河南大学民生学院 2023 届本科毕业论文（设计）

互联网研发岗位信息特征挖掘与分析

作 者 姓 名：郭名胜

作 者 学 号：1903691045

所 在 学 院：信息工程学院

所 学 专 业：数据科学与大数据技术

导 师 姓 名：李辉 杨杰

导 师 职 称：副教授 助教

2023 年 4 月 9 日

河南大学民生学院 2023 届本科生毕业论文（设计）

开题报告

论文（设计）题目	互联网研发岗位信息特征挖掘与分析		
所在学院	信息工程学院	专　业	数据科学与大数据技术
学生姓名	郭名胜	学　号	1903691045

一、本课题研究意义

信息化时代的高速发展与就业市场的激烈竞争，使众多的软件技术也随之不断交替更新。但由于众多应届毕业生缺乏工作经验以及部分高校教育相对滞后于市场发展，应届大学生在毕业时并不能达到招聘市场的一些要求，不能做到学以致用。为更好的推动大学生就业以及促进高校发展，本系统使用大数据分析的方法对网络上的招聘数据进行分析并给在校学生与高校更好的适应当今市场发展提供一定的借鉴意义。

二、国内外有关本课题的研究动态

企业招聘数据反映了企业对专业人才在知识、能力等方面的需求。通过对比国内目前的许多研究成果可以发现，对众多的招聘信息数据分析大都在宏观上提供了对岗位给求职方向的启示，但并未给学生的工作类型以及适合当今市场发展需要的相关知识水平、能力提供引导。目前我国对本科教育改革推行了一系列行之有效的举措，如校企合作，学生实习，课程改革等。但其中有些举措并不能普及全部学生，而课程的改革又使学生不能更好的审视自己的需要而盲从于学校的安排，许多学生在毕业时仍面临着就业难的问题，在学生独立选择工作岗位时发现自己的知识水平也许并不能适应当今市场的发展。

三、本课题研究的基本内容

本课题主要通过对互联网上的企业招聘数据的抓取，运用数据挖掘与大数据分析的方法分析不同岗位适时的相关技术栈，从而呈现出相应的关键词从而便于学生更好的找到当前适合自己未来发展需要掌握的相关技术栈，学生可以通过搜索引擎的帮助从而分析出自己想要从事的工作需要学习和掌握的相关技术。

四、本课题拟解决的主要问题

实现对招聘平台信息的抓取并分析得出更好的适应学生独立学习的需要，以及便于学生更好的在自主择业时达到企业的要求。

五、研究方法

信息研究方法

六、主要创新点

从学生自主择业与独立学习的方向进行分析从而更好的适应学生的独立发展的需要

七、主要参考文献

宋齐明.校园与工作场所：关于本科生可就业能力的研究[C].华东师范大学,2018.
黄承慧;印鉴;侯昉.一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J].计算机学报,2011,34(05):856-864.
张俊峰;魏瑞斌.国内招聘类网站的数据类岗位人才需求特征挖掘[J].情报杂志,2018,(06):176-182.
陈飞.应用型本科教育课程调整与改革研究[C].华东师范大学,2014.
钟晓旭.基于 Web 招聘信息的文本挖掘系统研究[C].合肥工业大学,2010.

具体时间及写作进度安排

2022.1.1-2022.2.28	项目设计
2022.3.1-2022.3.31	毕业论文写作
2022.4.25	毕业答辩

指导教师对开题报告的意见

通过

指导教师签名： 杨杰 2023 年 4 月 10 日

系（室）对本课题开题的意见

同意开题。

负责人签名： 吕永飞 2023 年 4 月 15 日

河南大学民生学院 2023 届本科生毕业论文（设计）

中期进展情况检查表

<table>
<tr><td>论文（设计）题目</td><td colspan="3">互联网研发岗位信息特征挖掘与分析</td></tr>
<tr><td>所在学院</td><td>信息工程学院</td><td>专　业</td><td>数据科学与大数据技术</td></tr>
<tr><td>学生姓名</td><td>郭名胜</td><td>学　号</td><td>1903691045</td></tr>
<tr><td colspan="4">毕业论文（设计）进展情况
目前已经将毕业设计的前期数据采集工作完全完成，并正在进行数据预处理部分的编码工作。</td></tr>
<tr><td colspan="4">毕业论文（设计）存在问题及解决方案
在进行毕业设计编写的过程中遇到的较为复杂的问题是“在数据量不够多的情况下训练出的结果并不能的出较好的提取结果。作者找到的一种解决方法是通过使用目前更为先进的自然语言处理模型从而免去对数据进行训练的步骤。</td></tr>
<tr><td colspan="4">三、指导教师对学生论文（设计）进展等方面的评价
通过

指导教师签字：杨杰 毕设专用　2023 年 4 月 10 日</td></tr>
<tr><td colspan="4">系（室）对本课题中期检查的意见
中期检查通过。

负责人签名：吕永飞 毕设专用　2023 年 4 月 20 日</td></tr>
</table>

河南大学民生学院 2023 届毕业论文（设计）

评阅教师评价表

<table>
<tr><td colspan="2">论文（设计）题目</td><td colspan="4">互联网研发岗位信息特征挖掘与分析</td></tr>
<tr><td colspan="2">所在学院</td><td>信息工程学院</td><td>专　业</td><td colspan="2">数据科学与大数据技术</td></tr>
<tr><td colspan="2">学生姓名</td><td>郭名胜</td><td>学　号</td><td colspan="2">1903691045</td></tr>
<tr><td colspan="2">评阅教师</td><td>张燕妮</td><td>评阅教师职称</td><td colspan="2">助教</td></tr>
<tr><td rowspan="8">评阅教师评分</td><td rowspan="2">序号</td><td rowspan="2">评分项目</td><td colspan="2">评分参考</td><td>得分</td></tr>
<tr><td>评阅指标（优秀标准）</td><td>满分（100 分）</td><td></td></tr>
<tr><td>1</td><td>选题质量</td><td>符合专业培养目标：有实际意义和推广价值。</td><td>20</td><td>18</td></tr>
<tr><td>2</td><td>文献资料应用能力</td><td>能独立查看文献，具有收集，加工各种信息及获取新知识的能力。</td><td>10</td><td>7</td></tr>
<tr><td>3</td><td>调查研究能力</td><td>能准确理解课题任务：研究方案设计合理:能独立从事调查研究，能综合运用所学知识发现与解决实际问题。</td><td>30</td><td>25</td></tr>
<tr><td>4</td><td>论文（设计）质量</td><td>格式，图表（或图纸）规范符合要求：结构严谨，逻辑性强 内容翔实，表达准确流畅：学术价值或使用价值高。</td><td>30</td><td>26</td></tr>
<tr><td>5</td><td>创新能力</td><td>观点独到，方法新颖，角度新颖。</td><td>10</td><td>8</td></tr>
<tr><td colspan="3">总得分</td><td colspan="2">84</td></tr>
<tr><td rowspan="3">评阅教师评定意见</td><td colspan="5">参照上述评价标准及论文（设计）内容，做出具体评价：</td></tr>
<tr><td colspan="5"></td></tr>
<tr><td colspan="5">评阅教师签名：张燕妮　　2023 年 4 月 20 日</td></tr>
</table>

河南大学民生学院 2023 届毕业论文（设计）

指导教师评价表

<table>
<tr><td colspan="2">论文（设计）题目</td><td colspan="4">互联网研发岗位信息特征挖掘与分析</td></tr>
<tr><td colspan="2">所在学院</td><td>信息工程学院</td><td>专　业</td><td colspan="2">数据科学与大数据技术</td></tr>
<tr><td colspan="2">学生姓名</td><td>郭名胜</td><td>学　号</td><td colspan="2">1903691045</td></tr>
<tr><td colspan="2">指导教师</td><td>杨杰</td><td>指导教师职称</td><td colspan="2">助教</td></tr>
<tr><td rowspan="9">指导教师评分</td><td rowspan="2">序号</td><td rowspan="2">评分项目</td><td colspan="2">评分参考</td><td>得分</td></tr>
<tr><td>评阅指标（优秀标准）</td><td>满分（100 分）</td><td></td></tr>
<tr><td>1</td><td>选题质量</td><td>符合专业培养目标；有实际意义和推广价值。</td><td>20</td><td>18</td></tr>
<tr><td>2</td><td>文献资料 应用能力</td><td>能独立查阅文献，具有收集、加工各种信息及获取新知识的能力。</td><td>10</td><td>8</td></tr>
<tr><td>3</td><td>调查研究能力</td><td>能准确理解课题任务；研究方案设计合理；能独立从事调查研究；能综合运用所学知识发现与解决实际问题。</td><td>30</td><td>27</td></tr>
<tr><td>4</td><td>论文（设计）质量</td><td>格式、图表（或图纸）规范，符合要求；结构严谨，逻辑性强；内容翔实，表达准确流畅；学术价值或实用价值高。</td><td>20</td><td>18</td></tr>
<tr><td>5</td><td>创新能力</td><td>观点独到，方法新颖，角度新颖。</td><td>10</td><td>9</td></tr>
<tr><td>6</td><td>工作量及态度</td><td>工作量饱满；能圆满完成任务书规定的各项工作。</td><td>10</td><td>9</td></tr>
<tr><td colspan="3">总得分</td><td colspan="2">89</td></tr>
<tr><td rowspan="3">指导教师评定意见</td><td colspan="5">参照上述评价标准及学生论文（设计）完成情况，做出具体评价：</td></tr>
<tr><td colspan="5"></td></tr>
<tr><td colspan="5">指导教师签名：杨杰　　2023 年 4 月 10 日</td></tr>
</table>

河南大学民生学院 2023 届本科生毕业论文（设计）

答辩成绩表

<table>
<tr><td colspan="2">论文（设计）题目</td><td colspan="5">互联网研发岗位信息特征挖掘与分析</td></tr>
<tr><td colspan="2">所在学院</td><td colspan="2">信息工程学院</td><td>专　业</td><td colspan="2">数据科学与大数据技术</td></tr>
<tr><td colspan="2">学生姓名</td><td colspan="2">郭名胜</td><td>学　号</td><td colspan="2">1903691045</td></tr>
<tr><td rowspan="4">答辩委员会（组）评分及评定结论</td><td>评分项目及分值</td><td colspan="6">答辩委员会（组）专家评分</td></tr>
<tr><td></td><td colspan="2">答　辩　情　况</td><td colspan="2">论　文　质　量</td><td rowspan="2">合计
（100 分）</td></tr>
<tr><td></td><td>内容表达情况
（15 分）</td><td>回答问题情况
（25 分）</td><td>规范要求与文字表达
（20 分）</td><td>论文（设计）质量和创新意识
（40 分）</td></tr>
<tr><td>得分</td><td>12.2</td><td>21.6</td><td>16.4</td><td>33.2</td><td>83.4</td></tr>
<tr><td>答辩委员会（组）评定结论</td><td colspan="6">理论正确，论文有一定水平，通过答辩

答辩委员会（组）签字：　沈夏炯、毛海涛、孙丽娜、袁帅、龚玲、段延超、杨杰</td></tr>
<tr><td colspan="7">毕业论文（设计）答辩成绩：　83.4　　分

答辩委员会（组）负责人签字：　沈夏炯</td></tr>
</table>

河南大学民生学院 2023 届本科生毕业论文（设计）

答辩纪要

论文（设计）题目	互联网研发岗位信息特征挖掘与分析		
答辩人姓名	郭名胜	**指导教师**	杨杰
答辩时间	**4 月 23 日上午**	**答辩地点**	**软件工程实验室**
答辩委员会(组)负责人	**沈夏炯**	**答辩委员会(组)成员**	毛海涛、孙丽娜、袁帅、龚玲、段延超、杨杰
答辩中提出的问题及回答的简要情况记录： 问题一：论文重点是什么？摘要需要改善 问题二：量化分析需要明确说明 问题三：预处理过程没写出来 问题四：工作中间过程基本没写 记录人签名：			

河南大学民生学院 2023 届本科生毕业论文（设计）

综合成绩表

<table>
<tr><td>论文（设计）题目</td><td colspan="5">互联网研发岗位信息特征挖掘与分析</td></tr>
<tr><td>所在学院</td><td colspan="2">信息工程学院</td><td>专　业</td><td colspan="2">数据科学与大数据技术</td></tr>
<tr><td>学生姓名</td><td colspan="2">郭名胜</td><td>学　号</td><td colspan="2">1903691045</td></tr>
<tr><td>指导教师评分</td><td>89</td><td>评阅教师评分</td><td>84</td><td>答辩评分</td><td>83.4</td></tr>
<tr><td>综合成绩</td><td>85</td><td>成绩等级</td><td>良好</td><td>是否推优</td><td>否</td></tr>
<tr><td>所在系（室）意见</td><td colspan="5">负责人签名：  2023 年 4 月 30 日</td></tr>
<tr><td>学院意见（仅推优论文填写）</td><td colspan="5">负责人签名（盖章）：</td></tr>
</table>

河南大学民生学院本科生毕业论文（设计）承诺书

论文（设计）题目	互联网研发岗位信息特征挖掘与分析				
姓 名	郭名胜	**学 号**	1903691045	**专 业**	数据科学与大数据技术
指导教师姓名	杨杰	**职　　称**	助教		
完成时间	2023 年 6 月 30 日				

承诺内容：

1、 **本毕业论文（设计、创作）是学生** 郭名胜 **在导师** 杨杰 **的指导下独立完成的，没有抄袭、剽窃他人成果，没有请人代做，若在毕业论文（设计）的各种检查、评比中被发现有以上行为，愿按学校有关规定接受处理，并承担相应的法律责任。**

2、 **学校有权保留并向上级有关部门送交本毕业论文（设计）的复印件和电子稿件。学校可以将毕业论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。**

备注：

学生签名： 郭名胜 **指导教师签名：** 杨杰

2023 年 4 月 20 日　　**2023 年 4 月 10 日**

互联网研发岗位信息特征挖掘与分析

摘要

本文旨在探究招聘岗位职位要求中的特征，通过对招聘岗位描述中的文本数据进行挖掘和分析，从中提取出最具代表性的特征词汇。本文通过对比分析各种特征选择方法的优劣，而后使用 Open AI 公司的 gpt-3.5-turbo 模型进行特征提取的特征选择方法，该方法可以自动筛选招聘信息中的特征词汇，而后作者又通过数据分析的方式生成招聘信息中字段之间的相关性分析结果。实验结果表明，在互联网行业，应届毕业生与企业要求的岗位技能之间存在诸多差距，并且提取出的特征词汇对于应届毕业生的求职存在支持作用。

关键词：特征挖掘；特征选择；相关性分析

Mining and Analysis of Information Characteristics of Internet R&D Posts

ABSTRACT

The purpose of this paper is to explore the characteristics of the job requirements of the recruitment positions, and extract the most representative characteristic words by mining and analyzing the text data in the job descriptions of the recruitment positions. This article compares and analyzes the advantages and disadvantages of various feature selection methods, and then uses the gpt-3.5-turbo model of Open AI Company to perform feature selection method for feature extraction. This method can automatically screen the feature words in the recruitment information, and then the author uses the data The analysis method generates the correlation analysis results between the fields in the recruitment information. The experimental results show that in the Internet industry, there are many gaps between fresh graduates and the job skills required by enterprises, and the extracted characteristic vocabulary can support the job hunting of fresh graduates.

Keywords： feature mining; feature selection; correlation analysis

目录

1 绪论

当前，随着人们对高质量的工作的需求不断增加，在招聘市场中，求职者之间的竞争也变得越来越激烈。对于企业来说，寻找到符合企业要求的候选人变得越来越困难，而对于求职者来说，了解企业并让自己达到企业的相关岗位要求也变得越来越重要。因此，深入探究招聘岗位任职要求的特征对于企业招聘与个人职业发展具有重要意义。对招聘岗位任职要求中的特征进行挖掘和分析，有助于求职者更好地了解企业的需求，帮助求职者更好地提高个人能力并匹配这些岗位要求。同时，通过使用更为优秀的技术方法，例如自然语言处理和机器学习，可以提高特征选择结果的准确性和效率。本文旨在探究招聘岗位任职要求中的特征，通过对招聘岗位描述中的文本数据进行挖掘和分析，并从中提取出最具代表性的特征词汇，从而为企业和求职者提供更有价值的信息。

1.1 研究背景及意义

随着互联网行业的迅速发展，企业相关研发类岗位的需求量也在不断增加，但应届毕业生就业难的问题却依然严峻。因此，对于求职者来说，如何更好地了解并掌握企业对于求职者所需技能的相关要求，成为了一个重要的问题。在这个背景下，通过对招聘岗位任职信息中的文本数据进行挖掘和分析，并从中提取出最具代表性的特征，对于求职者在求职过程中提高个人竞争力具有一定的意义。同时，近年来不断有学者探究利用自然语言处理技术来辅助招聘决策，但在特征选择方面仍存在诸多挑战。为此，通过使用基于 Open AI 的 gpt-3.5-turbo 模型接口的相关特征选择方法，筛选出对于招聘决策具有重要意义的特征，具有一定的研究价值和实用意义。因此，本文旨在探究招聘岗位任职要求中的特征，并分析其与应届毕业生技能的匹配度，为应届毕业生提供一定的支持和帮助，同时对于校企合作的发展和求职者的针对性学习也具有一定价值。

1.2 国内外研究现状

随着社会的发展和技术的进步，招聘岗位任职要求的特征研究也受到了国内外学者的广泛关注。总体来说，国内外学者在招聘岗位任职要求的特征挖掘分析研究方面都已经取得了较好的进展，但仍然存在一些亟待解决的问题。例如，如何将招聘岗位任职要求中的特征与求职者的实际情况相结合，并为求职者提供更具有针对性的求职建议等，

这些问题也将成为将来研究的重点。

1.2.1 国内研究现状

在国内，随着国家经济的快速发展，越来越多的毕业生进入就业市场，此时，企业招聘困难的问题却日益凸显。近年来，一些国内学者开始关注招聘岗位任职要求的特征研究，并在此基础上开展了一系列的分析。宋齐明博士早在 2018 年就已经在关于本科生可就业能力的研究中提出，在互联网行业，应届毕业生的实际技能与企业要求存在较大差距[1]，这也是造成应届毕业生求职难的主要原因之一。对于招聘岗位需求特征挖掘所用到的有关文本相似性度量技术，在 2011 年，黄承慧等人提出了一种结合词项语义信息和 TF-IDF 方法的文本相似性度量的方法[2]。张俊峰等人也与此基础上于 2018 年通过对国内招聘网站岗位的数据挖掘分析中提出了有关数据类学科人才培养方案中的一些看法[3]。

1.2.2 国外研究现状

在国外研究中，研究人员通常通过采用文本挖掘和数据分析技术来分析互联网研发岗位信息中的特征。文本挖掘技术可以自动化地从海量数据中提取出有用的信息，例如招聘信息中的技能要求、工作经验和薪资水平等。数据分析技术则可以对这些信息进行统计和分析，从而得出互联网研发岗位招聘信息的一些趋势和特征。

研究表明，互联网研发岗位信息具有一些特定的特征。例如，互联网研发岗位通常要求应聘者掌握多项技能，如编程语言、数据库技术和系统架构等。此外，对工作经验的要求也较高，通常需要具备数年的实际工作经验。在薪资水平方面，互联网研发岗位相对其他行业来说较高，但不同国家和地区的薪资水平存在差异。

总之，互联网研发岗位信息特征挖掘与分析的国外研究现状涉及到文本挖掘和数据分析技术的应用，可以为求职者和企业提供更好的就业和招聘信息。这个领域的研究已经在国外引起了广泛的关注，而且随着技术的不断发展，该领域的研究将进一步深入和广泛。陈飞博士也在 2014 年对德国应用型本科教育进行了具体分析，提出应用型本科教育课程应及时反应经济社会需求的观点[4]，同时提出了一些关于国内应用型本科教育改革的一些探讨方案。

1.2 主要工作

本文的研究主要分为以下若干个部分：

（一）、采用文本数据采集技术对特定网站进行了招聘信息数据采集，并将获取到的数据存储至数据库中。

（二）、通过调用 Open AI 公司提供的基于 gpt-3.5-turbo 模型接口，免去了对网页采集到的文本数据之后对数据进行特征挖掘时构建分词词典与对关键词的相关性分析步骤。从而实现了对数据库中的招聘要求字段进行特征词汇的提取工作。

（三）、最终，通过对提取的数据进行预处理操作，我们而后将进行更加深入的分析，并通过可视化技术展示其分析结果。

通过以上步骤，我们得到了一份针对某特定行业的招聘要求特征词汇的数据集，从中提取出了最具代表性的特征词汇，并得到一份其中字段的相关性数据分析结果，为进一步探究该行业的招聘趋势和任职要求提供了有力的支持。

1.3 本章小结

本文旨在运用自然语言处理模型对招聘岗位任职要求中的信息进行特征挖掘和分析，通过分析招聘岗位描述中的文本数据，提取出最具代表性的特征，并探究了招聘岗位各字段之间的相关性。为了实现这一目标，我们采用了一种基于 Open AI 公司的 gpt-3.5-turbo 模型的特征选择方法，该方法可以自动筛选并提取出对于招聘决策具有重要意义的特征词汇。

在实验过程中，我们以招聘网站上的招聘岗位数据为基础，通过文本挖掘技术进行信息特征挖掘和分析，并生成了不同字段之间的相关性分析结果。实验结果表明，通过此方式提取的特征词汇得出的实验结果可以为应届毕业生提供一定的信息支持。同时，其相应的数据分析结果对于校企合作的发展和应届求职者的针对性学习具有一定的参考价值。

2 数据处理

在传统的数据分析中需要大量的问卷信息统计以及人力进行统计分析，而在现代数据分析中，我们可以利用各种现代技术获取数据来源。在本研究中，我们将采用同于 Web 招聘信息的文本挖掘系统研究中[5]使用的网络爬虫技术从多个知名的招聘网站上获取有关招聘信息的数据。这些网站包括 51job、智联招聘、拉勾网等，它们拥有大量的求职者和企业用户，并提供了丰富的招聘信息。我们可以通过爬虫程序自动抓取这些网站上的招聘信息，并将其转化为结构化数据进行分析。

采集和预处理数据是数据分析中至关重要的步骤，因为它们的质量直接关系到后续分析的可信度和精度。在本研究中，我们需要对从招聘网站采集到的数据进行以下步骤的预处理：

（一）数据清洗：对采集到的原始数据进行初步的清洗和处理，去除无用信息和重复数据，纠正数据中的错误。

（二）数据转换：将从招聘网站采集到的非结构化数据转化为结构化数据，并按照一定格式进行整理和编排，例如将文本数据转换为表格形式，以方便进行后续的数据分析。

（三）数据集成：将来自不同来源的数据整合到一起，去除重复信息，从而得到一个完整的数据集的步骤。

（四）数据归约：对数据进行抽样、压缩和规约等处理，可以将数据集的规模压缩到合理的范围内。

（五）数据标准化：标准化处理是对来自不同来源的数据进行处理，以保证它们具有相同的格式和单位。

通过以上步骤的处理，可以为后续的数据分析工作提供可靠的基础数据支持，并将采集到的原始数据转换为结构化的数据。

2.1 数据采集

在过去的数据采集中，通常通过填写线下问卷的方式进行数据收集。然而，这种方式存在许多明显的缺点。首先，因为线下问卷需要印刷、分发和回收等过程，相较于在线调查等方式，收集数据的速度较慢。其次，由于问卷填写过程中存在人为因素，如回

答不认真、漏填、填错等情况，因此数据质量难以保证。此外，线下问卷收集的数据需要手动输入到电脑中进行数据分析，相较于在线调查等方式，效率较低，也容易出现误差。最后，线下问卷需要印刷、分发、回收等成本，而且需要人力进行数据录入和分析，相较于在线调查等方式，成本较高。

因此，在现代数据分析中，越来越多的人开始采用在线调查等方式进行数据收集。对于企业招聘信息的收集，线下问卷方式不仅难以实现，而且数据质量和真实性也难以保证。在这种情况下，互联网招聘网站上面向求职者的招聘信息就具有了非常高的参考价值。

2.1.1 数据来源概述

数据源是进行数据分析的基石，之后的所有分析结果都建立在数据的基础之上。高质量的数据不仅可以在分析的过程中减少错误和偏差，而且可以分析得出更为高质量的结论。因此，选择适当的数据源在数据分析中起到至关重要的作用。

选择适当数据源的几个关键点在于需要考虑到数据的可靠性、相关性、质量、数量、格式、保密性等。因此本课题选取了互联网中较为流行的一个招聘网站进行数据采集。以下为招聘网站中的某公司有关 JAVA 开发岗位的招聘信息：

图 2.1 某招聘网站的某一职位信息

如图 2.1，此类网站所提供的字段主要包括：职位的基本信息（职位名称、职位所在城市、职位工作经验要求、职位学历要求、职位发布时间等），企业的基本属性（企业名称，企业类型、企业所在行业、企业规模等），职位的描述（岗位描述、岗位要求等），而在岗位要求中包含多角度的需求特征，也是本研究的数据预处理过程中的主要处理信息。

2.1.2 数据采集过程

网页的数据展示主要通过客户端向服务器发送请求并接收服务器返回的包含 HTML、CSS 和 JavaScript 等前端技术的代码之后，再经过浏览器的渲染而实现。HTML 定义了网页内容的结构和语义，CSS 则定义了网页的样式和布局，而 JavaScript 则通过为网页添加交互性和动态性，使得网页更加丰富和生动。

在数据的采集的过程中则需要将网页代码转换为可供分析的结构化数据以供进行数据分析使用，因此则需要去除其中的各种 HTML 标签、CSS 样式、以及 JavaScript 代码部分并将其中我们所需要的数据提取保留。在 Python 语言中，我们可以使用它的第三方库 requests 来发送 HTTP 请求以获取网页的内容，但是如果目标网站存在反爬虫机制，使用 requests 很有可能会被服务器拒绝访问。为了解决这个问题，一种可行的方法是使用浏览器模拟请求的方式来请求网站。Selenium 是一种自动化测试工具，它可以模拟用户在浏览器上的各种操作，例如打开网页、输入关键字、点击按钮等。通过 Selenium，我们可以使用 Python 代码来模拟浏览器上的一些操作，同时获取到网页内容。

在使用爬虫抓取网页数据时，我们应该遵守网站的 robots.txt 协议，不应过度访问网站，以避免对服务器造成不必要的压力。同时，也不应使用爬虫来做一些获取敏感信息或者侵犯他人隐私的行为。

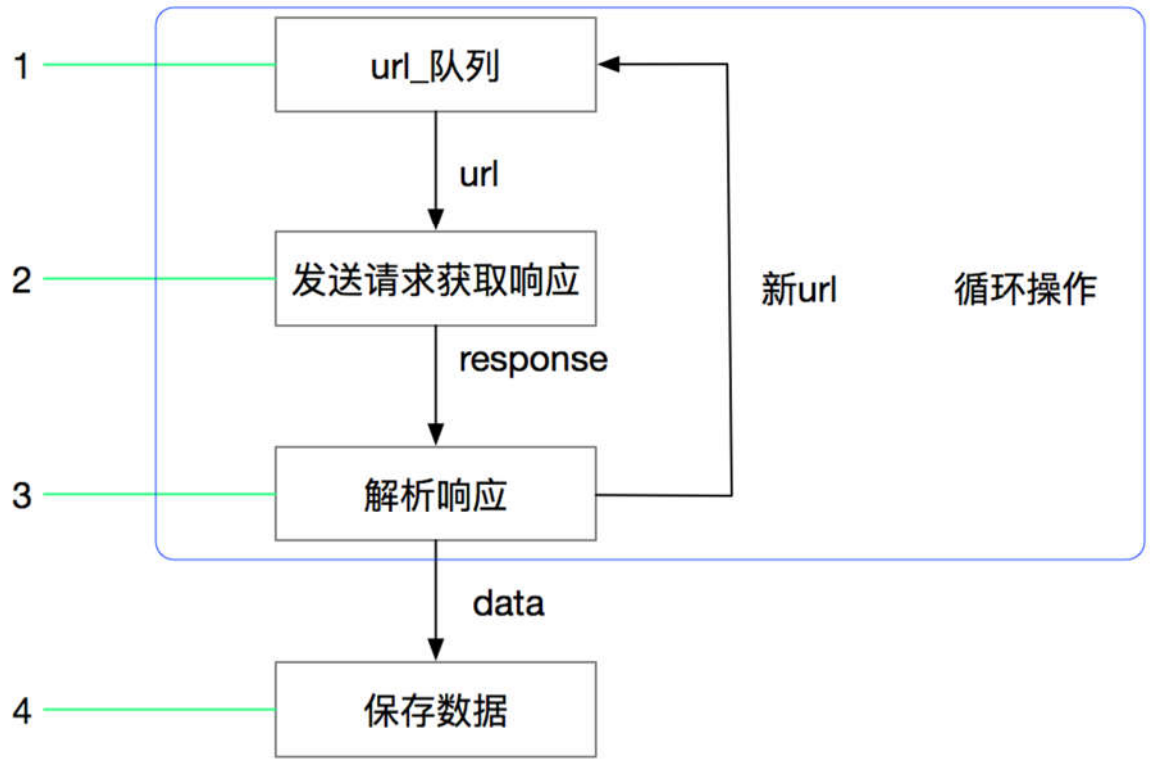


图 2.2 数据采集的主要流程

图 2.2 中介绍了数据采集的主要过程，将待爬取网页的数据链接插入至 url 队列中并向服务器发送请求获取响应，解析到数据之后再将其中的解析到的需要再次请求的新链接插入至 url 请求队列，以供再次请求。若解析到的是待分析的数据，则将其保存在数据库或本地文件中供数据分析使用。

需要注意的是，在进行网页数据采集时，被采集的网站可能会出现滑块验证的页面以验证是手动进行的访问，这是由于该网站的反爬机制。在正常的浏览器验证中，可以通过验证并继续获取网页数据。然而，在使用 selenium 进行人工滑动验证时，有可能仍然会弹出验证失败的提示。

经过测试，笔者找到一种可行的解决方法：通过修改 selenium 中某一标识字符，然后通过 selenium 模拟人工进行滑块验证，即可再次获取网页数据。

理论上，这时可以不断请求该网站获取到该网站的服务器数据。但是，在短时间多次对该网站进行页面的访问操作之后，发现仍然会被该网站检测到并封禁 IP，这时我们可以通过使用代理 IP 的方式来请求获取页面数据。这样，每次请求时，参与请求的 IP 地址都会发生变化，从而避免被网站检测到并拒绝进行访问。

2.1.3 数据的初步结构化处理

(1) 数据提取

通过对网页获取请求之后，我们得到的仅仅是存在众多繁杂的无用代码字符的数据据，这时候就需要提取出其中的结构化数据并存入数据库或本地文件中以供数据分析使用。

获取到的响应内容主要有以下两种：

（一）、结构化的响应内容：主要包括 json 字符串与 xml 字符串等，此类数据可以通过正则匹配的方法或使用其所对应的 jsonpath、lxml 等模块中的相关语法进行数据提取。

（二）、非结构化的响应内容：主要包括 html 字符串等，此类数据一般可以通过正则匹配的方法或使用 lxml 模块中的 xpath 语法进行数据提取。

(2) 数据存储

笔者使用了 MySql 数据库进行了数据的存储，并在插入之前对数据库中已有的数据行查询操作，以避免再次请求重复的网页数据和存储重复的待分析数据。数据表 job_table 主要包括的字段信息如表 2.1 所示。

表 2.1 供存储采集数据和中间处理数据的数据库表 job_table

序号	字段名称	字段类型	允许为空	最大长度	备注
1	unique_key	varchar	True	255	数据唯一标识符
2	job_name	varchar	True	99	职位名称
3	job_detail	varchar	True	9999	职位详情
4	job_skills	varchar	True	255	所需技能
5	job_academic	varchar	True	255	学历要求
6	job_year	varchar	True	255	工作经验要求
7	job_welfare	varchar	True	255	福利
8	job_compensation	varchar	True	255	工作薪资
9	job_city	varchar	True	255	工作城市
10	job_adress	varchar	True	255	工作地址
11	company_name	varchar	True	255	招聘公司
12	company_type	varchar	True	255	公司类型
13	company_industry	varchar	True	255	公司行业
14	release_time	varchar	True	255	发布时间
15	data_link	varchar	True	255	招聘数据链接
16	data_source	varchar	True	255	数据来源网站
17	job_module	varchar	True	255	数据所在模块
18	detail_result	varchar	True	999	特征提取结果

2.2 数据预处理

经过初步预处理之后的数据已经大致包括了我们所需要的信息，但是在数据分析的

过程中，仍需要将数据处理为可以被计算机直接识别并进行统计分析的数据。如将不同的工作城市统一至其所在的省份、不包含工作经验要求年限的数据类型统一为 int 型数据、不同的薪资计算方式统一为按月计算等。

在数据预处理的过程中，招聘信息中职位需求信息的挖掘，就需要在较多的文本中提取出其中最具代表性的技术类词汇，因此需要通过职位需求特征挖掘的相关技术进行分析获取结果。

2.2.1 职位需求特征挖掘关键技术

招聘信息中的岗位要求属于文本类数据，因此对于职位需求特征的提取处理过程就是要从文本信息中自动抽取出其中与文本主题相关的关键词或短语的过程，即从大量文本数据中提取有用信息的过程。文本关键词提取的方法主要有以下几种：基于词频的方法、基于 TF-IDF 的方法、基于文本分类的方法、基于主题模型的方法以及基于网络结构的方法等。

笔者主要介绍了本研究中三种可行的解决方法：基于 TF-IDF 的方法、基于 word2vec 模型的关键词提取方法和基于 gpt-3.5-turbo 模型的提取方法。

(1) TF-IDF

TF-IDF 是一项被广泛应用于文本挖掘和信息检索领域的技术。TF-IDF 技术根据词频和逆文档频率的乘积来确定每个词在文档中的重要性，即每个词都被赋予一个权重，用于衡量其对文档的影响程度。TF-IDF 的基本思想是，如果一个词在一篇文档中出现很多次，那么它可能在这篇文档中非常重要，但如果它在语料库中的其他文档中也同样频繁地出现，那么它可能是一个常见的单词，对于文档的区分能力就不大。

具体来说，TF-IDF 算法将会计算一个词在文档中的 TF 值，然后根据该词在整个语料库中出现的文档频率计算它的 IDF 值，最后得到两个值的乘积即为该词的 TF-IDF 值。公式如（2.1）所示：

$$TF-IDF(w)=\frac{\text{在文档中词}w\text{出现的次数}}{\text{文档中总词数}}*\log_e\left(\frac{\text{文集中文档总数}}{\text{包含词 w 的文档数}}\right) \tag{2.1}$$

$TF-IDF(w)$公式中$TF(w)$表示词频，是一个词语 w 在文档中出现的词频数，$TF(w)$通常用于计算一个词在一个文档中的重要程度或权重。其中，一个词在单个文档中多次出现，但在整个语料库中出现次数很少，那么它在该文档中的权重就会更高。公式如（2.2）

所示：

$$TF(w) = \frac{\text{在文档中词}w\text{出现的次数}}{\text{文档中总词数}} \tag{2.2}$$

$TF - IDF(w)$公式中IDF(w)指一个词语 w 在语料库中出现的文档频率的倒数，即逆文档频率（Inverse Document Frequency），IDF(w)用于评估一个词语在整个语料库中的重要性或权重，其中，一个词语在整个语料库中出现的文档频率越小，它的重要性或权重就越高。公式如（2.3）所示：

$$\mathrm{IDF(w)} = \log_{\mathrm{e}}\left(\frac{\text{文集中文本总数}}{\text{包含词 w 的文档数}}\right) \tag{2.3}$$

TF-IDF 是一种简单有效的评估文本中词语重要性的方法，它能够反映词语在文本中的重要性评分，并且对于词语的重要性可以进行排序，具有较强的可解释性。但是 TF-IDF 忽略了词语的顺序信息，对于长文本的处理存在一定的问题，而且对于专业术语或领域词汇的重要性评估可能需要进行特殊处理。

(2) Word2Vec

Word2Vec 是一种用于生成词向量（Word Embedding）的神经网络模型。它是由谷歌（Google）研发的一种无监督学习算法，用于将词语表示为实数值向量，并将这些向量用于自然语言处理（NLP）任务中。

Word2Vec 模型主要有两种实现方法，分别是连续词袋模型（Continuous Bag of Words，CBOW）和跳字模型（Skip-Gram）。CBOW 模型是一种基于上下文预测中心词的模型，而 Skip-Gram 模型则是一种基于中心词预测上下文词的模型。这两种模型的核心思想都是基于分布假设，即相似的词在上下文中出现的频率也是相似的。

Word2Vec 模型的训练过程是通过对大量文本数据进行神经网络训练，不断调整词向量的值，使得每个词语的向量在空间中能够相互接近和区分出来。训练完成之后，每个词语都会有一个固定维度的向量表示，这些向量就可以作为输入数据用于各种自然语言处理任务当中，例如文本分类、情感分析、语义相似性计算等。

Word2Vec 具有以下优点：能够将文本中的语义信息表示为实数向量，方便进行数学计算和比较；能够利用无监督学习的方法，自动地从海量文本数据中学习出词汇的特征表示，从而无需手动设计特征；可以以此来解决一些数据高维和稀疏的问题。然而，它也有一些缺点：对低频词不敏感、无法处理多义词、无法处理词序信息、对于长文本处

理较慢、对于一些语言表达形式的处理不够充分。

(3) GPT-3.5-turbo

ChatGPT 是一种基于 GPT（Generative Pre-trained Transformer）-3.5 的大型语言模型，GPT-3.5-turbo 是 OpenAI 公司向公众开放的可供开发者使用并处理自然语言处理任务的接口模型。GPT 是一种基于 Transformor 架构的神经网络模型，其核心思想是使用大量的未标记文本数据进行预训练，然后在特定任务上进行微调。

ChatGPT 使用了大量的非结构化文本数据进行预训练，这些文本数据包括维基百科、新闻、小说、社交媒体、网站等等。在预训练过程中，模型学习如何对输入的文本进行编码，并在不同任务上生成自然语言文本。

ChatGPT 可以用于各种自然语言处理任务，包括对话系统、语言翻译、文本摘要、问题回答等等。它的优势在于其可以自动生成自然语言文本，而不需要人工编写规则或手工标注数据，这使得它可以处理大量的非结构化数据，并具有很好的泛化能力。但是在某些情况下，ChatGPT 也会给出一些看似正确但实际是错误的结果，而且 OpenAI 公司提供的模型接口具有一定的 token 限制，因此并不能处理过长的文本任务。但在招聘要求中的任职要求关键词的提取工作中，ChatGPT 具有相对较好的提取结果。

2.2.2 职位需求特征挖掘过程分析与技术选型

对于职位需求特征挖掘主要的步骤有：特征词汇词典数据源选取，特征词汇词典构建，职位需求字段分词，Word2Vec 模型训练，特征词汇相关性分析与提取等过程。

(1) 特征词汇词典数据源选取

特征词汇词典数据源的选取可以通过在搜索引擎中搜索相关岗位特征、从专业领域相关书籍或 IT 培训机构课程大纲中获取信息等方式进行数据源选择。而在选择好数据源之后，其中在互联网上存在的数据源仍然可以通过网页数据采集的方式进行数据提取，同时对于采集到的结果数据中可能仍然是具有完整语义的句子，这时候还需要对其中需要用到的特征词汇进行词汇选取的工作。

(2) 特征词汇词典构建

对于此类特征词汇词典的构建来说，词典在建成之后的维护成本很低。因此我们主要可以人工从数据源中选取出以下几类词汇：在招聘岗位要求中出现较多的词汇、专业性较强的词汇、同义词的不同表达等。

而在中文的岗位要求中，有时会出现一些英文的技术类词汇，如 Mysql、Java、Python 等。对于此类词汇，我们可以直接通过正则匹配的方式进行选取，可以省去人工选择的一些步骤。

(3) 职位需求字段分词

完成特征词汇词典的构建之后，我们就可以对招聘信息中的岗位要求字段进行处理。笔者的一种处理思路是使用 jieba 分词将已构建完成的词典作为参数进行分词提取。在模型训练的过程中，我们可以使用 jieba 分词提供的全匹配模式。而在职位特征词汇提取的过程中可以使用精确匹配模式从而将其中的意义不完全的词语去除。

(4) Word2Vec 模型训练

完成特征词汇的选取之后，我们在提取的过程中可以使用全匹配模式进行模型的训练，这样可以提高训练的参数数量，从而可以在一定程度上提高训练模型的准确度。模型训练的关键代码如图 2.3 所示：

调用清理函数并生成训练

```
import gensim
from gensim.models import Word2Vec

# 得出训练词汇列表
lines = update(data)
# 清理停用词
lines = clean(lines)

# 调用Word2Vec训练
# 参数：size: 词向量维度；window: 上下文的宽度，min_count为考虑计算的单词的最低词频阈值
w2v_model = Word2Vec(lines, vector_size = 40, window = 2 , min_count = 2, epochs=7, negative=10,sg=1)
```

图 2.3 Word2Vec 模型训练的关键代码

updata 函数返回一个包含特征词汇子列表的列表，其中每个子列表包含一条职位信息中的特征词汇。然而，这些子列表中可能会包含许多与特征词汇无关的词汇。如果将这些无关词汇全部去除，可能会丢失一些并未在特征词汇词典中出现但具有重要意义的词汇。因此，可以通过使用 clean 函数清理停用词后再进行模型训练。

在进行模型训练时，需要传入参数来调整 word2vec 模型。不同的参数可能会导致不同的训练结果和精度，因此需要多次进行实验和比较，以找到最优的参数值。

(5) 特征词汇提取

在训练好数据模型之后，就可以通过调用 word2vec 中的 most_similar 方法获得与已知词汇相关性最高的一些词汇，如图 2.4 中通过调用该方法打印训练模型中与"mysql"相关性最高的五个词汇及其在该训练模型中的相关性元组：

```
w2v_model.wv.most_similar("mysql", topn = 5)
```

```
[('postgresql', 0.8797929286956787),
 ('nosql', 0.8716303706169128),
 ('mysql等', 0.8520218729972839),
 ('synapse', 0.8494269251823425),
 ('nosql数据库', 0.8479032516479492)]
```

图 2.4 获取与 MySQL 相关性最高的五个词汇

我们可以使用相同的方法来获取与特征词汇词典相关性较高的一部分尚未添加到特征词汇词典中的特征词汇。然后，我们可以将这些词汇转换为待分析词汇列表，并将其存储到数据库中。

(6) 提取结果分析

根据以上提取流程，我们已经初步得到了一组相对较好的关键词提取结果。然而，该流程对于构建特征词汇词典的质量要求较高。例如，如果我们的词典中不包括"sprint boot"这个词汇，结巴分词则会将其识别为"spring"和"boot"两个单词。这将导致分析数据的结果不包括"sprint boot"这个词汇，进而影响分析结果的准确性。同时，在总数据量较少的分组中，模型的准确度会很低，导致模型失去其该有的意义。

(7) 提取过程总结

为了克服以上职位需求特征词汇提取方法的局限性，作者采用了 GPT-3.5-turbo 模型提供的接口进行关键词提取。GPT-3.5-turbo 是一种先进的自然语言处理模型，能够生成高质量的文本，并用于文本分类、语言翻译、关键词提取等任务。笔者通过使用这个模型能够获得更为精确的关键词提取结果。

将数据分别调用模型的接口获取到了众多包含提取结果的词汇列表，而后我们对数据进行了预处理，包括去除停用词、无用字符和无关词汇等，以提高关键词提取结果的准确性。然后将获取到的关键词提取结果并将其保存到本地表格中，以便进行后续的数据分析和处理。

2.2.3 其他字段预处理

数据中的其他在数据分析过程中需要用到的字段主要有：工作经验要求字段、学历要求字段、薪资待遇字段和工作所在省份字段等。而主要是通过以下几种方式对数据进行预处理的操作：

（一）、将无经验要求的职位设置为零，再从通过正则表达式方式获取其中的数字并将其中取得的工作经验要求范围中的最小值作为待分析字段；

（二）、将学历要求字段中不包含学历要求字段信息的数据设置为无学历要求，从而在后续分组中可以将此类数据划分为一个分组；

（三）、将薪资待遇字段中的不同计算薪资方式通过正则匹配的方式将其统一为按月计算，并将其中的文字数据如万、千等词修改为数值型数据从而方便对此类数据进行求平均值的操作。

（四）、通过检索工作所在城市中的城市信息以及获取其在 json 文件中的某一个省份信息从而获取该职位的工作所在省份信息。

3 数据分析

处理完数据预处理步骤后，我们可以利用统计分析方法来分析数据。本文作者主要将分析工作分为以下几个模块，对每个表格字段与其他字段的相关性进行分析。

3.1 职位需求模块

在职位需求模块中，我们主要分析了职位需求特征与学历、职位类别、薪资和职位工作经验要求之间的关系。

3.1.1 不同学历下的职位需求特征分析

首先，使用关键字 JAVA 筛选包含 JAVA 的职位名称，然后按学历字段分组并筛选出关键字占比较高的数据。接下来，对每条数据中的职位需求特征词汇进行统计分析，并制作不同分组中的占比折线图如 3.1 所示。

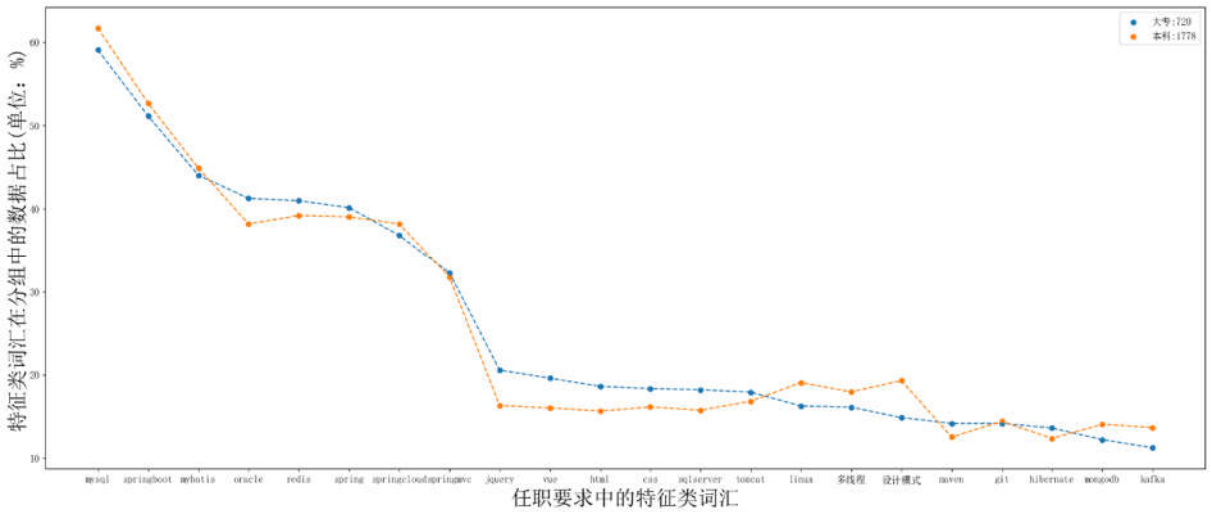


图 3.1 JAVA 类职位招聘要求中特征词汇占比图

根据图像，可以推断出 Java 类职位中涉及 MySQL 和 Spring Boot 技术的需求非常高，超过了其他流行技术的出现比例。这意味着，对于那些想要从事 Java 类职位的求职者来说，熟练掌握 MySQL 和 Spring Boot 技术将是非常有价值的。同时，对于那些已经掌握这些技术的求职者来说，他们在市场上将会有更多的机会来寻找合适的工作岗位。

然而，这并不意味着其他技术不重要。通过观察图像可以看到，其他较流行技术的出现比例也在 20%至 50%之间不等。因此，对于那些想要在 Java 类职位中获得成功的求职者来说，他们应该保持对各种适时技术的了解，并不断学习和提高自己的技能，以便

能够满足不同工作岗位的需求。

3.1.2 不同职位类别下的需求特征分析

对职位需求类词汇进行统计分析，可以通过将数据根据职位所在模块进行分组，得到如下图 3.2 的饼图和 3.3 的词云图。

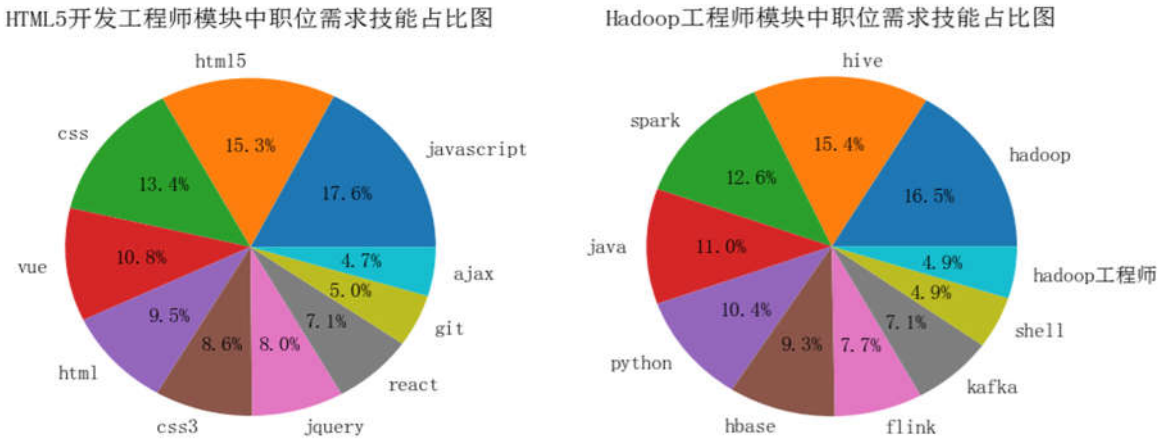


图 3.2 前端与 Hadoop 开发工程师的职位需求相对占比图

图 3.3 JAVA 与大数据开发工程师的职位需求词云图

可以看出其中哪些关键词技术是最常出现的，以及它们在其所在分组中所占的比例。这可以帮助您了解市场上最受欢迎的职位需求类词汇，从而更好地准备求职材料，提高自己的竞争力。

3.1.3 不同薪资下的需求特征分析

（1）技能价值折线图

根据如 3.1.1 中的分析方法可以得到 JAVA 工程师岗位中的可供参考的技能价值，技能价值公式如公式 3.1 所示：

$$\left(\frac{最高}{最低}\right)技能价值=\frac{该技能平均\left(\frac{最高}{最低}\right)薪资*该词汇在分组中的出现次数}{分组中的数据总数} \quad 3.1$$

其中技能价值表示该技能在该类岗位中的重要程度与可供参考的价值月薪，如 3.4、3.5 是 JAVA 类岗位的技能价值折线图：

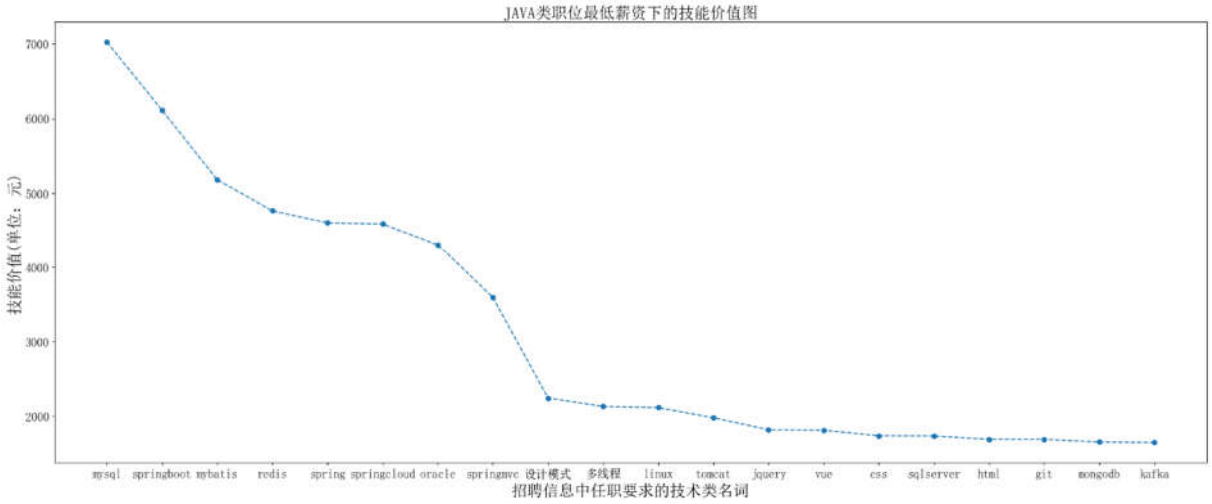


图 3.4 JAVA 类职位平均最低薪资下的技能价值

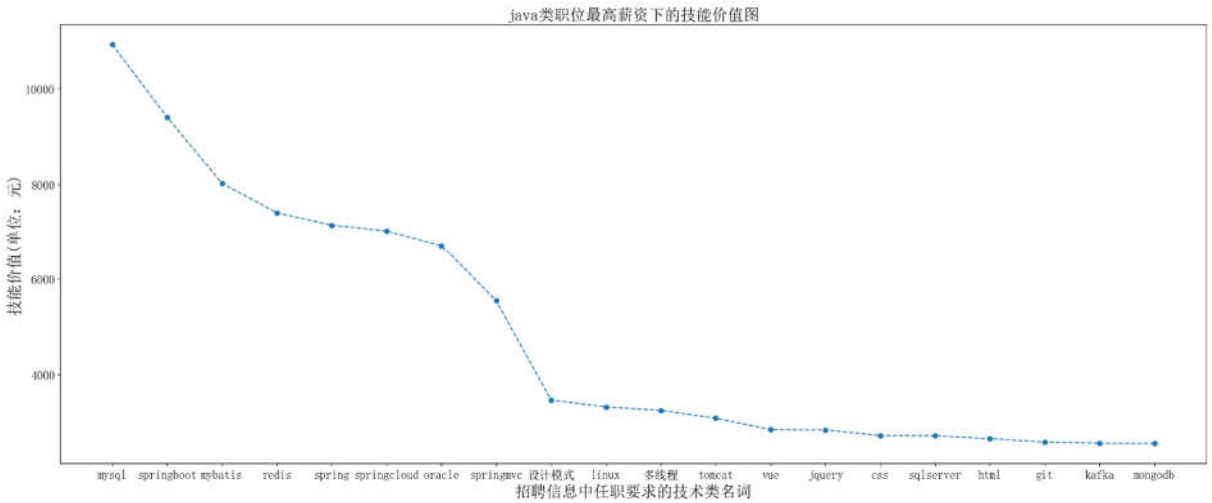


图 3.5 JAVA 类职位平均最高薪资下的技能价值

(2) 薪资范围饼图

图 3.6 不同薪资范围之间的职位需求占比饼图

通过分析图 3.6 展示的不同薪资范围的 Java 职位需求特征占比图，我们发现对 Redis 技术的要求占比显著提高。Redis 被设计用于提供快速、可扩展、可靠的数据存储解决方案，并可用作缓存、消息队列、应用程序数据库等。因此，对于个人职业发展来说，在掌握基本技能的基础上，也应该注重提升自身此类高性能、可扩展等的相关技能。

3.1.4 不同工作经验下的需求特征分析

根据如 3.1.1 中的分析方法可以得到 JAVA 工程师岗位中不同工作经验要求下的职位需求特征占比图如 3.7 所示。

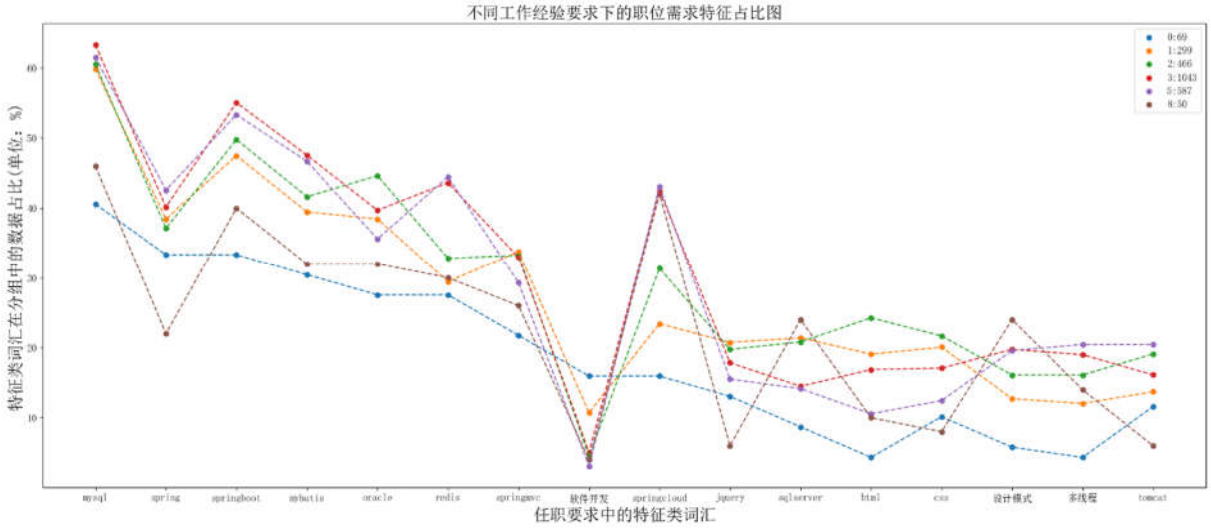


图 3.7 JAVA 工程师岗位中不同工作经验要求下的职位需求特征占比图

通过观察图像可知，随着工作经验要求的增加，Spring Cloud 技术的比重不断提高。由此可以得出结论，此类技能在个人职业生涯的发展历程中具有一定的积极作用。

3.2 薪资模块

在薪资模块中，我们主要分析了不同工作经验要求下的薪资与学历、职位类别、工作城市所在地区之间的关系图。

3.2.1 不同学历与工作经验要求下的薪资

经过对 JAVA 类岗位进行筛选，并按照学历进行分组后，可以得到图 3.8 和图 3.9 所示的折线图。

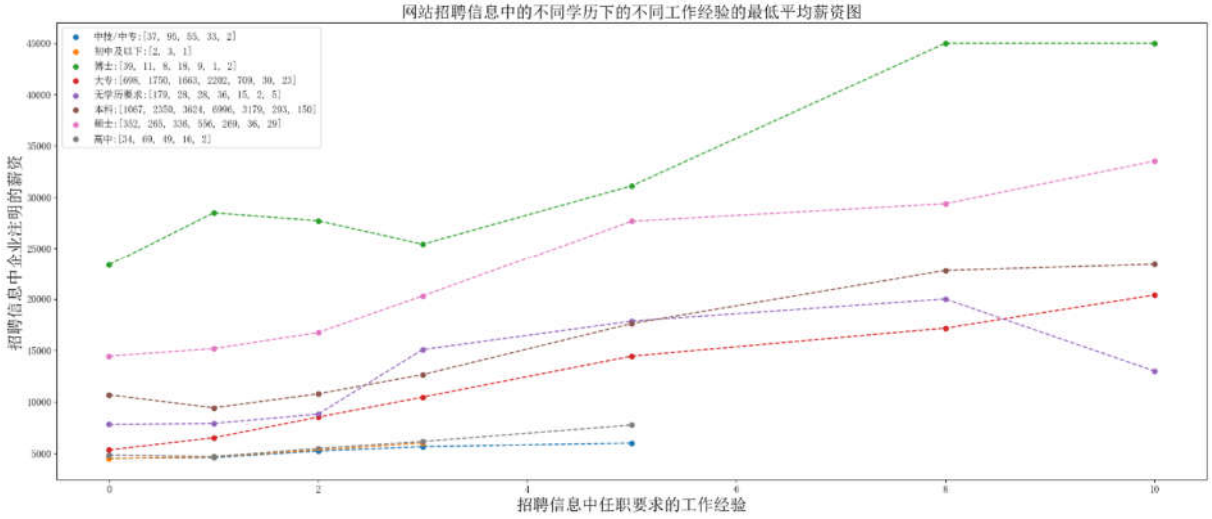


图 3.8 JAVA 类岗位不同学历下与工作经验要求下的最低平均薪资图

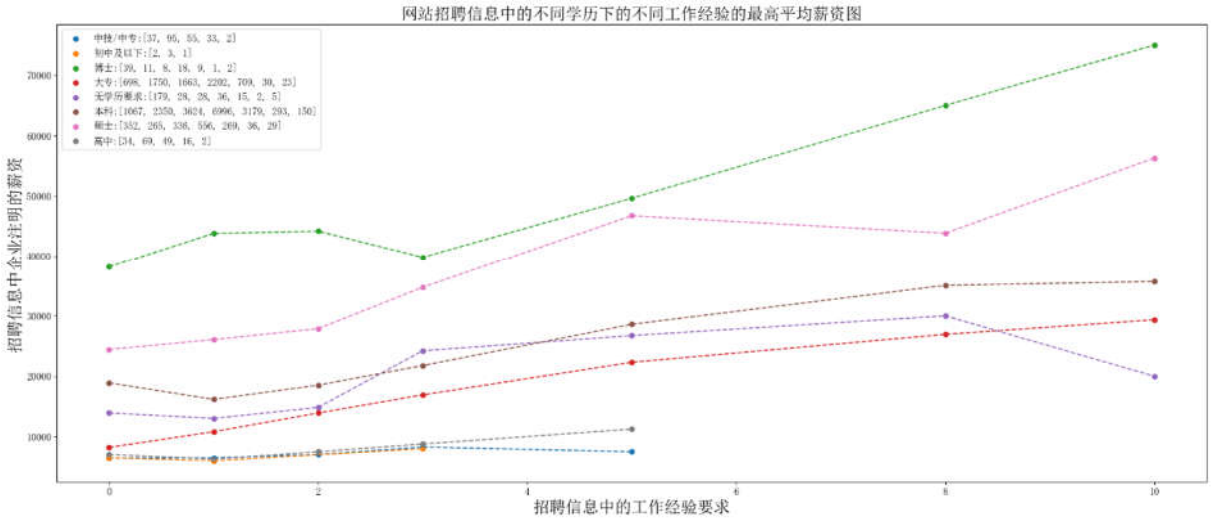


图 3.9 JAVA 类岗位不同学历下与工作经验要求下的最高平均薪资图

观察数据可以发现，具备博士学历的从业者在三年后的薪资涨幅呈现直线上升的趋势。这表明，在 JAVA 开发领域，高学历对职业发展同样也有着一定积极的影响。

3.2.2 不同职位类别与经验要求下的薪资

经过对不同职位类别的岗位进行筛选，并按照薪资进行分组后，可以得到图 3.10 和图 3.11 所示的折线图。

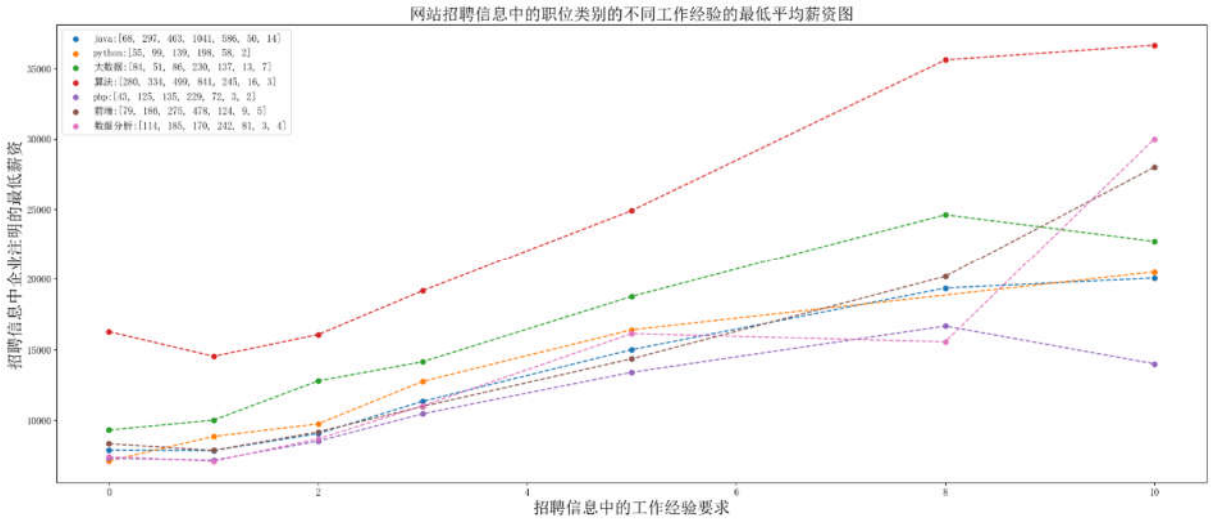


图 3.10 不同职位类别与工作经验要求下的最低平均薪资图

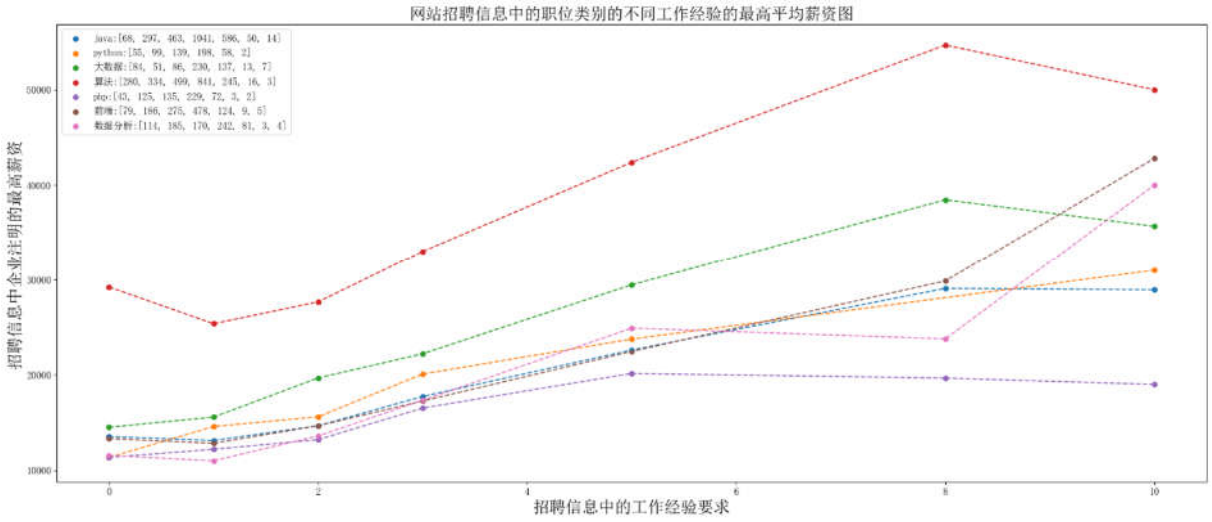


图 3.11 不同职位类别与工作经验要求下的最高平均薪资图

通过观察以上几种职位类型的薪资水平，我们可以发现算法岗位一直处于领先地位，这表明算法相关岗位是一种高薪且对个人能力要求较高的研发岗位。

3.2.3 不同地区与工作经验要求下的薪资

对工作所在城市的所在地区统计，并按照工作经验分组之后对薪资进行求平均值之后可以得到如图 3.12 和 3.13 所示的折线图。

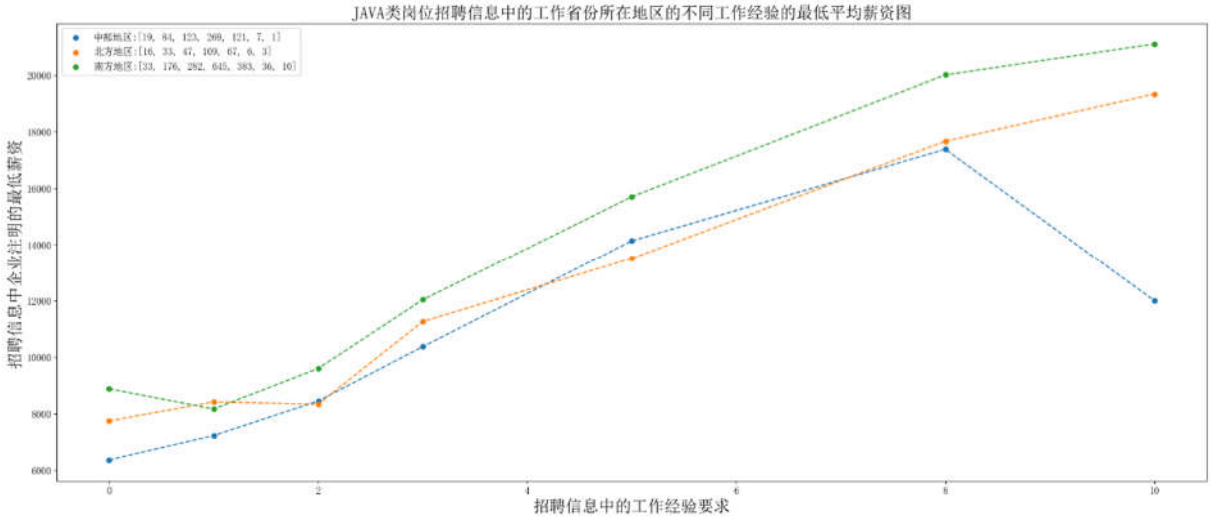


图 3.12 不同工作省份所在地区与工作经验要求下的平均最低薪资

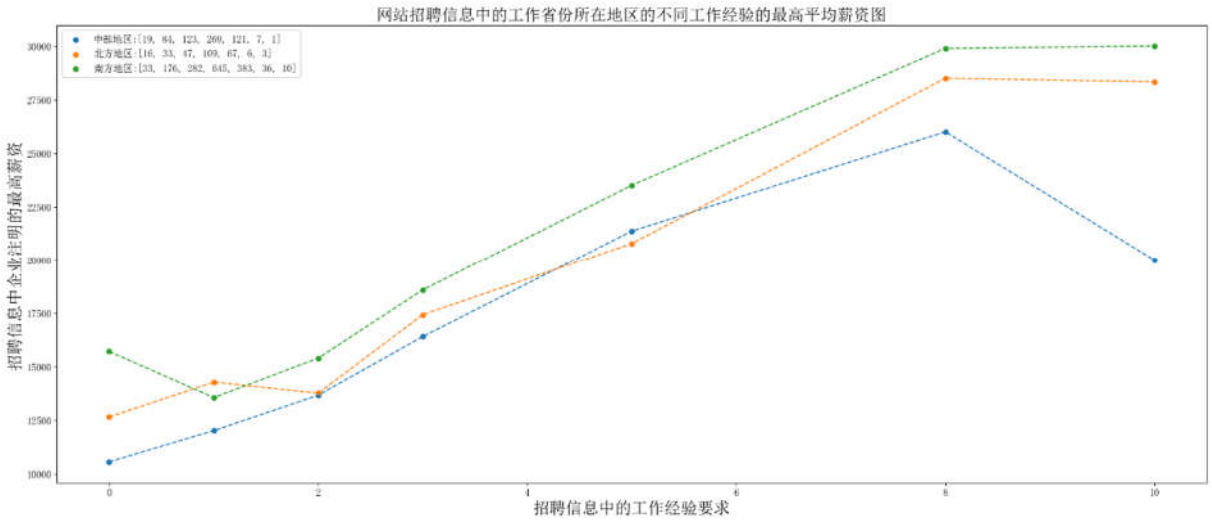


图 3.13 不同工作省份所在地区与工作经验要求下的平均最高薪资

可以看出，对于刚刚毕业的应届毕业生而言，在南方一些经济发展较为迅速的城市，要比中部地区的城市给出的薪资高出许多。

3.3 学历模块

在学历模块中，我们主要分析了不同职位类别、城市、公司行业和公司类型下的学历分布图。

3.3.1 不同职位类别下的学历分布

在不同的职位类型中根据学历字段进行分组之后即可获得该职位类型相应的层叠条形图如图 3.14 所示。

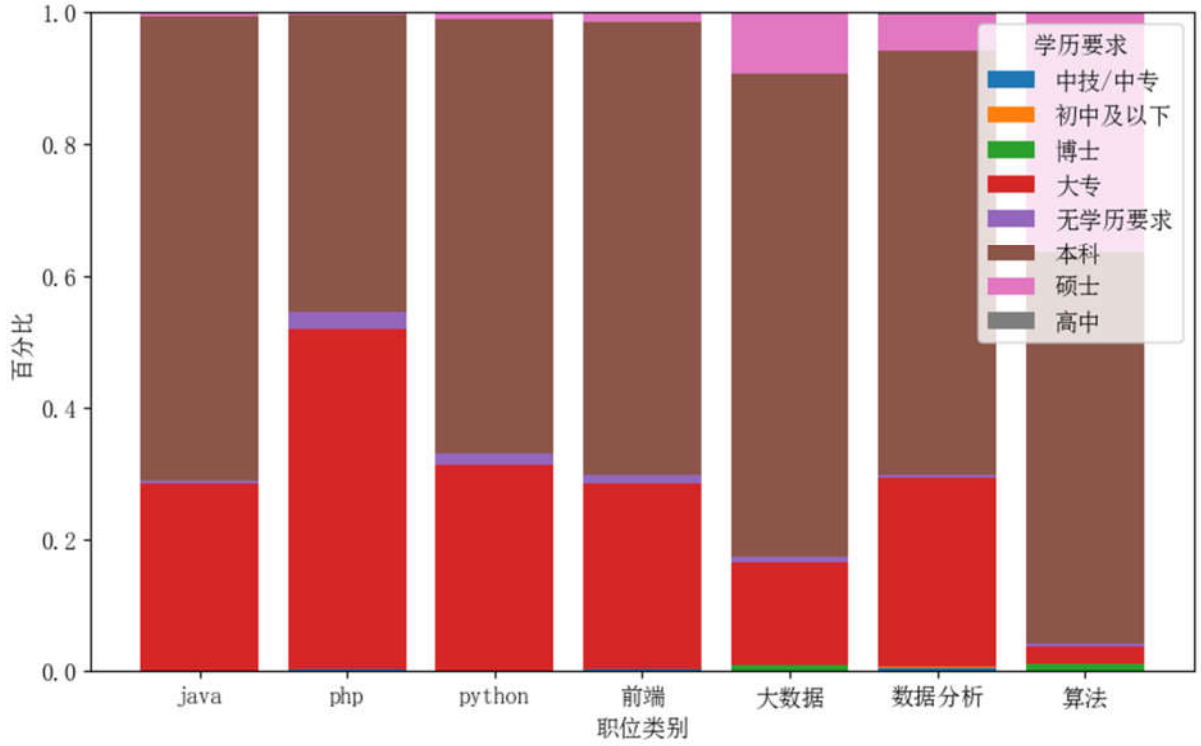


图 3.14 不同职位类别下的最低学历要求分布图

根据图像呈现的数据，算法相关的职位似乎更倾向于要求较高的学历，而其他职位则相对较少要求高学历。

3.3.2 不同省份的学历分布

在不同的职位省份中根据学历要求进行分组之后即可获得该不同省份的相应的层叠条形图如图 3.15 所示。

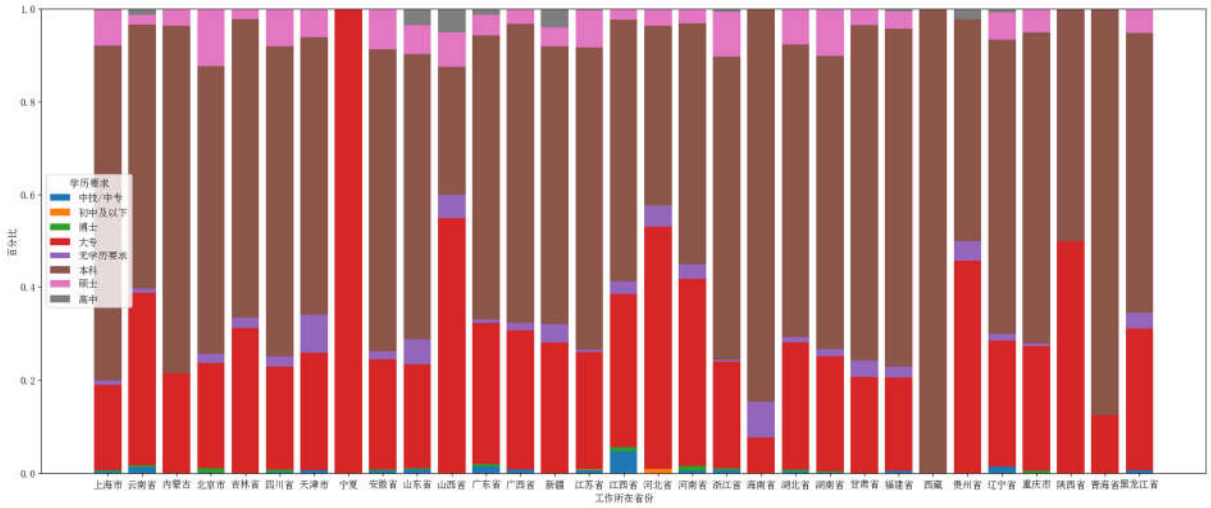


图 3.15 不同工作所在省份的最低学历要求分布图

3.3.3 不同公司行业下的学历分布

在不同的公司行业中筛选其中出现次数最多的公司行业，再根据学历要求进行分组之后即可得到不同公司行业下的学历分布图如图 3.16 所示。

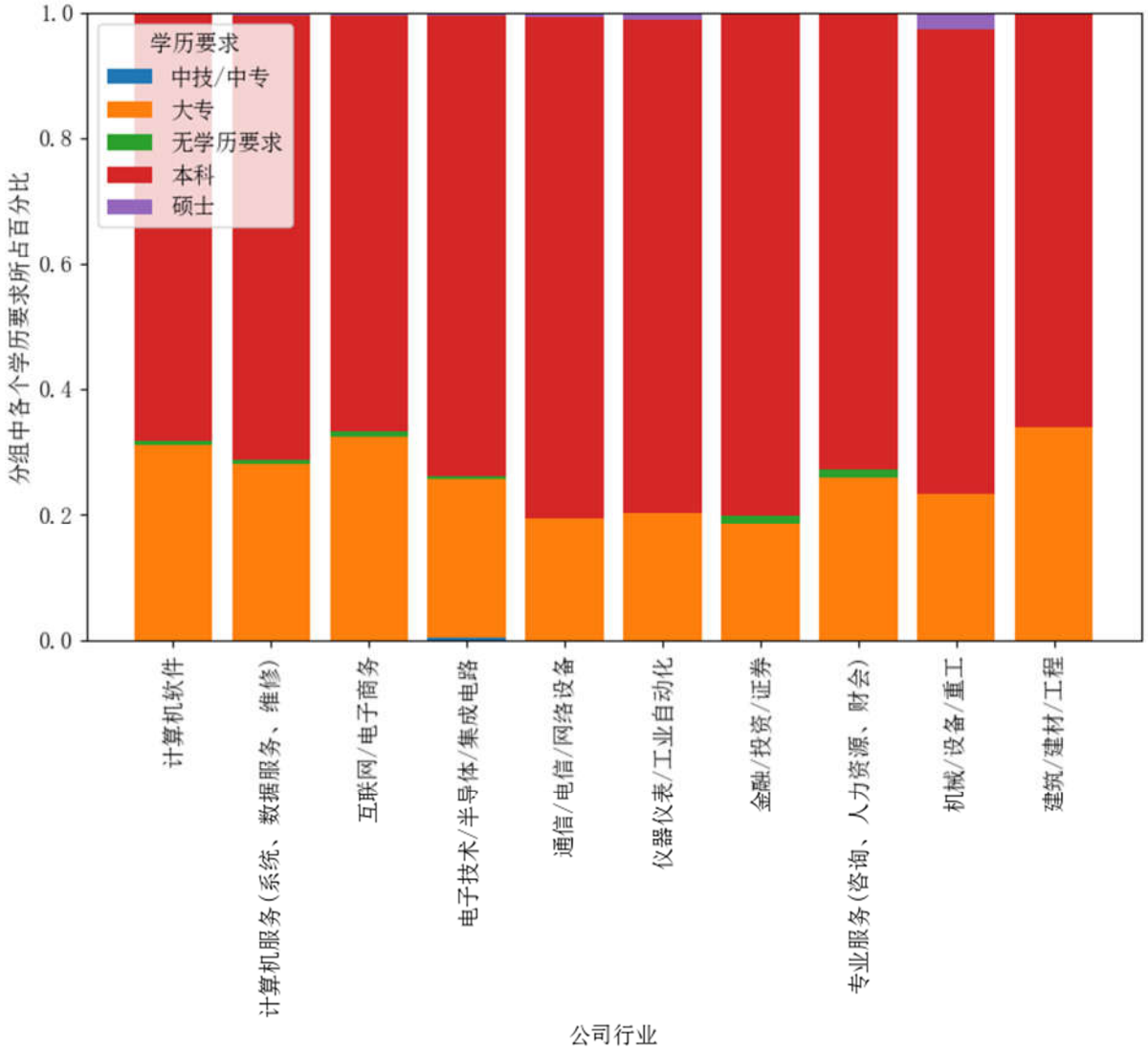


图 3.16 不同公司行业下的最低学历要求分布图

3.3.4 不同公司类型下的学历分布

根据公司类型与学历分组之后即可得到不同公司类型下相应的层叠条形图如图 3.17 所示。

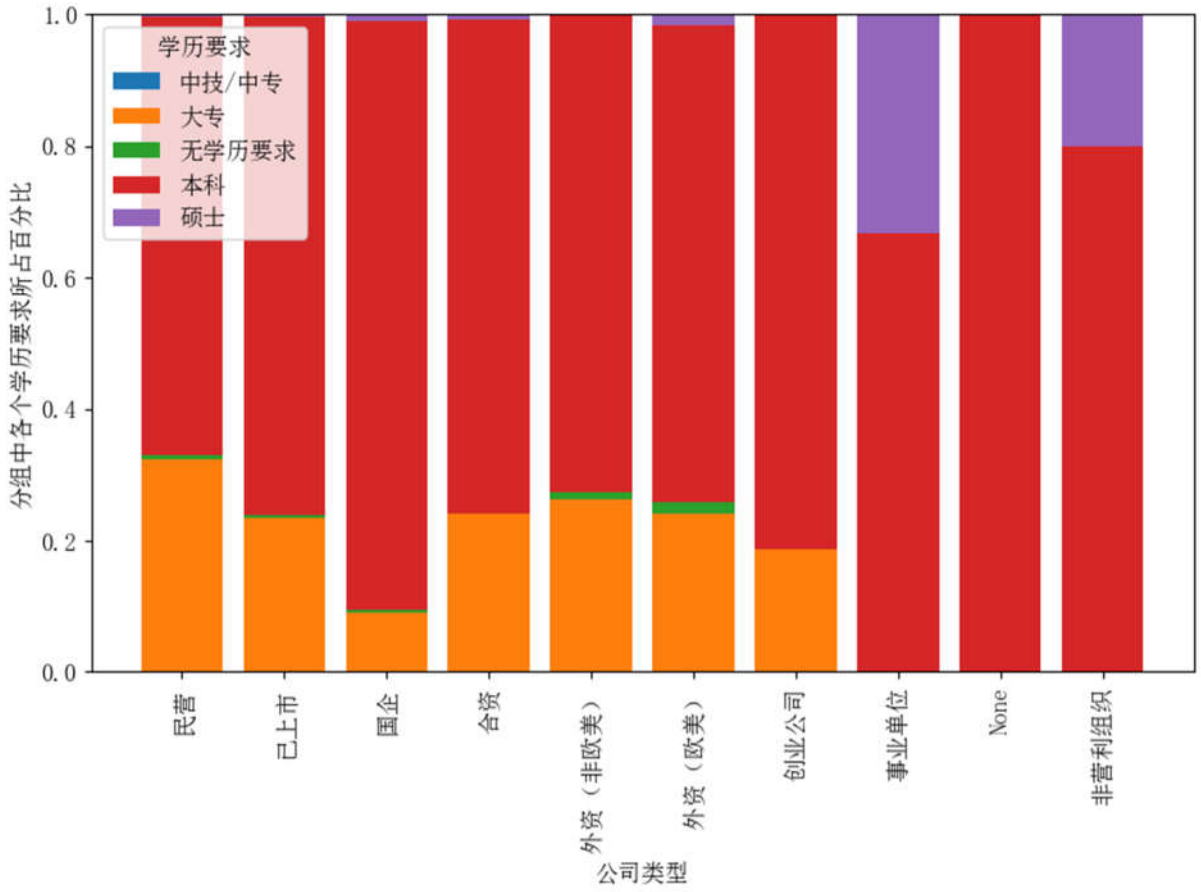


图 3.17 不同公司类型下的最低学历要求分布图

3.4 职位类别模块

在职位类别模块中，我们主要分析了不同城市、工作经验要求和公司类型下的学历分布图。

3.4.1 不同城市下职位类别在统计数据中的分布

根据职位类别、工作所在城市分组后并对数据进行统计分析可以得到如图 3.18 所示的层叠条形图。

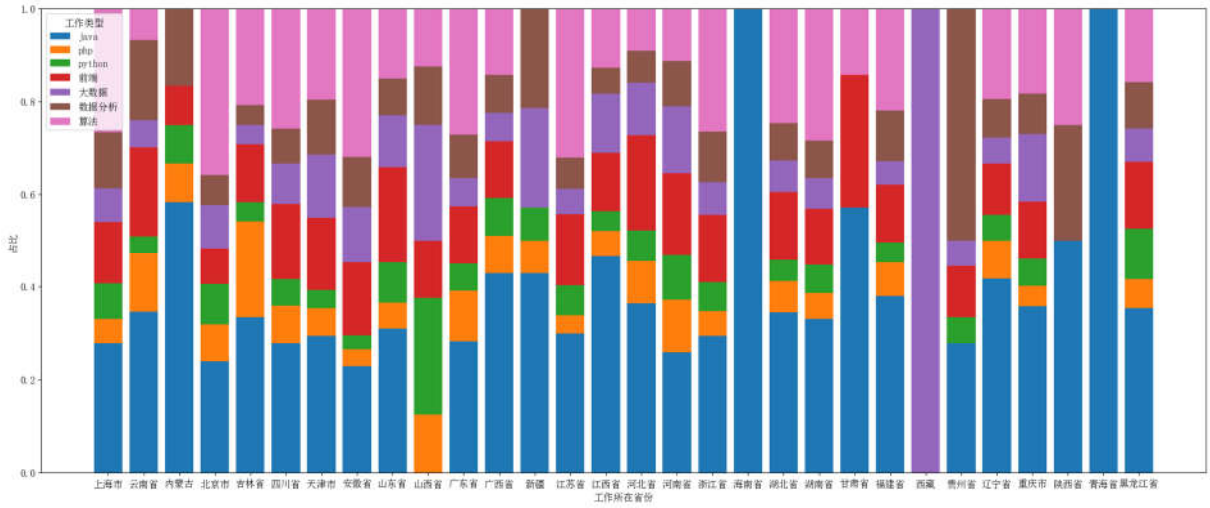


图 3.18 不同城市下职位类别在统计数据中的分布

可以看出其中 JAVA 类职位在大多数省份中具有较高的占比，而其他的如 PHP、数据分析类岗位相对较少。

3.4.2 不同职位类别的工作经验要求分布

根据不同职位类别分组并按照工作经验要求字段进行统计分析之后可以得到如图 3.19 所示的层叠条形图。

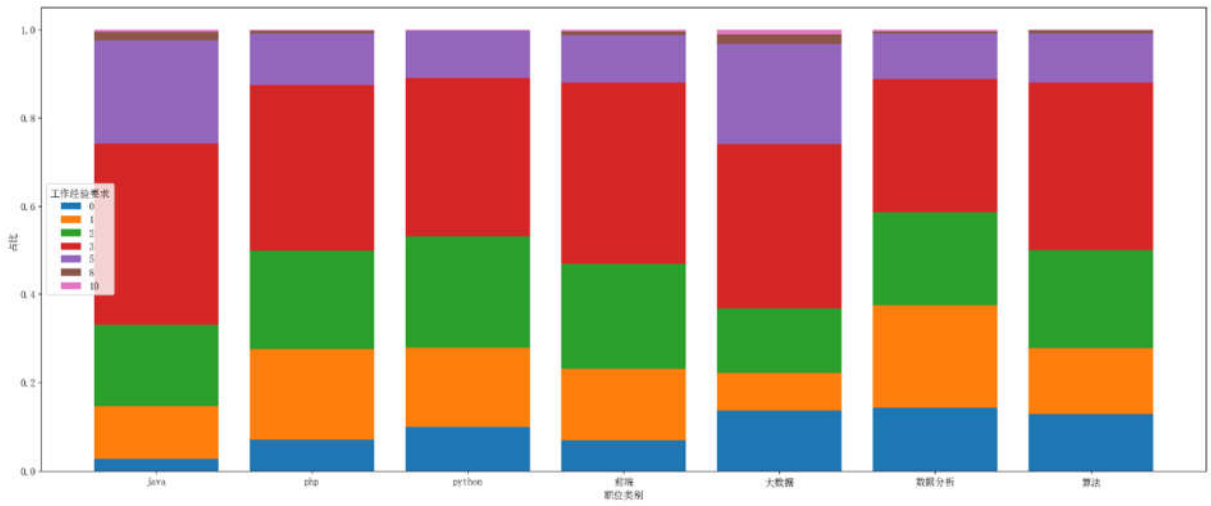


图 3.19 不同职位类别下的工作经验要求分布

可以看到对于大多数公司来说，最低工作经验要求在三年的职位最多，可见此类求职者在择业方面相对来说更容易受到企业的青睐。

3.4.3 不同职位类别下的公司类型分布

根据职位类别进行分组之后对其中的公司类型进行统计分析可以得到如图 3.20 所示的层叠条形图。

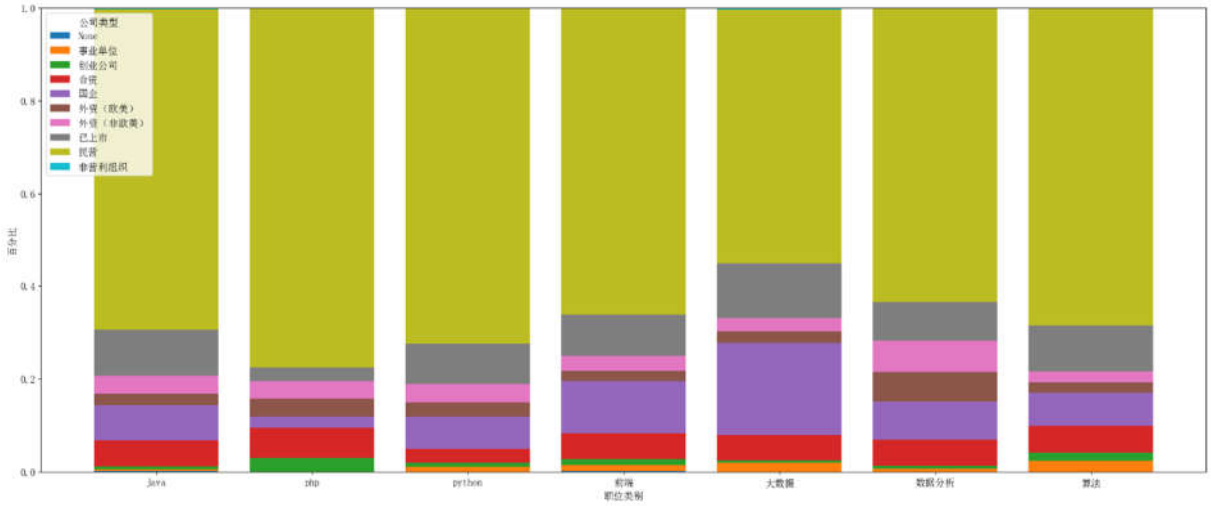


图 3.20 不同职位类别下的公司类型分布

可以看到国内民营类企业的招聘职位数量占有极其高的比重，同时对于大数据类岗位来说，国有企业招聘的岗位数量相对其他岗位也有着不小的占比。

3.5 城市模块

本模块主要分析了平均岗位薪资、工作数量在不同工作所在省份的分布情况。

3.5.1 不同城市的岗位平均薪资地图

对工作所在城市字段进行相关处理之后可以得到工作所在省份字段，而后对工作所在省份字段进行数据分组，并对各分组中的最低、最高薪资进行求平均值的操作之后可以得到如图 3.21、3.22 所示的分布地图。

可以从分布地图中看出其中一些经济发展相对较好的城市给出的职位薪资要比其他一些城市高出许多。尤其是北京、上海、广东、浙江几个省份给出应届生的平均最低薪资已经达到了 10k 以上，同时平均最高薪资也达到了很高的水平。

图 3.21 不同省份最低平均薪资分布图

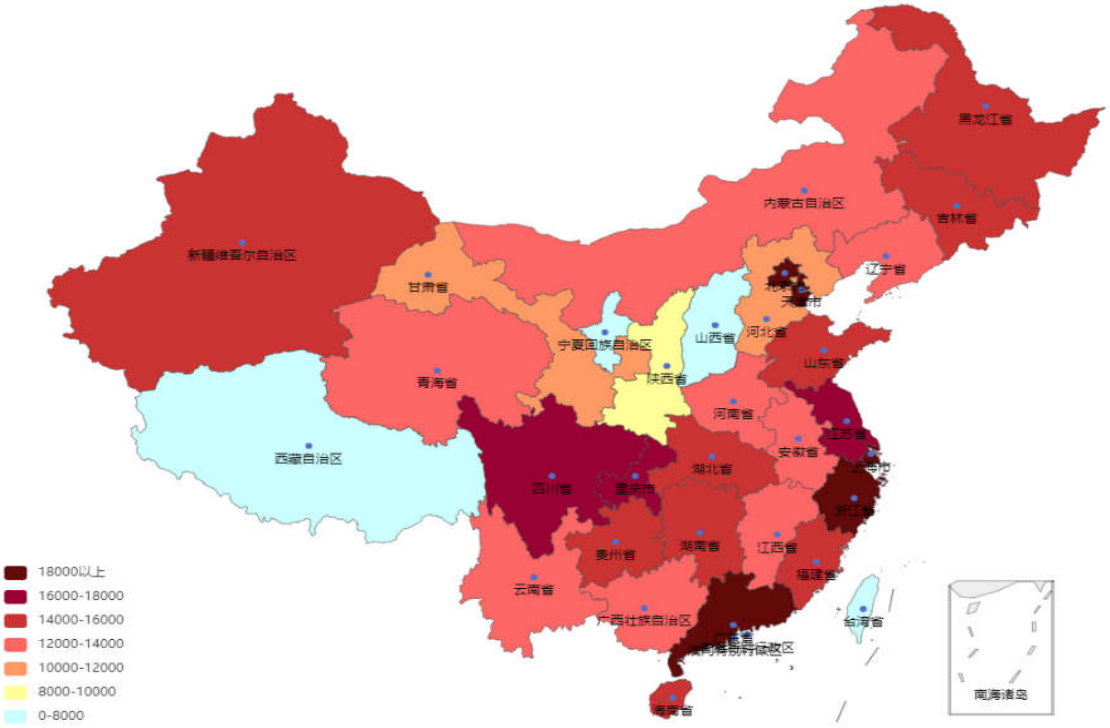


图 3.22 不同省份最高平均薪资分布图

3.5.2 不同城市的岗位数量分布地图

通过对职位信息中的工作所在省份进行统计分析，可以绘制出岗位分布地图，如图3.23所示。

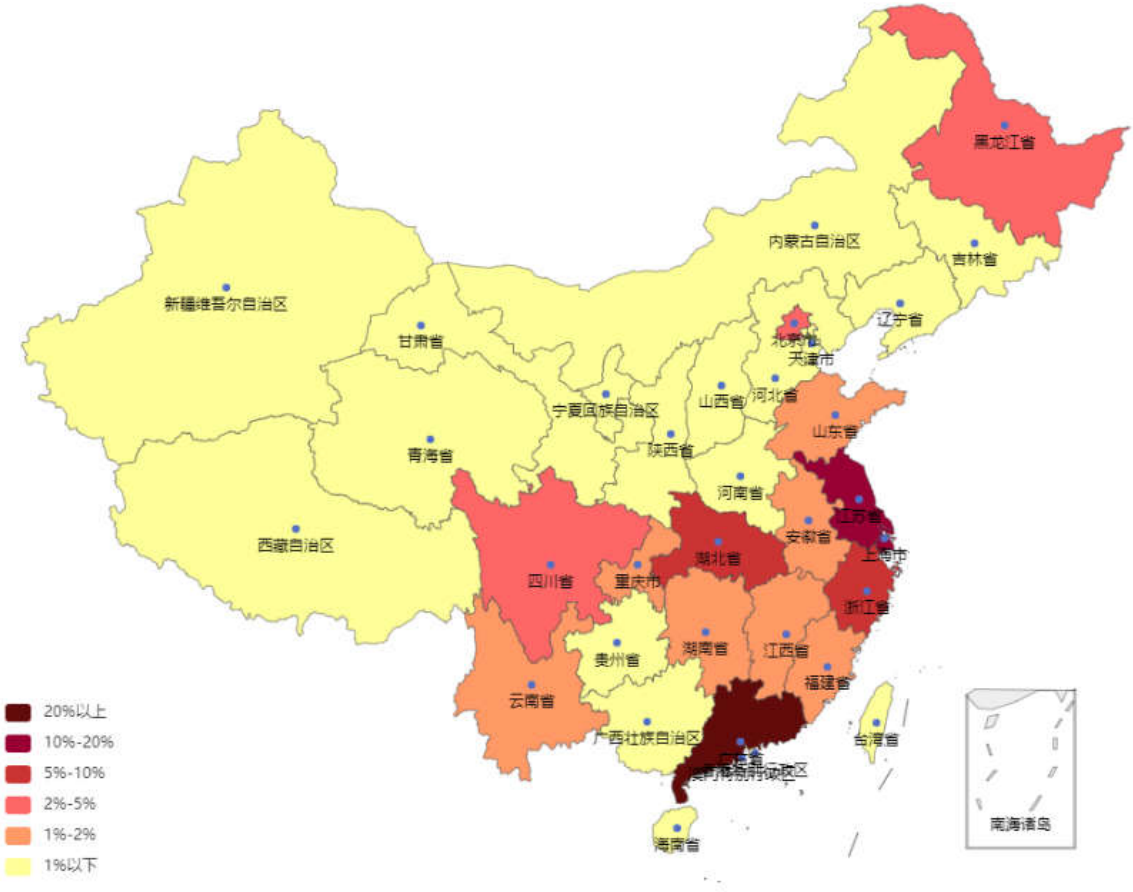


图 3.23 不同省份职位数量百分比分布图

由图像可以看出当前广东、江苏等地区的职位数量占比已经远远超过了河南、河北等众多的经济发展相对落后的地区。因此对于能力相当的应届毕业生来说，我们更容易在此类地区找到一份薪资水平相对较高的工作。

4 总结与展望

4.1 总结

本文主要针对某特定行业的招聘信息进行了数据采集、特征词汇提取和数据可视化分析等工作。通过采用文本数据采集技术，获取特定网站的招聘信息数据并存储至数据库中，使用 Open AI 公司提供的基于 gpt-3.5-turbo 模型接口，对数据库中的招聘要求字段进行特征词汇的提取，最后对提取后的数据进行了预处理和可视化分析。通过这些工作，我们得到了一份针对该行业的招聘要求特征词汇的数据分析结果，也可以从中得到许多关于个人与高校发展的启示。同时在本研究中运用的特征选择方案和进行数据分析的方式中，也依然存在着诸多问题。

4.1.1 启示

根据本研究的分析结果，我们可以得到以下几个启示：

在个人能力提升方面：对于那些对数学能力要求相对不高的职位，我们应更加注重学习与自己想从事的工作相关的专业技能知识，而不仅仅局限于一些可能已经过时的课本知识。此外，由于各种语言技术在不断的发展，以往流行的开发技术也有可能会被逐渐淘汰，因此我们应该积极了解最新的相关技术，而不是仍然处于被动学习的状态之中。

在个人职业生涯方面：我们应注重自身能够为企业和社会进步带来的价值，因为良好的企业发展和社会进步也将相对带来更好的职业发展机遇。同时，我们也应注重提高数学能力和算法能力，这也更能体现我们作为研发工程师的价值和水平。

在高校发展方面：高校应积极接受并推动新技术的教育，同时在新技术授课过程中从旧技术知识中汲取精华、去其糟粕，以更好地推动社会经济发展。此外，对于应用类课程，应采用更为科学以及更能适应社会经济需求的教育模式。

4.1.2 不足

本研究通过采用CPT-3.5模型获得词汇相关的关键词信息已经达到了相对较好的提取效果，但是在如今自然语言处理方面的一些技术难题同样也依然存在于本研究之中。

首先，对于完全同义的词汇整合方面，本研究通过采用一些对数据进行标准化处理的方式进行减少了部分由于词汇之间的空格以及大小写导致的数据不一致问题。但是在

中英文完全同义的词汇中多种不同表达方式上，本研究仍然有可能会丢失一些相关的词汇数据，进而有可能会导致部分关键词统计数量存在部分缺失的问题出现。其次，在本研究中所采用的分析方法也存在着诸多问题。例如，在众多的技术类词汇相关性分析中，本研究并没有通过更为准确的相关性数学分析方法进行词汇选择与相关性分析。此外，在各个字段的数据分析过程中，可能存在更为科学和精确的相关性分析方法，但由于研究人员技术水平有限，目前无法进一步探究和应用这些方法。

4.2 展望

对于本研究中存在的缺陷，我们可以通过以下几个方面来加以改进和优化。首先，可以在数据预处理环节中采用更加完备的同义词汇表和翻译工具，以增强对中英文完全同义的词汇的处理能力，从而尽可能减少数据不一致问题的出现。其次，在关键词的选择和相关性分析环节中，可以探究和应用更为准确的相关性数学分析方法，例如基于统计学的卡方检验等方法，从而提高分析结果的准确性和可靠性。此外，可以考虑引入更为先进和高效的文本数据挖掘技术和算法，以进一步提高招聘信息数据的提取和分析效果。

在未来的工作中，也可以进一步拓展和完善该数据集，增加更多的数据来源和多样化的特征词汇提取方法，提高数据的覆盖面和代表性。同时，可以将这些特征词汇应用于招聘信息的筛选和推荐，为求职者和招聘企业提供更加个性化和精准的服务。另外，该方法也可以应用于其他类别的文本数据分析和挖掘中，具备广泛的应用前景，有望在多个领域发挥重要作用。

参考文献

[1] 宋齐明.校园与工作场所：关于本科生可就业能力的研究[C].华东师范大学,2018.

[2] 黄承慧;印鉴;侯昉.一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J].计算机学报,2011,34(05):856-864.

[3] 张俊峰;魏瑞斌.国内招聘类网站的数据类岗位人才需求特征挖掘[J].情报杂志,2018,(06):176-182.

[4] 陈飞.应用型本科教育课程调整与改革研究[C].华东师范大学,2014.

[5] 钟晓旭.基于 Web 招聘信息的文本挖掘系统研究[C].合肥工业大学,2010.

致谢

十八年求学生涯即将告一段落，纵然并未曾获得太大的成功，但终于也算塑造了有一定自我认同感的三观和较为清晰的人生规划。感谢我的导师，在论文的编写过程中给予的帮助；感谢我的同学在论文的完成过程中给予的一些建议；感谢我的父母，含辛茹苦培养出了家里的三个大学生；感谢这一路走来曾与我同行过一程的所有人；感谢这一路走来对得起的、对不起的人；感谢这一路走来出现的所有善意与恶意；感谢这一路走来所有的开心与难过；感谢这一路走来所有可以让我越来越了解世界的一切事物。

许多往事仍历历在目，但人终究还是要往前走。

二十一又半载光阴，真正学会开始思考一些事情也不过是十年之前而已。但十年转瞬即逝，所了解的事物也仍只是冰山一角，曾未想通的事情可能也仍未想通。但还好事物发展历程都是螺旋式上升，不论经历是否圆满，也终将剧终。

词不达意，感激之情仅能以此来聊表心意。

谨以此篇论文，纪念我逝去的青春，迎接我也许将会极其璀璨绚烂的人生。

G·格子达论文检测报告【简版】

报告编号:55958FD908BC43A1A07DC4FF475B964A

送检文档:互联网研发岗位信息特征挖掘与分析

作 者:郭名胜　　送检单位:河南大学　　送检时间:2023-05-08 21:10:41

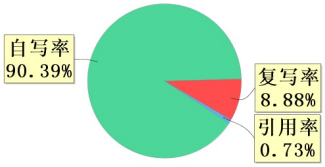
比对索引库

1989-01-01至2023-05-08

学术期刊库	报纸资源库	本科论文共享库	格子达公识库
学位论文库	互联网资源库	专利库	机构自建库
会议论文库	格子达多元库		

检测结果

总相似比:9.61%(总相似比=复写率+引用率)



查重检测指标:自写率90.39%复写率8.88%引用率0.73%（含自引率0.0%）

其他类型检测结果:去除引用后总相似比:8.88% 同校同届总相似比:1.56%

相似片段:复写片段28 同届片段5 引用片段2

指标名称	学校要求	指标检测结果	系统判定
总相似比	不超过20%	9.61%	符合
复写率	不超过20%	8.88%	符合
同届比	不超过20%	1.56%	符合
论文总字数	不少于800字/单词	17722字符	符合

其他检测结果：

指标名称	识别数量	系统判定
代码块检测	0	--

复写率索引来源

学术期刊:（3.96%）　　学位论文:（3.44%）　　本科论文共享库:（1.48%）

引用率片段来源

学术规范引用（0.73%）　自我引用（0.0%）　其他引用（0.0%）

免责声明

1、 本报告为G · 格子达系统检测后自动生成，鉴于论文检测技术及论文检测样本库的局限性，G · 格子达不保证检测报告的绝对准确，您所选择的检测资源范围内的检验结果及相关结论仅供参考，不得作为其他任何依据；
2、 G · 格子达论文检测服务中使用的论文样本，除特别声明者外，其著作权归各自权利人享有。根据《中华人民共和国著作权法》等相关法律法规，G · 格子达网站仅为学习研究、介绍、科研等目的引用论文片段。除非经原作者许可，请勿超出合理使用范围使用本网站提供的检测报告及其他内容。

联系我们

防伪二维码　关注微信公众号

官方网站:co.gocheck.cn
客服热线:400-699-3389
客服QQ:800113999