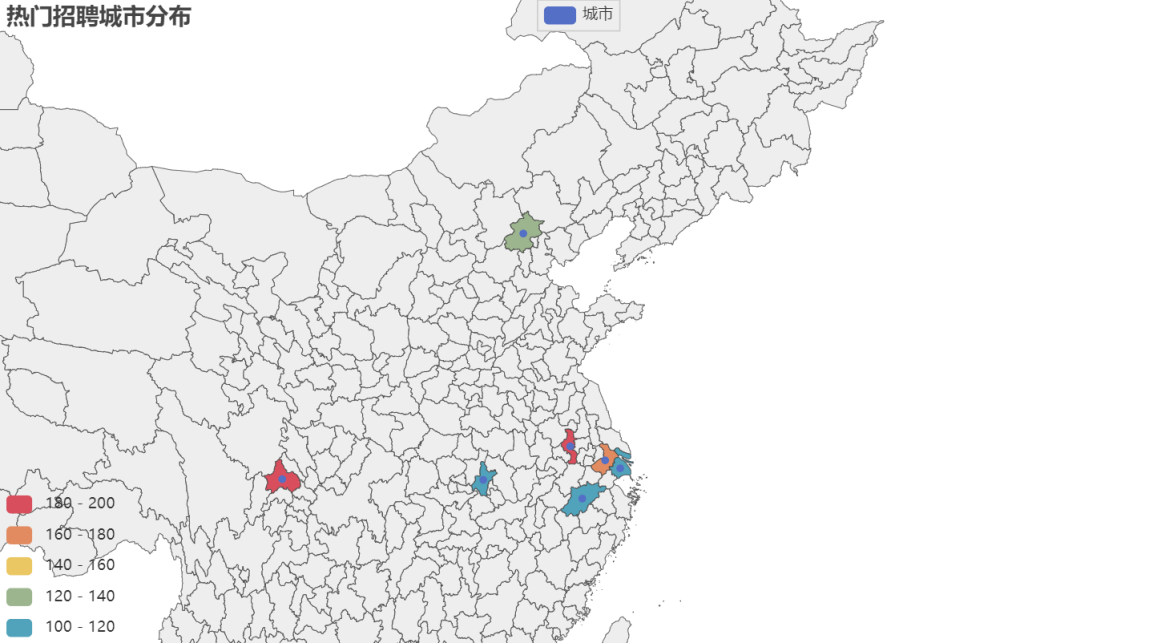


图4-10 招聘公司所属行业的分布情况

## 4.11学历和工资的关系

对于毕业生，薪资水平肯定是我们需要考虑的一方面，利用柱状图Bar()和折线图Line()我们可以大致了解以后所处工作的大致薪资水平。以df[['最低工资','学历']].groupby(by='学历').mean()为例，按照学历对最低工资进行分组，得到学历最低工资均值，然后再设置相关参数。而折线图是以平均工资反应学历和工资关系。我们将分别探究“最低工资”、“最高工资”、“平均工资”三大类别与学历的关系，探究学历对薪资水平是否有影响，学历和薪资之间大致呈现什么关系。

具体结果如图4-5所示，我们可以大致看出学历和工资两者之间的关系大致呈线性关系。我们可以清楚的看出，博士的平均工资最高，初中及以下的平均工资最低，且在“高中”及以下学历薪资情况基本一致，但最高都不超过10000元/月，平均薪资也在5000元左右，不会太高。在“大专”以上学历，薪资水平越来越高，符合我们正常的生活认知，学历越高，薪资越高，最高达到30000元/月以上。可见学历的重要性！

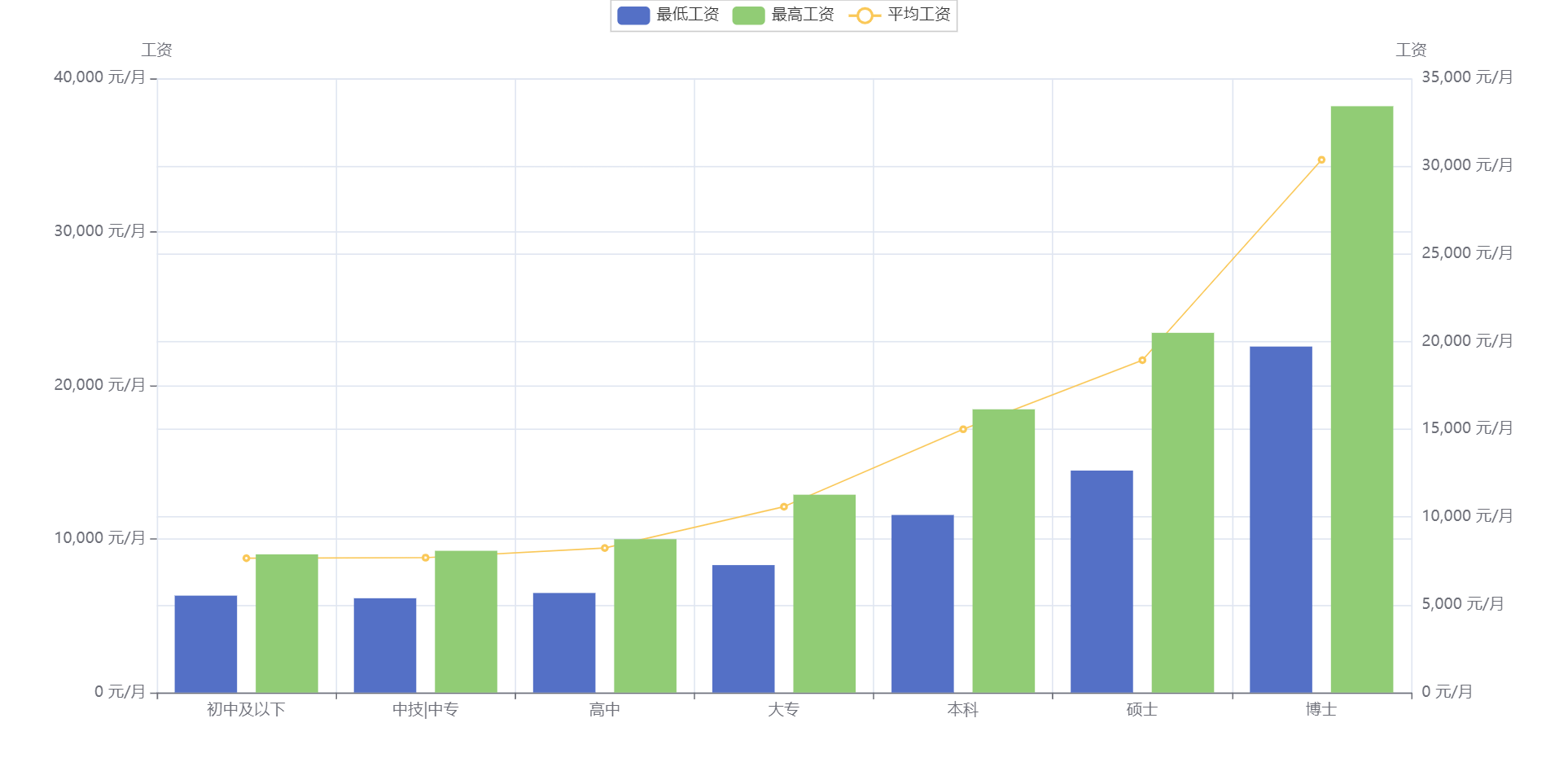


图4-11 招聘公司所属行业的分布情况

# 5数据多元回归建模预测薪资

## 5.1多元线性回归

LinearRegression线性回归是一种回归学习方法，基于最小二乘法的线性回归[18]。一般用于处理连续性变量，算是机器学习的入门算法[19]。尽管线性模型的形式很简单，可是线性模型的思想是很重要的，非线性模型通常可在线性模型的基础上通过引入高维映射得到[20]。该模型通过数据拟合，得到一个相对符合实际规律的数学表达式，寻找数据之间的相互关系[21]。例如对于工资的预测。

多元线性回归分析是指使用两个或更多自变量的回归分析。

由于许多现象与多个因素有关，该方法采用多个独立变量的最佳组合来预测或估计因变量。相比于只使用一个自变量来进行预测或估计，多元线性回归更为准确实用[22]。

优点:建模速度快，不需要复杂计算；可解释性。

缺点:不适用与非线性数据；可能出现过拟合。

拟合是指构建一种算法，使其能够匹配真实数据。从机器学习角度来看，线性回归需要构建一个线性函数，以使函数与目标值之间的匹配度最高。从几何角度来看，就是让函数的直线（或平面）尽可能接近空间中的所有数据点（点到直线的距离之和最小）。线性回归的输出为连续值[23]。

现实中的数据可能是比较复杂的，自变量也可能不止一个，例如，影响工资也很可能不止学历一个因素，可能还有经验、地区等诸多因素。不过，这些因素对工资影响的权重是不同的，因此，我们可以使用多个权重来表示多个因素与工资的关系：

其中x：影响因素，即特征；w：每个x的影响力度；n：特征的个数；y^：工资的预测价格。

## 5.2提取相关字段

导入所需sklearn相关LabelEncoder标签编码器是用来对分类型特征值进行编码，即对不连续的数值或文本进行编码，LinearRegression线性回归，提取处理后的job\_info2.csv数据文件中的工作地点、工资、经验、学历字段单独建立jobs表格。

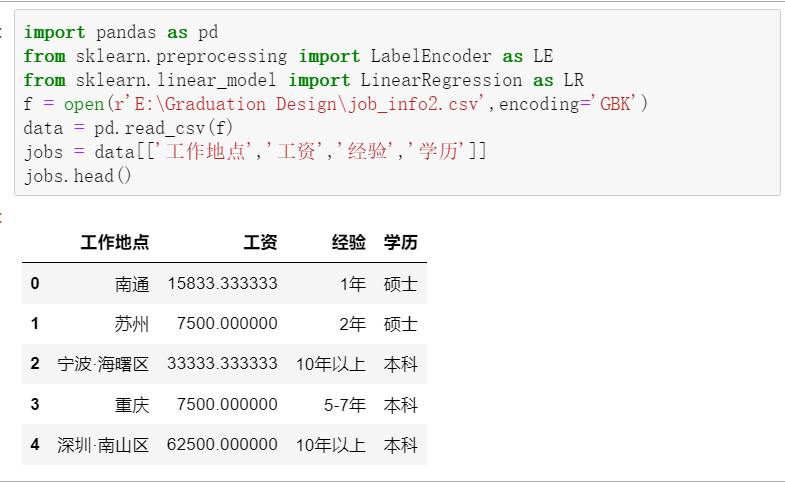


图5-1 提取相关字段

## 5.3字符串函数处理

定义函数get\_area(area)，提取工作城市名字，用split('-|·')进行提取，若存在则返回第指定字符前面的文字，若不存在指定字符则直接提取，定义函数get\_exp(exp)。

处理经验字符串，split('-')进行提取，若存在则返回第指定字符前面的文字，若不存在指定字符则直接返回0，代表无需经验，定义函数get\_wage(wage)，因为工资字符串在前面已经处理完毕，这里就不进行处理，仅定义函数，然后将上述三个函数应用到对应列。如图5-2所示。

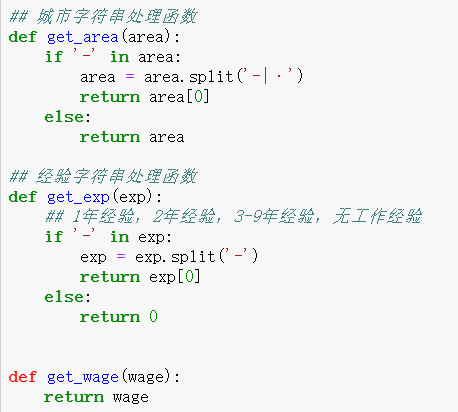


图5-2 字符串函数处理

## 5.4建立模型

先将各字符分类变量重编码为数值分类变量，利用fit\_transform将城市列重新排列后再进行标准化即重编码，再利用replace()将学历重编码，分别编码为中技/中专：0，高中：1，大专：2，本科：3，硕士：4，博士：5，初中及以下：6，再进行特征选择jobs指定工作地点、经验、学历为X，工资列为结果集y。

调用sklearn线性回归算法步骤:

1. 导入模型。调用逻辑回归LinearRegression()函数，创建分类器。
2. fit()训练。调用fit(X,y)的方法来训练模型拟合数据点，其中X为数据的属性即工作地点、经验、学历，y为所属类型即工资。
3. predict()预测。LabelEncoder可以将标签分配一个0—n\_classes-1之间的编码，将各种标签分配一个可数的连续编号，再利用训练得到的模型对数据集进行预测，返回预测结果。如图5-3为模型建立的预测函数。

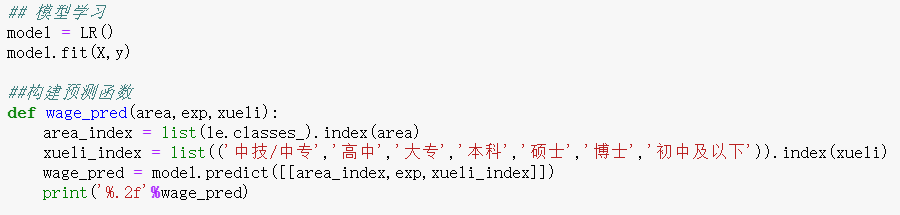


图5-3 模型建立

然后测试代码，输入城市、经验、学历三个参数进行预测，这里以“深圳，2，本科”为例，得到的结果是14289.76的工资水平，测试完毕。

# 结论

这篇论文深入剖析了网络爬虫方面的知识，总结出理论与实践相辅相成的编程实践策略，即多加上机编写程序，提升编程动手能力和独立思考的能力。同时，此次论文也使我对网络爬虫产生了兴趣，特别是在通过查书籍资料以及网上学习资源的情况下能够使程序运行成功，得到了前所未有的成就感。

在爬虫方面，selenium爬虫可以帮我们避免繁琐的通信流程，通过驱动浏览器，完全模拟浏览器的操作，比如跳转、输入、点击、下拉等，从而拿到网页渲染之后的结果，进行自动执行自定义的逻辑代码，通过代码完全模拟成人类使用浏览器自动访问目标站点并操作，爬虫通过里面的webdriver等操作进入浏览器设置和启动Chrome浏览器。

在数据的预处理方面，利用Pandas的数据框数据的处理，整合数据，分割数据，利用正则表达式查找我们想要的字符和汉字等。

在可视化方面，没有利用传统的matplotlib库，而是利用Pyecharts库作为可视化的核心库，它能动态的展示数据变化并以html的形式存储到本地，美观，实用，和matplotlib有着异曲同工之妙。

在多元回归建模方面，利用sklearn相关库对模型进行准备，最后做出输入地区，经验，学历，预测薪资的模型。

在论文过程中，我也遇到了很多困难，例如翻页使用自动按钮后发现如果在某一页出了问题，重新启动爬虫会仍从第一页开始，这使得爬虫时间增加以及资源浪费，为此我的处理方式是选择将自动按钮更改为页面跳转，每次进入网页时清理数据，填入需要的页码进行爬取，这样可以在指定的页面进行爬取，并以追加的方式存入数据。

结合数据可视化结果，以下是我有关大数据专业学生的几点建议：

1. 加强学校专业课的学习，针对每门专业课做到基础夯实，为以后工作选择的行业夯实基础。
2. 针对毕业之后想要就业的同学，根据分析，南方沿海城市或者北京、武汉等地针对数据相关的人才需求较高，热门城市为上海、深圳、广州，地理位置较好，机遇与挑战并存。
3. 对于本专业的学生，本科毕业工资平均为16000元/月，博士的工资为30000元/月，我们可以看出学历是决定薪资的关键因素，所以如果有机会可以选择继续深造，提升自我学历，创造更优越的价值。
4. 针对公司的选择大部分为民营公司，且大多数为小公司，说明这类公司对数据相关人才需求较高，但是我们也发现，大公司体制比较完备，各有各的优点及缺点，如果有机会，有能力，挑战大厂也不失为一个好的选择。

# 参考文献

1. 杨月.Python网络爬虫技术的研究[J].电子世界,2021(10):2.
2. 李泽鑫.基于互联网大数据的实时房屋租赁价格指数编制与实证研究——以山东省为例[D].山东财经大学,2019.
3. Razzaq Asif,Yang Xiaodong.Digital finance and green growth in China: Appraising inclusive digital finance using web crawler technology and big data[J].Technological Forecasting & Social Change,2023,188.
4. CHEN R,CHEN Y,DENG C, et al. Design of Data Capture Program Based on Web Crawler Technology[C]//Advanced Science and Industry Research Center.Proceedings of 2018 International Conference on Information, Electronic and Communication Engineering(IECE 2018).DEStech Publications,2018:229-232.
5. Rui-xing CHEN,Yi CHEN,Chun-yuan DENG,Di MO. Design of Data Capture Program Based on Web Crawler Technology[P].DEStech Transactions on Computer Science and Engineering,2018.
6. 王碧瑶.基于Python的网络爬虫技术研究[J].数字技术与应用,2017(5):1.
7. 王申.大数据时代下的图书馆建设与应用研究——个性化实时推荐系统的建立[C]// 2017年全国图书馆新型服务能力建设学术研讨会.2017.
8. 鞠文飞.Python结合大数据技术分析教务数据[J].2019.
9. 马佳琪,滕国文.基于Matplotlib的大数据可视化应用研究[J].电脑知识与技术：学术版,2019,15(6Z):2.
10. 汤琳纯.人工智能教育课程的教学探索与实践[J].福建电脑,2021.
11. 吴蕙羽.基于Python技术分析小说人物关系和社交网络[J].电脑编程技巧与 维护,2020(6):3.
12. 龚丽媛.宁夏联通固移融合"在网有礼"系统的设计与实现[D].电子科技大学.
13. 周永江.基于机器学习的多源数据雾霾预测模型研究[D].武汉大学.
14. 毛遂,毛红霞.基于51job网站招聘信息的爬取与分析——以Python技术岗位为例[J].网络安全技术与应用,2021(04):47-49.
15. Ziwen Yu,Jianjun Zhang,Wenwu Tan,Ziyi Xiong,Peilun Li,Liangqing Meng,Haijun Lin,Guang Sun,Peng Guo.Design of a Web Crawler for Water Quality Monitoring Data and Data Visualization[J].Journal on Big Data,2022,4(2).
16. 杨刚,杨凯.大数据关键处理技术综述[J].计算机与数字工程,2016,44(4):6.
17. 张宗霞,陈玉忠.高校大数据人才培养与企业需求实证研究[J].湖南工业职业技术学院学报,2020,20(5):6.[8]
18. 薛方,苏芮锋,杨升等.基于回归分析的公路货运量预测[J].汽车实用技术, 2019(15):5.[10]
19. 刘淼.基于Spark并行框架的推荐系统研究[D].长春理工大学.
20. 章金成.基于网络口碑和搜索数据的汽车销量预测研究[D].青岛大学.
21. 黄欣荣.大数据对科学哲学的新挑战[J].新疆师范大学学报：哲学社会科学版, 2016, 37(3):7.
22. 黄俊聪.南方电网区域大气腐蚀等级图绘制--剂量响应函数建立及绘图方法优化[D].华南理工大学.
23. Luo Mi,Lin Jianzong. Research on the impact of Credit, Brand and Service Recovery on Online Sales Based on Web Crawler Technology and Regression Analysis Methods[J].Journal of Physics:Conference Series,2021,1955(1).

# 致谢

行文至此，意味着我的大学生涯已然落幕，这是我刚进校园时梦想到达的时刻，却又是如今留恋不舍的时刻，当初的自己初到大学的茫然无措，看什么都觉得新鲜，总觉来日方长，却不知时光匆匆。

在本篇论文即将收尾之时，我想由衷的感谢我的论文指导老师，在作此次毕业设计的过程中，得到了袁旭东老师的精心指导，正是因为老师不断解疑答惑，使我在论文内容方面不断完善更进。

感谢数学与大数据学院的每一位老师，有了诸位恩师的教导才有了如今的我们，是老师们教导我，为学之要贵在勤奋，贵在坚持，贵在有恒。

感谢我的大学室友，是你们朝夕陪伴了我的大学四年，一起哭，一起笑，花开花落万物道，聚散离别终有时，这段弥足珍贵的青春岁月，我将用一生来怀念。