分类号_____　　密级_____
U D C _____　　编号　10741

硕 士 学 位 论 文

（专业学位）

论文题目　基于文本挖掘的数据类岗位人才需求分析

研　究　生　姓　名：　王宇辰

指导教师姓名、职称：　黄恒君　教授

学　科、专　业　名　称：　统计学　应用统计硕士

研　　究　　方　　向：　大数据分析

提　　交　　日　　期：　2022.05.30

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：王宇辰 签字日期：2022年5月30日

导师签名：黄恒君 签字日期：2022年5月30日

导师(校外)签名：杜荣 签字日期：2022年5月30日

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：王宇辰 签字日期：2022年5月30日

导师签名：黄恒君 签字日期：2022年5月30日

导师(校外)签名：杜荣 签字日期：2022年5月30日

Demand Analysis of Data Jobs Based on Text Mining

Candidate : Wang Yuchen

Supervisor: Huang Hengjun

摘 要

随着大数据时代的到来，高等教育的发展与普及，高校数据类人才数量不断增加；在大数据与人工智能技术的发展过程中，数据类岗位的市场需求也逐渐增大。但是，仍然会出现毕业生难以找到较为满意的工作、企业难以招聘到理想人才的情况。本文旨在探究就业市场上数据类人才的招聘情况，对数据类岗位的侧重点进行对比，挖掘出企业对数据类人才的要求，为希望从事数据类岗位的求职者提供一些参考。

本文在前程无忧招聘平台上以“数据分析”、“数据挖掘”、“数据开发”、“数据运营”等四种关键词进行检索，并使用网络爬虫技术获取岗位信息进而对数据类岗位展开研究。首先，采用描述性统计与可视化方法，对这四类岗位的城市需求分布、工作经验要求、学历要求进行比较，探究不同地区数据类岗位的需求差异，并在此基础上分析岗位薪资与工作经验要求、薪资与学历要求的走势情况；其次，对四类岗位的招聘文本信息进行文本数据预处理与文本向量化，构建 LDA 主题模型，根据模型评价指标确定最优的主题数，并输出各类主题词；最后，利用 Word2Vec 模型对 LDA 模型提取的主题词进行拓展，提取语义近似的词项，并结合二者筛选出关键词绘制词云，从岗位职责、职位场景、技能需求等方面对四类岗位进行比较。

研究结果表明：数据类岗位需求较大的城市包括华东地区的上海、杭州、南京等，华南地区的广州、深圳、东莞等，华南地区对数据挖掘与数据开发岗位的需求更大；工作经验要求方面，绝大部分岗位都要求就业人员具备 1 年以上相关工作经验；学历要求方面，本科成为大部分数据类岗位的起步学历，寻求数据挖掘与数据开发工作时，本科学历不具备竞争优势。LDA 主题模型与 Word2Vec 模型的主题词提取结果显示，数据分析岗位强调数据敏感性与搭建指标体系的能力；数据挖掘岗位注重数据分析与挖掘算法的设计与优化，需要具备大数据平台的使用能力；数据开发岗位侧重大数据平台、数据计算架构的研发与运维；数据运营旨在通过数据与业务相结合，指导业务增长。数据分析和数据运营偏向业务，数据挖掘和数据开发更偏重技术。

根据上述结果，本文旨在帮助致力于向数据类岗位发展的求职者，有助于求职者、应届生和在校生根据自身情况选择合适自己的岗位类型、就业地区，根据市场需求有针对性的补强自身实力，提高就业率。

【关键词】数据类岗位； 文本挖掘； LDA 主题模型； 岗位需求对比

Abstract

With the advent of the era of big data and the development and popularization of higher education, the number of data talents in colleges and universities is increasing; in the process of development of big data and artificial intelligence technology, the market demand for data positions is also gradually increasing. However, there will still be situations in which it is difficult for graduates to find more satisfactory jobs and for companies to recruit ideal talents. This paper aims to explore the recruitment situation of data talents in the job market, compare the focus of data positions, and discover the requirements of enterprises for data talents, so as to provide some references for job seekers who want to engage in data positions.

In this paper, four keywords, such as "data analysis" "data mining" "data development" and "data operation", are used to search on the 51job.com recruitment platform, and the web crawler technology is used to obtain job information and then expand the data-based jobs. Research. First of all, using descriptive statistics and visualization methods to compare the urban demand distribution, work experience requirements, and education requirements of these four types of jobs, explore the differences in demand for data jobs in different regions, and analyze job salaries and work experience on this basis. The trend of requirements, salary and educational requirements; secondly, text data preprocessing and text vectorization are carried out on the recruitment text information of the four types of positions, the LDA topic model is constructed,

the optimal number of topics is determined according to the model evaluation index, and various types of topics are output. Subject words; finally, the Word2Vec model is used to expand the subject words extracted by the LDA model, extract terms with similar semantics, and combine the two to filter out the keywords to draw a word cloud. class positions for comparison.

The research results show that cities with large demand for data jobs include Shanghai, Hangzhou, Nanjing, etc. in eastern China, Guangzhou, Shenzhen, Dongguan, etc. in southern China, and there is a greater demand for data mining and data development jobs in southern China; work experience In terms of requirements, most positions require employees to have more than 1 year of relevant work experience; in terms of academic requirements, a bachelor's degree has become the starting degree for most data-related jobs. When seeking data mining and data development jobs, a bachelor's degree does not have a competitive advantage. The results of the subject heading extraction of the LDA topic model and the Word2Vec model show that the data analysis position emphasizes data sensitivity and the establishment of an index system; the data mining position focuses on the design and optimization of data analysis and mining algorithms, and requires the ability to use big data platforms; data development The research and development and operation and maintenance of major data platforms and data computing architectures on the job side; data operations aim to guide business growth through the combination of data and business. Data analysis and data operations are more business-oriented, while data mining and data development are more technology-oriented.

According to the above results, this paper aims to help job seekers who are committed to the development of data-related positions, help job seekers, fresh graduates and school students to choose suitable job types and employment areas according to their own conditions, and to provide targeted solutions according to market demand. Strengthen its own strength and increase the employment rate.

Keywords： Data Jobs; Text Mining; LDA Topic Model; Job Requirements Comparison

目 录

1 引言

1.1 研究背景

伴随着大数据时代的到来，数据渗透到生活中的各个方面，日常生活中涌入大量的数据，不再局限于传统的数值形式，数据还可以表现为音频、文本、视频等等非结构化的形式。随着互联网的迅速发展，网络中时刻产生着海量的文本信息，这些海量文本信息看似杂乱无章，实际上隐含着数据对象的潜在规律。如何从海量的、存在噪声的文本信息数据中挖掘出有价值的信息与结论，为决策者提供判断依据是文本挖掘的重要领域之一，也是广大研究者的研究方向之一。

随着互联网的迅速发展，网络招聘平台近些年如雨后春笋般出现，企业招聘人才、工作者求职的形式也发生巨大的变化。网络招聘平台费用低且信息更新及时迅速，逐渐已经成为招聘者发布岗位信息、应聘者获取岗位信息的主要途径。通过网络招聘平台，企业可以在移动端随时发布工作的招聘需求，求职者们可以浏览工作岗位的相关招聘信息，也可以将个人状况公开在平台上供企业筛选，这种形式在很大程度上节省了双方的时间，提高了求职过程的效率。与此同时，国内对数据类人才的需求呈现出爆发式的增长，社会需要数据类人才深入各行业发挥数据的价值。然而，国家信息中心发布的《中国大数据发展报告（2017）》曾指出，由于大数据技术近些年在我国由出现、兴起到广泛应用的历时较短，数据类人才培养模式不够成熟，培养速度较为缓慢，常常形成就业市场上数据类岗位从业者技能与工作岗位不匹配的情况。此外，由于“大数据”等概念在社会受到热捧，部分求职者对数据相关工作的盲从，使得企业数据类岗位缺口较大、人才掌握的技能与岗位需求不匹配的现象愈发普遍，对数据类人才顺利求职、企业招贤纳士和国家发展大数据产业造成不利影响。

对于高校应届毕业生而言，对各职业类型所属的行业情况、城市需求分布及薪资的影响因素等综合性信息缺乏较为全面的了解，对于个人能力是否满足招聘方的技能要求也缺乏理性公正的认识，从而造成“毕业生与企业在求职问题上的不同步”的现象产生。通过细致分析可以发现，一方面是由于不同区域、不同行业对求职者的技能与学历要求程度不尽相同，并且根据国家政策的不断调整，企业对人才的要求随着时代的发展发生变化，但是高校环境与求职者对人才市场需求等重要的市场信息缺乏深入了解。

因此，针对数据类岗位大量且多样化的岗位招聘信息，探索挖掘出数据类岗位需求，

帮助企业定位人才，帮助人才积累企业所需技能，具备现实意义。

1.2 研究意义

求职过程实际上是市场需求和供给匹配的结果，招聘方和应聘方都有自己的需求，双方在网络招聘信息发布平台完成市场交易。然而，这种信息匹配的过程受到诸多因素影响，使得招聘企业在网络招聘网站上发布了招聘岗位的相关信息与用人需求，但是求职者不能够较为充分、较为全面地在海量繁杂的数据中获取具有重要价值的信息。

首先，招聘网站发布的网络招聘信息往往是文本形式，具备海量和非结构化的特点，传统数值型数据的分析方法略显不足。帮助人才对数据类岗位所在地区和区域等因素有基础的认识，不仅能够明确不同类别的数据类岗位对人才需求的侧重点，而且能够发现企业在招聘时对数据类人才的要求及规律，将分析结果与实际情况相结合，能够帮助求职者对数据类岗位所需技能及要求有更为深刻的认识，从而查漏补缺，根据自身知识储备向市场需求的方向靠拢。

其次，对于培育人才的高校而言，对招聘信息进行深层挖掘，及时了解数据类岗位市场需求动态，把握社会对数据类人才的需求变化，可以帮助提升高校毕业生的技能水平与专业素养，为求职者们提供贴合市场需求的就业指导。此外，高校能够紧跟时代步伐设置相应的数据类人才培养方案和课程，搭建相应的数据类人才培养模式和体系，承担起为社会培育更符合时代发展趋势的人才的重任。

最后，充分利用网络招聘信息的分析和挖掘的结果，可以对职位特点进行刻画，能够描述社会各行业对人才需求的情况，可以更加明确各类企业对求职者的任职要求，为企业提供招聘现状及企业对人才的需求情况。

显然，充分挖掘网络招聘信息对于求职者与企业进行联动分析，解决社会就业难的问题具有重要意义。

1.3 国内外研究现状

1.3.1 文本挖掘的研究现状

文本数据如今成为我们工作中最常见的数据类型之一，与我们的生活紧密相关，我

们讲话、书写、网络聊天的载体都是文本，其本质可以理解为一种以文本形式存在的数据。随着数据存储方法的改良、信息传输速度的提升，文本大数据挖掘方法逐渐成为数据挖掘领域的研究重点之一，代表方向为自然语言处理。

目前，网络上文本形式的数据数量繁多且内容复杂，文本挖掘技术旨在从海量文本内容中挖掘出有价值的潜在信息，最终将这些信息应用于个人、企业以解决实际问题。文本挖掘最早于 1995 年受到关注，最早由 Feldman[1]（1995）等研究非结构化数据问题时所提出。在国外的研究中，Beil[2]（2002）等提取短文本中的频繁集的关键词得到频繁特征集，发现在相同主题的短文本中存在较多类似的频繁特征，最后通过对频繁特征聚类表示短文本的聚类结果。Ding[3]（2012）等做出文本挖掘算法使得各语言之间的数字文档交流的通用性得到增强的判断，并认为文本挖掘方法将在知识发现领域中得到更深层的发展。Goswami[4]（2013）等使用模糊逻辑算法对文档数据进行聚类，提升了聚类效果；Hao[5]（2018）等使用文档主题模型、知识图谱、地理信息可视化等方法深入对文本挖掘领域的研究成果进行深层挖掘。

我国的文本挖掘研究领域起步较晚但成果明显，在短短数年间取得长足的进步和发展。在国内学者诸多研究领域中，文本挖掘方法与其特定领域相结合并发挥了巨大价值。杨亚楠[6]（2019）等采用多视图协同学习的方法，对具有时序特征的政策文本进行分析，探索政策内容的深层规律；沈健[7]（2018）等在生物领域运用文本挖掘技术，构造生物领域专业语料库，研究该领域中文本数据的相似性度量方法，并提出一种实例抽取与特征选择的方法，改善了多种环境下的文本检索效率；沈艳[8]（2019）等对文本大数据信息提取方法进行综述研究，并对文本大数据分析在经济学和金融学中的应用进行梳理说明；倪志恒[9]（2021）采用集成学习算法对电商评论中的虚假评论文本数据进行识别，对比多种机器学习算法在处理虚假评论识别问题的效果；戴德宝[10]（2019）等将文本挖掘算法应用到股票数据分析中，将文本情感分析、Granger 因果关系检验等方法构建投资者情绪综合指数，显著提高了股价走势预测的精度。

LDA 模型分析是文本挖掘中主题模型的重要方法之一，国内不少学者运用 LDA 模型进行各自学科文献的主题发现。李伦珑[11]（2021）使用 LDA 模型对各朝代的诗词主题进行挖掘，并沿着时代变迁分析诗词主题的变化；杨文清[12]（2021）使用 LDA 主题模型和 VAR 模型研究投资者情绪对股票收益率的影响，发现不同的投资者情绪变动方向对股票收益率存在显著差异；李梦杰[13]（2018）等将 LDA 主题模型和层次聚类算法相结合，对某在线教育平台内的课程信息进行挖掘，挖掘出不同主题课程间的潜在关联

以及用户课程选择时的关注特征；谭春辉[14]（2021）等采用 LDA 主题模型对数据挖掘领域核心期刊论文文本数据进行研究主题抽取，并基于生命周期识别热点主题，探究国内外数据挖掘领域研究热点主题演化的异同点；周云泽[15]（2021）等使用 LDA 模型与共享语义空间方法在多源数据下识别新兴技术主题，对新兴技术主题进行合并；黄琳[16]（2022）等提出一种改进的 LDA 模型，将客户-产品服务画像得来的先验知识整合进 LDA 模型中，进而引导模型学习出与产品服务相关的特定主题。

1.3.2 网络招聘信息的研究现状

网络招聘信息的日渐丰富，学者们纷纷对网络招聘数据进行研究。Shenoy[17]（2018）等梳理企业网络招聘的流程，探究网络招聘模式对企业、求职者和机构的所产生的影响。Litecky [18](2010)等使用文本分析提取工作技能术语，结合系统聚类和 k-means 聚类算法把 IT 岗位划分为 20 类，并针对各岗位类型提出技能需求组合。Turrell[19]（2019）等搜集数千万条的网络招聘广告并开发一种由招聘数据映射到标准职业分类编码中的算法，该算法可以为英国提供完善的劳动力需求图表。Sodhi[20]（2009）构建运筹类专业关键词词典，分析运筹类专业岗位招聘信息，着重探究各行业岗位对运筹类专业技能需求的差异；俞琰[21]（2019）等对比分析各专业领域技能衡量指标，基于传统的术语抽取算法对网络招聘文本中技能的相关信息完成抽取；许艳丽[22]（2019）对多种行业的网络招聘信息进行文本挖掘，发现人工智能行业对求职者综合能力要求较高，并认为高职院校培养人才时应结合行业特点；杨文泽[23]（2019）重点分析网络招聘岗位的薪资水平，探究薪资水平如何影响劳动力流动情况，并对改善劳动力资源分配提出建议；詹川[24]（2017）等对电商行业的招聘岗位进行研究，构建电商体系、技能指标以及专业技能词典，对电商各岗位的技能需求进行挖掘。

关于数据类岗位人才的研究中，张俊峰[25]（2017）使用互联网公司 BAT 和通讯设备企业华为共 4 家企业的招聘信息，首先构建专业技能分词词典，据此提取招聘信息中的高频词，并使用 Word2Vec 算法扩展语义近义词，总结出企业对人才的要求主要围绕专业背景、专业知识、计算机掌握程度等维度；刘睿伦[26]（2017）对前程无忧、看准网等多个招聘平台上的数据产品经理、数据分析师等 7 个数据类相关岗位的招聘信息进行挖掘，使用 jieba 包自定义词典，根据 k-means 算法确定文本向量的最佳维度和聚类个数，最终围绕能力要求、学历要求、工作经验三个维度对这 7 类岗位的异同点展开讨论。谭

云鹤[27]（2019）选择全国范围内的 8 个数据类岗位，尝试分析该 8 类岗位间的相似性，最终通过聚类算法将这 8 类岗位划分为数据分析类、数据开发类、数据算法类，并对这三类岗位的技能需求、岗位地区分布、薪资状况等维度进行分析；刘畅[28]（2019）选定职位名称、公司名称、工作地点、薪资待遇、所属行业等特征对数据类岗位信息开展分析，并从教育背景等 4 个不同的维度对数据类岗位对应的主题词归纳总结，并针对每种岗位内的需求构造共线，探讨不同岗位之间的任职需求差异，深层分析岗位内部技能间的联系；朱爱璐[29]（2020）以数据分析岗位所处地区为切入点，分析 7 个地区 297 个行业相关的数据分析岗位，在每个地区的数据中构建 LDA 主题模型，提取不同地区的岗位要求的关键词；黄崑[30]（2016）以数据分析、数据管理、数据挖掘等为关键词进行检索，并围绕这 3 类岗位的基本信息、主要职责、任职要求进行分析；杨静[31]（2019）分析了山东省以及大数据相关岗位的岗位要求，通过 TF-IDF 算法与 LDA 模型构建职位画像，对大数据类岗位相关的人才要求进行探讨；郑思雨[32]（2020）将工作城市所属地区分为东部、西部、中部和东北部地区，分别对 4 个地区内的数据进行关联规则分析和 k-means 聚类，挖掘岗位要求字段间隐藏的关联信息。

不同于网络招聘信息内容的深层挖掘，有学者着重研究招聘信息分析的算法。郭欢欢[33]（2020）将大数据方法引入精准招聘过程中，实现雇主画像和技能词典的构建，并基于此构建精准招聘系统；吴汉龙[34]（2021）等重点研究爬虫与信息匹配算法，提出一种网络招聘信息特征提取方法，将招聘信息以更快速度、更高效率从 Web 端获取并保存，以 Web 开发工程师为例进行测试；李寿清[35]（2020）使用机器学习算法对数据分析师、机器学习工程师、数据挖掘工程师、深度学习工程师这 4 类岗位的薪资影响因素进行挖掘，研究表明集成学习算法执行效率高，预测效果更优；俞琰[36]（2019）基于网络招聘的文本数据，提出“岗位-课程-知识点”的三元模型，构建了 Java 开发工程师的课程知识模型。

1.3.3 文献述评

通过上述研究现状的梳理，可以发现有关文本挖掘的研究内容主要从挖掘算法与算法应用的角度展开，并且文本分类和聚类的算法都发展较为成熟。关于网络招聘信息数据的研究主要分为两类：一是对获取到的网络招聘信息进行分析，探索网络招聘数据与某种社会经济现象之间的联系；二是对网络招聘数据的获取、建模及算法进行研究，提

升并改善各环节的效果。

现有研究缺少探讨数据类岗位的相关因素的分布与岗位所需技能，对于岗位要求所需技能缺少综合考量与横向对比，岗位要求缺少较为明确的分类描述，值得进一步探讨。

1.4 研究内容与技术路线

本文总共分为六章，按照章节划分如下：

第一章为文章的引言部分，主要阐述数据类岗位人才需求分析的研究背景、研究意义、国内外研究现状、研究内容和本文主要的技术路线图，最后简述本文研究的创新点。

第二章为文本挖掘相关技术理论，包括文本挖掘的定义与应用、文本数据预处理方法、文本表示方法等文本挖掘任务必备知识点。此外，着重介绍主题和主题模型相关立论、LDA 主题模型原理以及最优主题数目的确定方法。

第三章为数据采集与预处理过程，对本研究所使用的数据集来源、获取数据的方法、数据清洗过程等环节进行说明，为之后章节的数据探索性分析、数据可视化与文本主题挖掘过程实现做好准备。

第四章为数据类岗位特征探索性分析，对经过清洗过程的数据进行可视化分析。分别对“数据挖掘”、“数据分析”、“数据开发”、“数据运营”等 4 类岗位数据从不同角度进行数据分析与可视化呈现，主要包括岗位的薪资、工作经验需求、学历要求、城市需求的分布状况，并对岗位任职要求整体的需求信息进行基础的可视化展现。

第五章为基于 LDA 主题模型的数据类岗位招聘需求分析，主要是对“数据挖掘”、“数据分析”、“数据开发”、“数据运营”等 4 类数据类相关岗位的任职要求信息进行文本主题发现，尝试将岗位任职要求整体内容划分为若干个主题领域，并探讨最优的主题模型数目，最终对每类岗位中的每个主题内容进行深层次的分析，明确岗位需求内容，并使用 pyLDAvis 可视化技术呈现结果。

第六章为全文的总结与展望部分。首先总结全文主要的工作内容，梳理了数据类岗位人才招聘需求的分析流程，针对每部分的分析内容与每类岗位数据分布特点，结合实际情况对数据类人才提出相关建议，并希望通过基于文本数据挖掘的研究，为企业招聘、人才求职数据类岗位的顺利进行提供启迪。

本文的技术路线如图 1.1 所示。

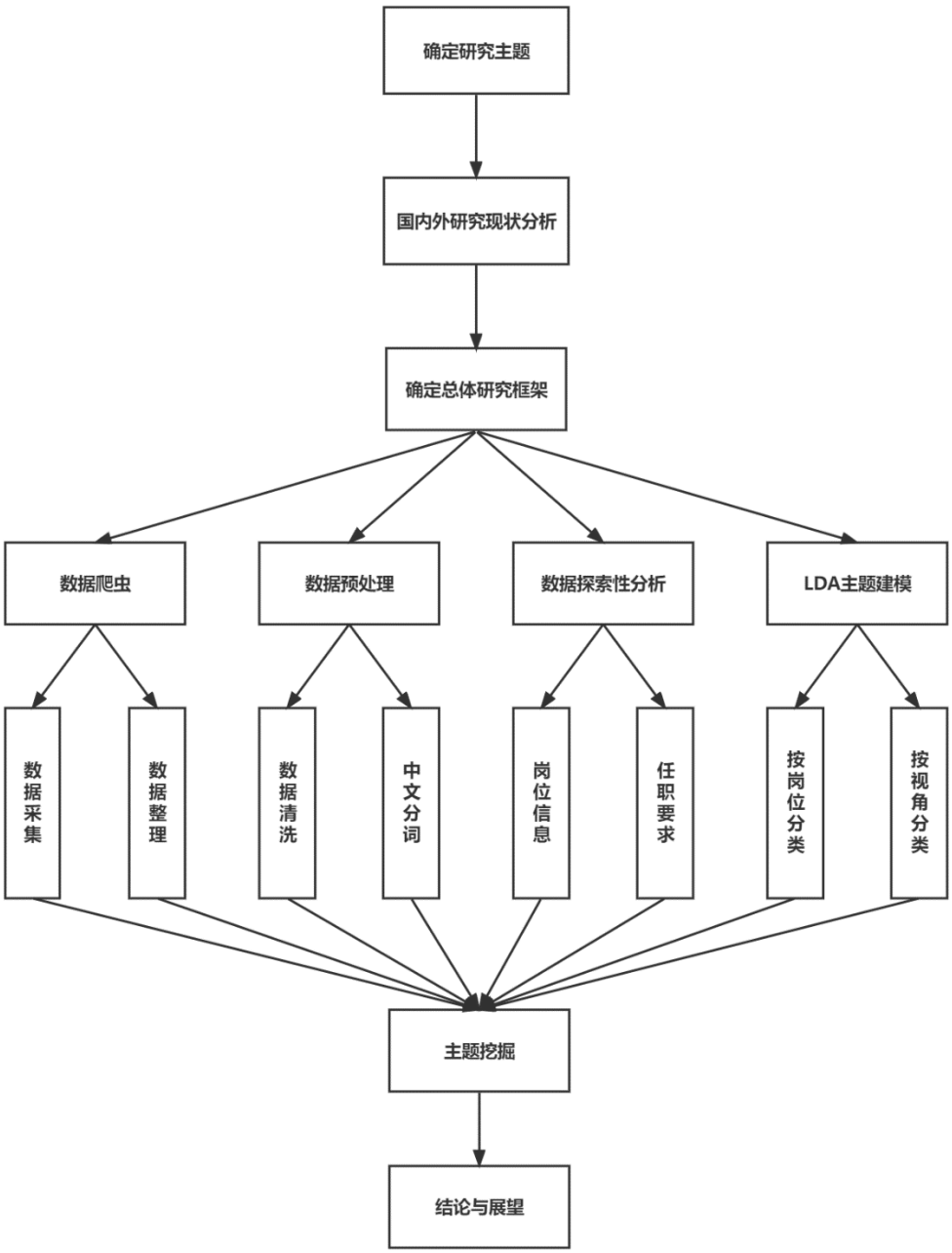

图 1.1 本文技术路线图

1.5 本文创新点

随着文本挖掘技术的不断成熟，对岗位需求的信息获取和挖掘过程的时间物力成本逐渐减低，有关网络招聘信息与数据类人才岗位需求分析的研究得到较大的发展。本文采用爬虫与文本挖掘、文本可视化技术，对岗位招聘信息进行分析，得到市场有关数据

类岗位人才的要求，本文的创新点有如下两个方面：

（1）研究方法的创新：现有的研究数据类岗位文本信息的文献，对于岗位分布特征描述或者岗位信息挖掘算法各有侧重，研究方法没有将传统的数值型数据与文本类的非结构化数据分析方法相结合，本文基于 LDA 主题模型算法，尝试将岗位分布特征和岗位信息挖掘过程相融合，进行主题内容的挖掘。

（2）研究视角的创新：现有的研究数据类岗位人才需求的文献，大多是以“大数据”为关键词对整体岗位进行分析，或是针对某一类岗位（如“数据分析师”）的需求信息进行分析。本文选取“数据分析”、“数据挖掘”、“数据开发”、“数据运营”等 4 类主流岗位为代表，对数据类岗位整体的人才需求进行分析与研究。

2 文本挖掘相关技术理论概述

2.1 文本挖掘概述

文本数据是指不能直接参与算数运算的任何字符，也成为字符型数据，如英文字母、汉字、不作为数值使用的数字和其他可输入的字符。文本数据不同于传统数据库中的数据，它具备自己的特点。文本数据主要特点包括以下几点：

（1）半结构化：结构化和非结构化的数据可能同时出现在同一条文本中。

（2）高维：文本向量的维数通常可达上万维，传统数据挖掘的方法在处理文本数据时存在失效的可能性。

（3）数据海量：通常，文本库中会同时存在成千上万个文本样本，文本数据预处理过程的工作量会非常庞大。

（4）语义性：文本数据具备其特定的语义内涵，并与相邻文本信息存在一定程度的联系，如在时间和空间上的上下文相关等情况。

文本数据中隐藏的内容信息可以通过文本挖掘方法进行分析，但是从复杂的自然语言中获取可用信息是较为困难的。随着文本挖掘技术不断被提出，有关文本数据的研究程度也愈发深入。1995 年，Feldman 和 Dagan 在非结构化数据研究中首次提出文本挖掘的相关概念，他们认为文本挖掘是一种基于机器设备将信息检索、信息提取、自然语言处理技术、数据挖掘、机器学习和统计相结合的文本分析方法。1995 年至今，诸多学者通过理论研究和实际应用证明了文本挖掘技术在从海量非结构化数据中挖掘并获取重要信息的领域具备强大的作用，使用文本挖掘技术能够较为科学合理地对文本信息开展分析，并寻找出具有价值的深层信息。

近些年，文本挖掘产生多个研究方向的分支，各分支侧重点不同但相互联系。陶洁曾检索并分析 Web of Science 中与文本挖掘相关文献，发现文本挖掘的研究主要涉及情感分析与主题分析、文本挖掘理论及主要算法模型、生物医学研究、概念与语义关系发现及其他领域应用等 5 大类。

2.2 文本预处理

网络上获取到文本信息形式复杂且多样，往往不够规范、整洁，计算机不能够直接

有效处理。因此，进行文本挖掘任务时首先要进行文本预处理。中文文本挖掘的预处理步骤包括分词、词性标注、去停用词、文本向量化表示等等，具体步骤的组成需要根据分析目的和文本数据的实际情况而定。

其中，分词过程是对中文文本进行断词，将一句文本分成若干词项和短语；词性标注是对分词后的单词和短语进行词性的信息标注；去停用词是把经常出现，但与文本主题内容相关性不大的词项进行删除。研究表明文本预处理过程在文本挖掘流程中起着重要作用，较高质量的文本预处理结果是后续分析效果的基本保障，下图 2.1 为文本预处理流程的基本步骤。

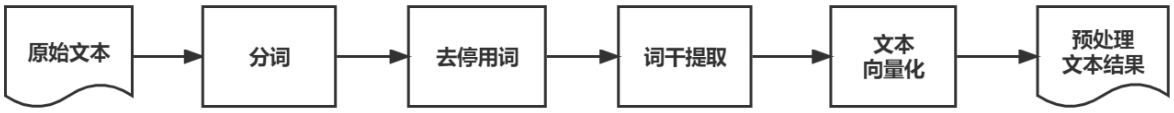

图 2.1 文本预处理流程图

2.2.1 中文分词

不同语言的文本分词方法往往存在差异。英文文本中每个词汇以空格间隔开，空格可直接作为切分点，但是中文的语法、语言习惯以及文本形式与英文有着很大的不同，要将中文的一段文字切分成一个个词的形式，需要根据中文语句的实际情况仔细分析。中文分词过程的难点主要有两处，第一点是分词结果容易产生歧义，不同的切分方法所产生的切分结果及其含义是不同的。如“我说不好”可以分词为“我\说不好”和“我说\不好”这两种情况，显然两种分词情况表达的含义是存在区别的。另一点是无法对专业词语直接分词进行识别，导致分词结果不准确。如“大数据”是指代海量的数据效果，但分词的结果却是“大数”、“据”两个词，“数据分析师”是指代使用数据完成分析任务的职业岗位，但分词的结果是“数据”、“分析师”，此时最重要的信息被切分开来，无法表示最初中文文本想要表达的信息。

实现中文文本分词可以借助 jieba 分词工具包，通过 jieba 分析后的中文语句会变为若干个单词和词汇。当分词后的文本包含表情符号、标点符号等干扰文本分词效果的内容时，可以通过 python 的 re 库调用正则表达式将其过滤。

2.2.2 停用词处理

中文文本中包含着如的、了、吗等频率较高，但对文本语义理解产生不了重要作用的词汇，这类词统称为停用词，剔除停用词可以提升文本分析的效率和效果。停用词处理的过程中，首先根据语料库构建停用词表，包含需要去除的词、符号和内容等自定义的停用词项目。通过加载停用词表遍历分词文本中的每个词语，当分词文本在停用词中出现，则该词被视为停用词，将该词从分词文本中删除，否则保留该词，最终保留的文本内容就是停用词处理后的结果。分词任务中经常使用的分词表有哈工大停用词表、四川大学停用词表、百度停用词表等。

2.2.3 文本向量化

文本表示是文本挖掘中的基础环节，对文本挖掘的效果和性能产生直接影响。文本向量化是通过一系列格式转换的方法，将文本表示成能够表达文本语义，且能够被计算机所识别从而进行运算的过程。本文对常见的文本向量化方法以离散表示和分布式表示进行简要说明。

2.2.3.1 离散表示

离散表示方法部分将对 One-hot 编码、词袋模型（BOW）的原理进行简要说明。

1. One-hot 编码

One-hot 编码是最基础的方法，关键步骤是构造文本分词后的字典。例如对“我喜欢吃苹果，苹果有营养”这句话构造字典，效果如下：

｛“我”：1，“喜欢”：2，“吃”：3，“苹果”：4，“有”：5，“营养”：6｝。

接下来，我们根据分词效果对每个词语进行向量化表示，用 0 和 1 来代表这个词是否出现，如“我”和“吃”这两个字的表示结果为：

“我”：[1, 0, 0, 0, 0, 0]和“吃”：[0, 0, 1, 0, 0, 0]。

One-hot 编码方式构造虽然简单，但往往不是一个好的选择，它有明显的缺点：

（1）维数过高。上面引用的例子表述每个词语已需要通过 6 维向量，并且随着语料的增加，维数将越来越大。

（2）矩阵稀疏。词向量的表述结果中只有 1 个维度取值为 1，其他维度上的数值都是 0，One-hot 编码容易产生矩阵稀疏的问题。

（3）无法保留语义。使用 One-hot 表示方法无法保留句子中的位置信息，进而会丢失掉句子的部分语义信息，如“我喜欢你”和“你喜欢我”的向量化结果并没有什么不同，但是两者表达的含义还是存在区别的。

2. 词袋模型

词袋模型（BOW）将文本看成是一系列单词的集合。在词袋模型中，一个文档的单词顺序和语法、句法等要素被忽略掉，只看作是若干个词汇的集合，文档中每个单词的出现都是独立的，不依赖于其他词语是否出现。

以“我喜欢吃苹果，苹果有营养”和“吃苹果身体好，可以多吃”这两句话为例，若将两句话视为一个文档集，将文档中出现的所有词语构造成一个字典。

｛“我”：1，“喜欢”：2，“吃”：3，“苹果”：4，“有”：5，“营养”：6，“身体”：7，“好”：8，“可以”：9，“多”：10｝。

然后，将句子向量化，向量的维数和字典大小保持一致，第i维上的数值代表 ID 为i的词语在这个句子里出现的频次，此时两个文本的表示结果分别为：

“我喜欢吃苹果，苹果有营养”：[1, 1, 1, 2, 1, 1, 0, 0, 0, 0]；“吃苹果身体好，可以多吃”：[0, 0, 2, 1, 0, 0, 1, 1, 1, 1]。

词袋模型不像 One-hot 编码那样导致维数非常大，但仍然存在无法保留语义信息、易产生高维与稀疏性的问题。

2.2.3.2 分布式表示

离散表示方法容易导致数据稀疏问题，分布式表示的方法应用更加广泛，其核心思想在于对一个词使用其附近的词进行表示，通过该词周边词汇共同构成其精确的语义信息。本文对共现矩阵和 Word2Vec 两种分布式表示方法进行说明。

1. 共现矩阵

共现矩阵主要挖掘词语共同出现的情况，词文档的共现矩阵主要用于主题发现。以“我喜欢吃苹果”、“吃苹果身体好”这两句为例，假定分词后的结果为｛“我”，“喜欢”，“吃”，“苹果”｝和｛“吃”，“苹果”，“身体”，“好”｝。此时，设置对称滑窗大小为 2，

可以得到{“我喜欢”,“喜欢吃”,“吃苹果”,“苹果吃”,“吃苹果”,“苹果身体”,“身体好”}等一系列词语，从而得到一个对称的共现矩阵，如表 2.1 所示。

表 2.1 共现矩阵计算结果说明

计数	我	喜欢	吃	苹果	身体	好
我	0	1	0	0	0	0
喜欢	1	0	1	0	0	0
吃	0	1	0	3	0	0
苹果	0	0	3	0	1	0
身体	0	0	0	0	0	1
好	0	0	0	0	1	0

表 2.1 中的每个元素表示的行和列组成的词组在词典中共同出现的次数。比如“吃苹果”出现在第 3、4、5 句话中，共出现 3 次，所以“吃苹果”在共现矩阵中的值为 3。对称窗口的含义为“苹果吃”在共现矩阵中的值也为 3。

2. Word2Vec

Word2Vec[37]是由 Google 于 2013 年提出的一种常用的词嵌入模型，能够将文本语料中的特征词转化为多维的实数向量。Word2Vec 可理解为一种包含输入层、投影层和输出层共三层结构的神经网络模型，它包含两种网络结构形式分别是 CBOW 和 Skip-gram，以下这两种模型结构的详细说明。

（1）CBOW 模型

CBOW 可用于输入已知上下文，输出对当前单词的预测。CBOW 把中间词当作 y，把窗口中的其他词当作 x 输入，然后通过一次隐藏层的求和操作，并通过激活函数 softmax 计算出每个单词的生成概率，然后训练神经网络的权重，使得语料库中所有单词的整体生成概率达到最大化，而最终求得的权重矩阵就是文本表示词向量的结果。CBOW 模型过程如式(2.1)-(2.2)所示，CBOW 模型框架见图 2.2。

假设有一个句子结构为 $w_{n-2},w_{n-1},w_n,w_{n+1},w_{n+2}$，CBOW 模型是通过输入 $w_{n-2},w_{n-1},w_{n+1},w_{n+2}$ 来预测 w_n 的词向量。CBOW 模型的表达式如下：

$$p(w_n \mid w_{n-k},w_{n-k+1},\cdots,w_{n+k-1},w_{n+k}),k\text{为滑窗大小} \tag{2.1}$$

CBOW 根据上下文对目标词进行预测，对 $w_1, w_2, \cdots, w_n$ 训练语料，预测公式如下：

$$p(w \mid context(w)) = \frac{\exp(e(w)^T x)}{\sum_{w' \in V} \exp(e(w')^T x)} \tag{2.2}$$

其中， $x = \sum_{i=1}^{n} e(w_i)$ ， $e(w_i)$ 为词 w_i 的词向量。

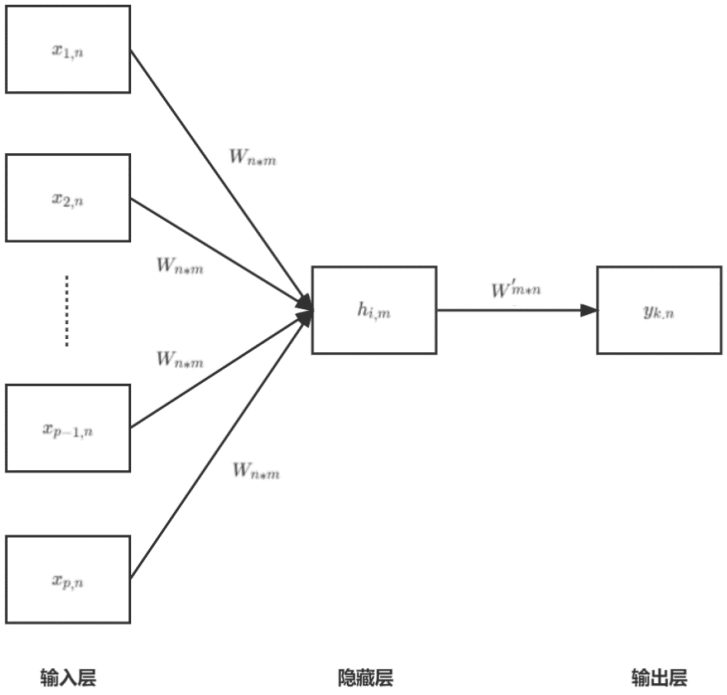

图 2.2 CBOW 模型结构

（2）Skip-gram 模型

与 CBOW 恰好相反，Skip-gram 把当前词当作 x，窗口中其他词当作 y，主要用于已知某个词语来预测周围的词语。模型计算的主要过程与 CBOW 类似，如式(2.3)-(2.5)所示，Skip-gram 模型框架见图 2.3。

假设有一个句子的结构为 $w_{n-2}, w_{n-1}, w_n, w_{n+1}, w_{n+2}$，Skip-gram 模型是通过输入 w_n 来预测 $w_{n-2}, w_{n-1}, w_{n+1}, w_{n+2}$ 的词向量。Skip-gram 模型的表达式如下：

$$p(w_i \mid w_n), n-k \le i \le n+k, 2k+1 \text{为运行滑窗大小} \tag{2.3}$$

Skip-gram 模型预测特征词上下文的方法如下：

$$p(content(w)\mid w)=\prod_{u\in content(w)} p(u\mid w), w\text{为当前词}, u\text{为周围词} \tag{2.4}$$

则有

$$p(u\mid w)=\frac{\exp(e(u)^T e(w))}{\sum_{w^{'}\in V}\exp(e(u)^T e(w))} \tag{2.5}$$

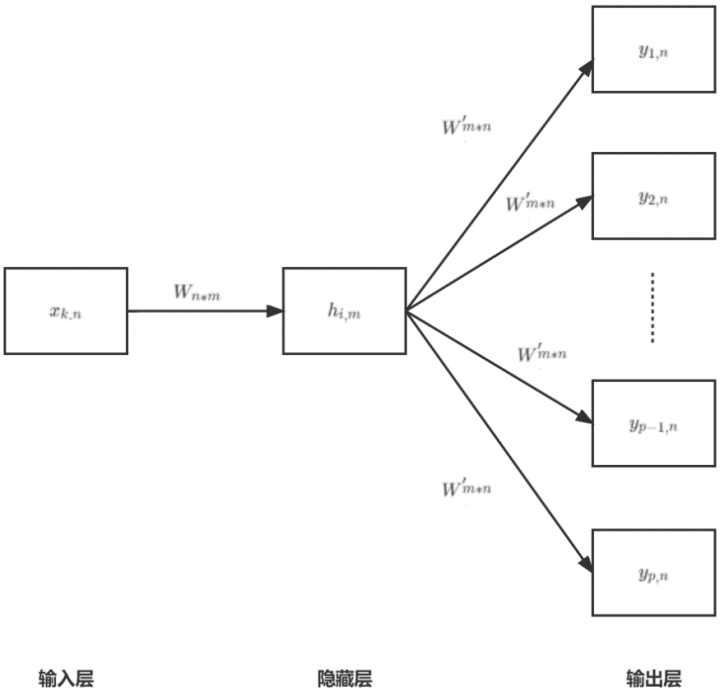

图 2.3 Skip-gram 模型结构

对比两种模型框架可以发现，Skip-gram 模型通常进行预测的次数更多，训练时间也比 CBOW 更长。同时，由于 Skip-gram 模型中的每个词都会受到周围词的影响，对于选定的某个词作为中心词时，都会进行 k 次的预测和调整，这个过程使得词向量相对更加准确，效果可能会更好。不过当数据量较大时，Skip-gram 模型的计算成本较大，耗费的时间更长，具体方法的确定要结合实际情况进一步判断。

2.3 LDA 主题模型

2.3.1 主题和主题模型

我们首先需要明确主题[38]的概念。主题是指社会活动或者文艺作品等所要表现出的中心思想，也可称为主要内容。通俗来讲，主题可以理解为一系列相关的词语所表示的主要内容，主题模型就是对这一系列文本的隐含主题进行挖掘的建模方法。

在传统的信息检索领域，可以通过统计文章中每个词语出现概率的大小来判断词语成为主题词的可能性，其隐含的观点是词频越高的词语成为主题词的可能性越大。但是，随着文档内词语数量的增大，这种仅关注词频而忽略词语语义的方法效果不佳，无法充分准确地表示主题内容。

在文本挖掘与自然语言处理领域，主题可反映为词项的实际概率分布，通过主题模型利用词项完成文档信息的提取，获得语义内容相关且类似的主题集合，同时把词项中的相关文档变换为主题空间，最终在文档的主题空间中进行表达，可理解为对文档进行降维处理。相较于传统的计量经济学等处理结构化数据的经典方法，主题模型能够深层次挖掘词语语义之间的关系，例如对新闻、文学作品、文献计量等进行研究，进行主题发现与预测等任务。

一个主题模型通常包含 5 项内容：主题模型的输入、基本假设、主题模型的表示、参数估计和新样本推断。常见的主题模型包括概率潜在语义分析（PLSA）模型、潜在狄利克雷分布（LDA）模型以及扩展模型。

2.3.2 LDA 模型理论基础

潜在狄利克雷分布[39]（Latent Dirichlet Allocation，简称 LDA）模型由 Blei 等人于 2003 年提出，是一种包含词项、主题和文档三层结构的贝叶斯模型。LDA 主题模型首先寻找文档的主题分布，并通过计算分词文档的主题分布和主题对应的词分布，完成主题词提取与主题发现。

LDA 模型假设每篇文档中的每个词都服从一定的概率分布，文档到主题、主题到词都服从多项式分布。词语间存在关联性，不同的词语可以划分归属为不同的主题，使用出现概率最大的一个或者多个主题表示该篇文章的主题含义。模型中的主要参数为α 和

β，详见表 2.2。其中，α 是狄利克雷先验的中心化参数，表示文档-主题集中度。α 取值较大时，文档被假定为由更多的主题组成，每篇文档生成更复杂的主题分布。类似的，β 是相同的中心化参数，表示文档-词语集中度。β 取值较大时，主题被假定为由大部分词汇组成，每篇文档生成更复杂的词汇分布。

表 2.2 LDA 模型主要参数

符号	含义
M	文档个数
K	主题个数
V	词项个数
α	θ_m 的先验分布超参数（K 维向量）
β	φ_k 的先验分布超参数（V 维向量）
θ_m	第 m 个文档的主题分布参数
φ_k	第 k 个主题的词项分布参数
N_m	第 m 个文档的长度
$z_{m,n}$	第 m 个文档第 n 个词对应的主题
$w_{m,n}$	第 m 个文档第 n 个词对应的词项
$z_m = \{z_{m,n}\}_{n=1}^{N_m}$	第 m 个文档对应的主题序列
$w_m = \{w_{m,n}\}_{n=1}^{N_m}$	第 m 个文档对应的词项序列
$w = \{w_m\}_{m=1}^{M}$	文档集对应的词项序列
$z = \{z_m\}_{m=1}^{M}$	文档集对应的主题序列

对 LDA 模型算法主要过程进行说明，首先作如下假设：假设共有 M 篇文档，共包含 K 个主题；每 m 篇文档的总次数为 N_m，每篇文档都有各自的主题分布，且服从参数为 α 的多项分布，α 是主题分布的 *Dirichlet* 先验参数；每个主题都有对应的词分布，词

分布服从参数为 β 的多项分布， β 是词分布的 *Dirichlet* 先验参数。

第 m 篇文档中的第 n 个词的生成过程如下：首先，按照概率选择一篇文档 m ，从 *Dirichlet* 分布 $Dir(\alpha)$ 中抽样生成文档 m 的主题分布 θ ，从主题分布 θ 中抽样生成第 m 篇文档的主题 z ，从 *Dirichlet* 分布 $Dir(\beta)$ 中抽样生成主题 z 的词分布 φ ，从词分布 φ 中抽样最终生成单词 w ，不断重复这个过程，直到 M 篇文档全部生成。生成流程如图 2.4 所示。

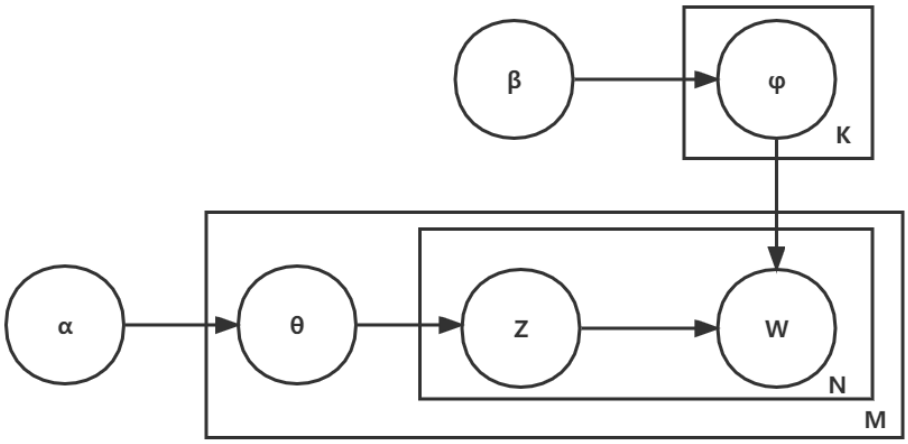

图 2.4 LDA 生成模型示意图

已知的数据是词向量 w ，对于主题 z 的分布是未知的。此时，可先求得词向量与主题的联合分布 $p(w,z)$ ，再求得某词 w_i 的主题是 z_i 的条件概率分布 $p(z_i=k|w,z_{-i})$ ，其中 z_{-i} 表示除去当前主题 z_i 后的主题分布。得到该条件分布之后，可以采用吉布斯采样进行抽样求解。

吉布斯采样[40]（Gibbs Sampling）可看作是马尔科夫蒙特卡罗算法的一个特例，该算法可以求解主题模型的参数。求解方法是计算给定其他坐标维度的变量值的条件下，采样得到当前坐标维度的值，每次选择概率向量的一个坐标维度，算法不断迭代直至收敛，最后得到主题模型的待估参数。算法思路如下：一开始随机给文档中的词分配主题，然后统计每个主题 z 出现词 t 的数量和每篇文档 m 出现主题 z 的数量，计算概率 $p(z_i|z_{-i},d,w)$ ，该概率衡量用其他所有词的主题分布来估计当前词分配主题的效果；用此概率分布，给当前词采样一个新的主题；重复上述步骤，得到下一个词的主题，直到每

篇文档的主题分布θ_i和每个主题的词分布φ_j收敛，此时分别可以得到所需主题分布和词分布的参数。

2.3.3 最优主题数的确定方法

评估主题模型的效果也是值得重点讨论的领域，困惑度（Perplexity）是 LDA 模型常用的评价方法之一。对于一篇文档d，训练出的模型对于文档d属于哪个主题的不确定程度就是困惑度。困惑度取值越低，说明 LDA 模型效果越好，在当前主题数目下的模型泛化能力越强。

困惑度的计算公式如式 2.6 和 2.7 所示：

$$perplexity(D)=\exp\left\{-\frac{\sum_{m=1}^{M}\log_D P(w_m)}{\sum_{m=1}^{M}N_m}\right\} \tag{2.6}$$

$$P(w_m)=\sum_d\prod_{n=1}^{T}\sum_{j=1}^{T}p(w_j\mid z_j=j)p(z_j=j\mid w_m)p(d) \tag{2.7}$$

其中，D表示语料库中的测试集，M表示文档数量，N_m表示文档m中的单词数量，$P(w_m)$表示w_m产生的概率。

显然，可以通过困惑度来确定主题数目。通过不同的主题数构建模型，求得困惑度最小值对应的最优主题数目。不难发现，在其他条件不变的情况下，主题数越多，困惑度就越小，但随之而来的问题是过拟合，因此合适的困惑度大小是确定主题数的重要参考标准。

除主题困惑度之外，一致性得分（Coherence）也是评估 LDA 模型效果的常用指标。主题一致性得分主要通过衡量主题间得分较高单词之间的语义相似程度，进而为单个主题打分，这种度量方式可用于区分语义上可解释的主题。具体的，用于计算评估 LDA 模型的一致性得分方法主要是 UCI 度量和 UMass 度量，这两种方式都是通过计算主题词集合V的相似性得分的和，将其作为一致性的度量方法，如式(2.8)所示：

$$coherence(V)=\sum_{(v_i,v_j)}score(v_i,v_j,\varepsilon) \tag{2.8}$$

其中，V 是一系列描述主题的词汇集合，ε 是确保一致性得分为实数的平滑算子。

UCI 度量通过计算两个词汇间的逐词的互信息，以求得词汇的配对得分，见式(2.9)。

$$score(v_i, v_j, \varepsilon) = \log \frac{p(v_i, v_j) + \varepsilon}{p(v_i) * (v_j)} \tag{2.9}$$

其中，词汇概率的计算方式是对计算语料库上的滑动窗口中的共现频率。

UMass 度量是基于文档共现情况的得分，见式(2.10)。

$$score(v_i, v_j, \varepsilon) = \log \frac{D(v_i, v_j) + \varepsilon}{D(v_j)} \tag{2.10}$$

其中，$D(x, y)$ 记录包含词 x 和 y 的文档数量，$D(x)$ 记录包含词 x 的文档数量。

对比 UCI 度量和 UMass 度量方式可以发现，UCI 度量可以看作是对外部的语料库与已知语义的比较，而 UMass 度量方式侧重于对训练主题模型的原始预料库的比较。

3 数据采集与预处理

3.1 数据采集

本文研究对象为全国范围内数据类招聘岗位，主要研究内容是数据类岗位招聘信息内容与需求情况。网络招聘信息可以便捷地了解企业对人才的需求，从而反映出数据类人才市场的需求特点。因此，本文主要通过采集网站上的招聘信息，对招聘岗位信息进行分析与挖掘，从而得到所关心的数据类岗位人才需求结果。

3.1.1 数据来源

本文通过第三方招聘网站前程无忧网来获取招聘信息，该招聘平台积累用户数量较多，是使用量较大的传统类型的招聘网站。

前程无忧网站发布的招聘信息主要涉及到三个方面：第一个方面是招聘单位的相关信息，包括企业名称、企业类型、企业规模与所属行业等等；第二个方面是岗位的相关信息，包括岗位名称、工作地点、工作经验和学历等要求，岗位的薪资水平、福利待遇、招聘人数以及岗位信息发布日期等等；第三个方面是岗位的需求信息，主要包括岗位职责和任职要求，岗位职责描述的是岗位的工作内容，任职要求主要描述任职者所需要具备的工作技能、业务知识和个人技能等实力信息。

如图 3.1 所示，以使用“数据分析”作为关键词检索为例，在前程无忧网中进行检索，可以得到相关岗位的招聘信息；点击某条招聘信息，可以进入岗位信息的详情页面，该页面中记录着包括岗位的任职要求等详情信息，如图 3.2 所示。本文后续分析的数据都将来自于该网页内容。

图 3.1 前程无忧招聘岗位基本信息页面

图 3.2 前程无忧招聘岗位详细信息页面

3.1.2 数据采集过程

网络爬虫技术可以从网页的海量信息中爬取数据信息并存储到本地。相比于实地访谈调研等人工获取数据的方式，网络爬虫技术可以实现海量数据的获取，同时降低了数据采集的成本，避免了人工整理数据时产生的非系统性误差。通过 Python、R 等编程语言可以实现网络爬虫任务，但对操作人员的编程水平要求较高。在遵循爬虫原理的基础上，弓箭手、八爪鱼等数据采集软件通过流程操作使得爬虫任务更加便捷化，对于初学者非常友好。本文使用八爪鱼采集器进行数据类岗位招聘信息的采集工作，八爪鱼采集器模拟人访问页面、爬取页面时的思维方式，对网页源码各数据位置进行精确定位，从而对用户所需要的数据进行准确采集。

3.1.3 数据采集结果

笔者于 2021 年 10 月在前程无忧招聘网站中检索全国范围内的“大数据”相关职位，分析检索得到的招聘信息，发现数据类岗位存在各行各业中，不仅包括互联网等新兴产业，传统制造业中也呈现出对数据类人才的需求，如工业领域中的异常检测问题就需要人才对海量工业数据的异常模式进行挖掘。

通过对数据类岗位描述的初步判断，发现数据类岗位可能存在众多的应用领域与发展方向。为了能够覆盖数据类岗位的大部分领域，本文对 2021 年 10 月 16 日前程无忧网站的招聘信息进行了采集，岗位地区设定为全国范围，以“数据分析”、“数据挖掘”、“数据开发”、“数据运营”这 4 个关键词进行检索，共采集到 19437 条招聘数据。采集结果中，数据分析的相关招聘信息有 7846 条，数据挖掘为 5178 条，数据运营为 3978 条，数据开发为 2435 条。

本文所采集到的数据字段包括：招聘岗位的名称、薪资、所在地、工作经验要求、学历要求、专业要求、福利标签、职位要求、公司名称、公司类型、公司规模、所属行业。这些岗位的招聘信息可以分成两部分：与岗位信息相关的分类型数据、数值型数据以及与岗位任职要求相关的文本数据, 如图 3.3 所示。

关键词	招聘岗位	薪资	所在地	工作经验	学历要求	专业要求	福利标签	职位要求	公司名称	公司类型	公司规模	所属行业
数据分析	商业化数据分析师（	9-12万/年	宁波-鄞州区	2年经验	本科	02-17发布	五险一金 专业	职责1、监控、分析用户运营数据，针对业务问题/需求，分析产品、机构、客户等数据，建立统计，根据运营数据提出产品构想、策略及计划；2、负责挖掘并分析行业的现状及需求，负责研究市场及竞品，进行分析对比，提供产品策略和运营建议；3、定期或结合重要项目输出数据分析报告，包括用户分层、成长和激励体系数据，给出明确的优化方案，提升用户体验，提升新用户转化率、老用户的活跃率和活跃度运营，降低用户转化成本；4、日常数据监控，及时反馈数据异常，发现并报告问题，对源数据进行收集、整理、存档，完善部门数据支撑平台。任职资格1、本科及以上学历；2、熟悉各种数据处理软件和分析工具，有结合内容营销的数据运营能力，有电商、APP数据分析经验优先考虑；3、有良好的沟通能力，优秀的逻辑思维能力，归纳总结能力	宁波艺羽供应链管理有限公	民营公司	500-1000人	外包服务 银行
数据分析	供应链物流管理主管	9-12万/年	广州-天河区	3-4年经验	本科	02-18发布		工作职责:1. 负责供应商合同/订单对接，跟进发货计划及进度，以及进出口货物的单据制作、审核、货物运输、通关的相关工作。2. 负责供应商、报关公司、货代、仓储物流管理，确保货物运输时效及各供应商KPI考核要求，定期输出KPI报告。3. 监控团队日常操作流程，控制物流成本，协助组员处理突发问题。4. 定期分析进出口各环节数据，根据数据呈现的问题及配合公司部门目标，输出制定优化及完善方案。5. 提升团队能力，确保供应链物流各环节运作规范、高效，优化工作流程，提高各相关部门之间工作沟通效率。任职资格:1. 本科及以上学历；具备两年以上供应链物流进出口团队管理经验优先。2. 具备三年或以上进出口货运（海/空/陆运）相关经验，能独立完成订单由发货到收货全流程。有鞋服配进出口经历者优先3. 熟悉进出口贸易及通关单证制作，特别是东盟进出口贸易4. 良好的沟通能力及数据分析能力。5. 熟练ERP系统及WORD/EXCEL/FOXMAIL等办公软件。	成都斯凯奇贸易有限公司	外资（欧美）	150-500人	批发/零售

图 3.3 数据类岗位招聘信息内容

3.2 数据预处理

本文所采集到的岗位信息中，除了职位描述，其它都是结构化数据，在进行后续分析之前，首先要对数据进行预处理。数据预处理是开展后续分析中重要的一步，需要耗费较大的精力，数据质量优劣程度会直接影响分析效果和可靠性。

3.2.1 数据清洗

数据清洗包括去除信息中重复的、有误的数据，数据清洗过程非常有必要。本文数据清洗过程主要包括以下步骤：

（1）删除包含错误采集信息的数据，这部分可以通过人工的方式进行初步筛查，如采集结果是否包含串行、大面积缺失等等情况。此外，采集数据可能包括不是数据类岗位的招聘信息，这部分数据也要剔除。分析数据采集结果发现，有些岗位采集结果会出现把“公司规模”结果保存在“学历要求”字段的内容中。我们分析发现“公司规模”字段内容都以“xx-xx 人”的形式存在，为解决该问题，我们使用 python 对“学历要求”字段内容进行遍历判断，将“学历要求”中出现以“xx 人”结尾的数据所在条目删除，完成学历要求中异常数据的清洗过程。

（2）重复值处理：招聘信息中存在大量的重复数据，公司在没有寻找到合适的人才时，可能会持续的将招聘信息放在招聘网站、并且可能会多次发布类似或相同的岗位信息，导致招聘信息重复。因此，要对由同一公司发布、职位描述相似的招聘信息剔除，

剔除重复数据时保留最新更新的信息。

（3）缺失值处理：在采集的数据中，有些字段可能会存在空值，比如招聘网页中岗位福利等非必填项，企业在发布岗位信息时可能会忽略掉此项内容。对于存在缺失值的数据，可以选择直接删除或者缺失值填充的方式。直接删除的方式简单高效，但是可能会把其他有价值的信息同时删掉，数据量较大时可以选择此种方式；另一种是对空值进行估计，估计方法也有很多，但是对于招聘岗位信息而言，大部分数据字段都是文本形式，数值型数据求均值等填充方法不适用，因此本文对于包含缺失值的数据将直接删除。

3.2.2 数据规范化

企业在网站中发布岗位信息时没有统一的数据录入标准，造成我们所采集到的数据可能存在单位、格式不统一的问题。比如岗位薪资字段存在“万/月”、“万/年”、“千/月”、“元/天”等 4 种不同的形式。为了后续能够进行横向纵向等多维度的比较，我们要在文本预处理阶段解决薪资格式不统一的问题。

具体的解决思路如下：首先将数据导入 python 中，构造以“万/年”、“千/月”结尾的文本字符，再对每条岗位信息的薪资字段遍历进行逻辑判断，具体使用的 re 模块中的 re.findall 方法进行匹配，把符合条件的数据单位统一转换为“万/月”，如将“8-9 千/月”转换为“0.8-0.9 万/月”、“9-12 万/年”转换为“0.8-1 万/月”等等。对于“元/天”结尾的数据直接剔除。

3.2.3 中文分词

本文通过调用 jieba 分词的精确模式完成中文文本分词，可以实现对中文文本内容基本的分词。此外，本文采集的岗位描述中记录着大量的数据类专业词汇。例如，“随机森林”在数据挖掘中指一种集成学习算法，但是 jieba 内置普通的分词模块会将该词划分为“随机”和“森林”，显然会对后续分析产生一定的误导。为避免这种情况出现，本文在使用 jieba 库分词时添加自定义词典。根据本文采集的招聘岗位与数据类相关信息，分别添加编程软件、专业知识、业务相关和基础条件等四类词汇，这种方法将有效提高招聘岗位信息文本分词的效果。数据类岗位技能词典部分内容如表 3.1 所示。

表 3.1 数据类岗位技能词典部分内容

技能维度	技能指标
编程软件	R，Python，SAS，SPSS，sql，MySQL，Java，C， C++，Matlab，Hadoop，Hive，Oracle，MongoDB， Julia，Scala，Kafka，Storm，Octave，Spark， Shell，Flink，HBase，Druid，ETL
专业知识	数据分析，数据结构，数据思维，数据敏感，数据挖掘， 数据运营，数据开发，数据结构，数据仓库，数据库， 数据建模，机器学习，深度学习，图像处理，语音识别， 算法设计，开发语言，前端开发，网络爬虫，技术框架， 数据汇报，数据清洗，数据可视化，自然语言处理，NLP
业务相关	业务数据，业务需求，业务分析，工作经验，思维能力， 开发能力，团队精神，抗压能力，客户需求，推荐系统， 数据产品，精准营销，上进心，理解能力，逻辑思维， 商业智能，学习能力，沟通，协调，工作能力，领导能力 语言表达能力，合作，进度把控，竞品分析
基础条件	专科及以上学历，本科及以上学历，经验不限，学历不限， 计算机相关，统计相关，经济管理相关，软件工程， 应用统计，相关专业优先，硕士优先，博士优先， 统计学，人工智能相关，英语四级，英语六级

3.2.4 去停用词

分词后的文本主要包括若干短句或词汇，但有些分词结果是我们不必使用且不关心的。为了降低分析文本的维度，减少对关键词频的影响，在进行中文分词之后，还需要去停用词处理。

本文主要通过哈工大停词表完成停用词处理。此外，由于本文主要研究的是数据类岗位的招聘信息，因此将诸如“背景”、“要求”、“岗位职责”、“优先”等对于岗位详细信息分析没有价值的词也加入停用词表中，以得到更优的停用词处理效果。

4 数据类岗位描述性统计分析

在对岗位描述进行深层挖掘分析之前，本章首先对清洗后的数据类岗位信息进行描述性统计分析，从需求岗位的城市分布、经验要求分布和学历要求出发，进而讨论薪资与工作经验要求、学历要求的变化情况，最终使用 python 的 matplotlib 和 pyecharts 库绘制可视化图形，展示分析效果[41]。

4.1 需求岗位的城市分布

通过对清洗后数据的公司名称、公司地点进行统计，分别得到全国范围内数据分析、数据挖掘、数据开发和数据运营岗位城市需求情况。如图 4.1 所示，图中刻度条颜色从冷色调到暖色调代表该地区对应岗位需求人数逐渐增大，红色点代表岗位需求比较大，蓝色点代表岗位需求比较小。不难发现这四种岗位需求较大的城市分布在我国的华东、华南地区。整体而言，岗位需求较大的城市普遍位于我国东部地区。

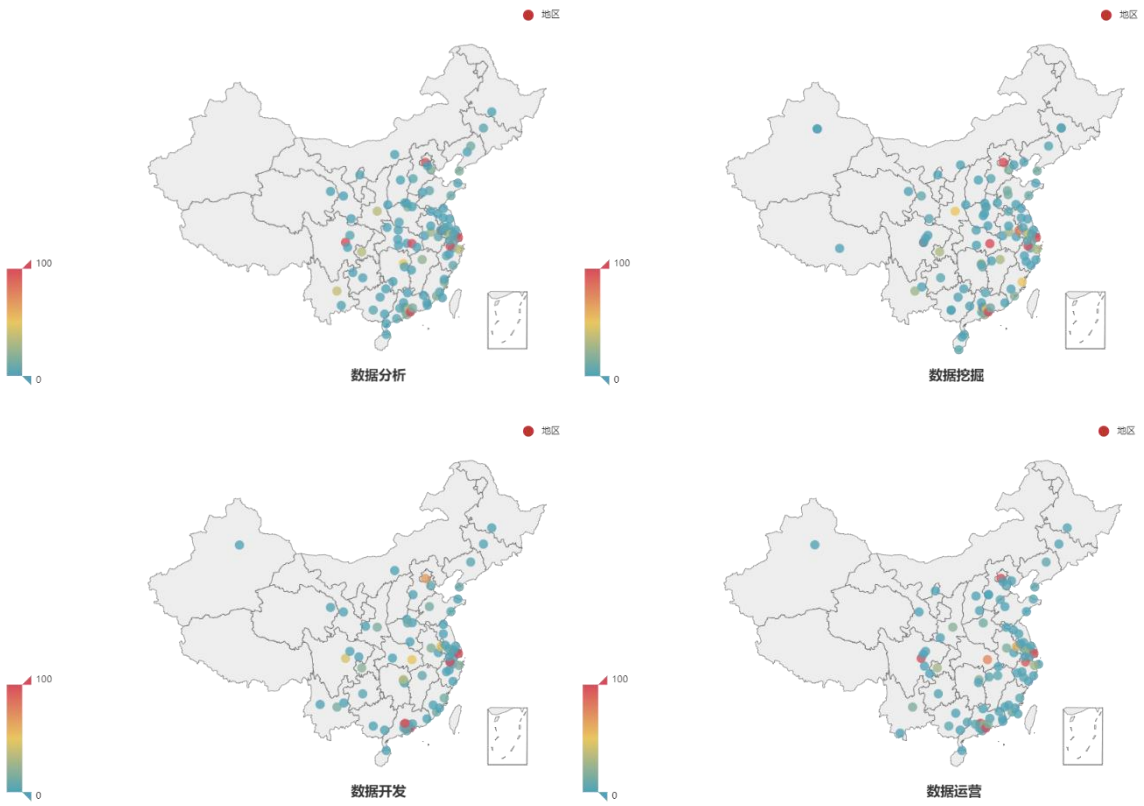

图 4.1 数据类岗位全国城市需求分布

为了对进一步探究岗位所在城市需求，将这四类岗位城市需求的前 10 位城市及其岗位需求数量进一步整理，见表 4.1。

表 4.1 数据类岗位全国城市需求前 10 位

岗位	城市需求前 10 位	频次	岗位	城市需求前 10 位	频次
数据分析	上海	575	数据开发	上海	267
	武汉	117		广州	135
	杭州	115		深圳	132
	北京	113		杭州	107
	成都	112		北京	105
	南京	93		南京	52
	苏州	57		武汉	49
	长沙	51		成都	44
	东莞	43		苏州	30
	西安	35		宁波	16
数据挖掘	上海	522	数据运营	广州	362
	深圳	386		上海	357
	北京	192		深圳	207
	杭州	133		北京	144
	武汉	116		成都	102
	成都	97		杭州	96
	南京	82		武汉	75
	福州	51		南京	60
	西安	50		苏州	36
	东莞	49		重庆	28

在这四类岗位需求城市的前 10 名中，出现频次较多的城市有：上海、广州、深圳、北京、杭州、武汉、成都、南京、苏州等等。不难发现，数据类岗位需求较大的城市是在我国的一线城市（北京、上海、广州、深圳）。其中，上海市在这四类岗位的需求排名中是最高的，出现了 3 次需求数量第一位的情况，说明上海市数据类岗位不仅需求大，并且需求岗位侧重点较为均衡，与数据类相关的工作机会均较多。其次，广州在数据开发和数据运营岗位的行业需求较大，同属于广东省的兄弟城市——深圳，则在数据挖掘、数据开发等更偏向技术领域、硬件行业的岗位需求更大。北京作为我国首都，各行业经济水平发展较为领先，是数据类岗位需求较大的城市中唯一的北部城市。

在北京、上海、广州、深圳这四个一线城市之后，新一线城市中数据类岗位机会较多的城市包括杭州、成都、南京、武汉和苏州。杭州地理位置优越，紧邻上海，江浙沪地区经济活动较为频繁，民营企业发展起步较早，能够提供就业机会的中小企业较多，

是一线城市外寻求数据类岗位的最佳选择。南京和苏州与杭州相似，但是在需求数量上会略少些。此外，成都和武汉作为举各自全省之力发展的城市，人口规模较大，知名企业较多，提供的工作机会较多，人才引进政策较优，发展势头迅猛。

4.2 经验要求分布

求职者的工作经验与项目经验是求职者找工作时企业较为看重的，通过对这四类岗位的招聘信息中的经验要求字段分析发现，无需经验的要求在数据分析、数据挖掘、数据开发、数据运营思维岗位的比例分别为 8.45%、9.08%、6.20%和 12.49%，数据运营岗位的经验要求较低、数据开发岗位的经验要求较高；对求职者要求有 3-4 年工作经验的岗位数量是最多的，这与数据类岗位自身特点有关，一名数据类岗位从业者的工作价值会随着工作年限与工作经验的逐步累积增大，这个特点在数据类岗位领域非常明显。此外，将“1 年经验”、“2 年经验”、“3-4 年经验”要求所占比例进行汇总，发现提出“1-4 年工作经验”要求的数据类岗位占比情况：数据分析为 83.44%、数据挖掘为 74.83%、数据开发为 77.98%与数据运营的 74.84%。数据分析岗位对求职者提出一定的结合业务知识与编程的能力，因此对工作经验较为看重。

表 4.2 数据类岗位经验要求分布

岗位	经验要求	频次	百分比(%)	岗位	经验要求	频次	百分比(%)
数据分析	无需经验	223	8.45	数据开发	无需经验	74	6.20
	1 年经验	789	29.90		1 年经验	188	15.75
	2 年经验	673	25.50		2 年经验	255	21.36
	3-4 年经验	740	28.04		3-4 年经验	488	40.87
	5-7 年经验	198	7.50		5-7 年经验	170	14.24
	8-9 年经验	11	0.42		8-9 年经验	14	1.17
	10 年以上经验	5	0.19		10 年以上经验	5	0.42
数据挖掘	无需经验	247	9.08	数据运营	无需经验	226	12.49
	1 年经验	441	16.21		1 年经验	430	23.77
	2 年经验	583	21.43		2 年经验	408	22.55
	3-4 年经验	1012	37.19		3-4 年经验	516	28.52
	5-7 年经验	383	14.08		5-7 年经验	199	11.00
	8-9 年经验	37	1.36		8-9 年经验	4	0.22
	10 年以上经验	18	0.66		10 年以上经验	26	1.44

此外，对这四类岗位每类岗位经验要求最多的岗位及其数量进行统计，发现数据分

析岗位对于求职者拥有 1 年经验的要求最多（29.90%），数据挖掘岗位对于 3-4 年经验的要求最多（37.19%），数据开发岗位对于 3-4 年经验的要求最多（40.87%），数据运营岗位对于 3-4 年经验的要求最多（28.52%）。对于这四类岗位里每类岗位的经验要求状况及其人数绘制漏斗图，漏斗图向下的方向代表该范围的经验要求状况在逐渐减少，如图 4.2 所示。

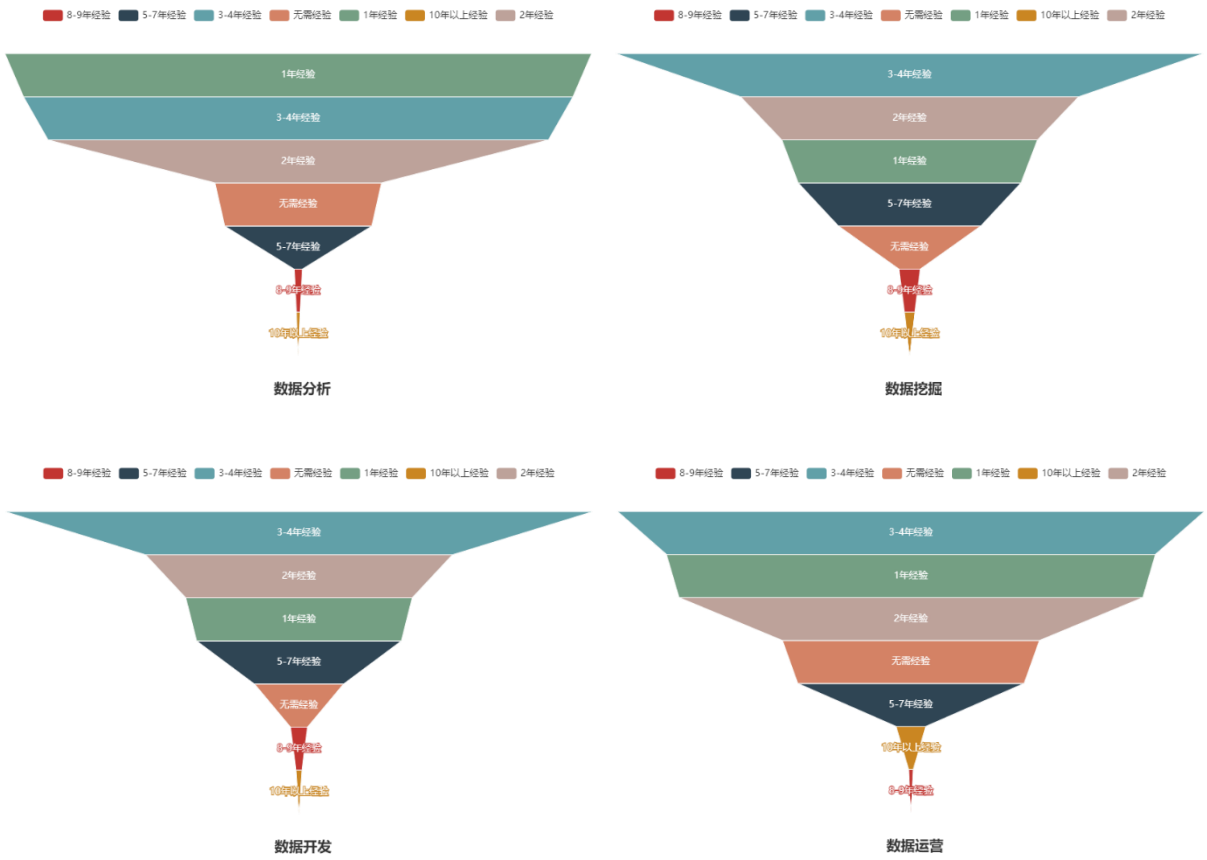

图 4.2 数据类岗位经验要求漏斗

4.3 学历要求分布

对比数据类岗位学历要求分布状况可知，如表 4.3 所示，发现这四类岗位关于学历要求最多的是本科，大专学历要求数量次之，大专和本科构成学历要求中占比最大的部分：数据分析岗位为 94.69%，数据挖掘为 94.46%，数据开发为 96.06%，数据运营为 92.76%。由于学历字段具有定序属性，不难发现，相比于数据运营和数据分析，数据挖掘和数据开发岗位的就业门槛会略高一些，对求职者的学历要求更高，而数据运营岗位对于学历要求则较低，这四种数据类岗位各自对应的学历要求情况见图 4.3。

表 4.3 数据类岗位学历要求分布

岗位	学历要求	频次	百分比(%)	岗位	学历要求	频次	百分比(%)
数据分析	高中	21	0.80	数据挖掘	高中	3	0.11
	中专	26	0.99		中专	4	0.15
	大专	1034	39.18		大专	615	22.58
	本科	1465	55.51		本科	1958	71.88
	硕士	88	3.33		硕士	125	4.59
	博士	5	0.19		博士	19	0.70
数据开发	中专	3	0.25	数据运营	高中	23	1.27
	大专	283	23.70		中专	88	4.86
	本科	864	72.36		大专	600	33.17
	硕士	40	3.35		本科	1078	59.59
	博士	4	0.34		硕士	20	1.11

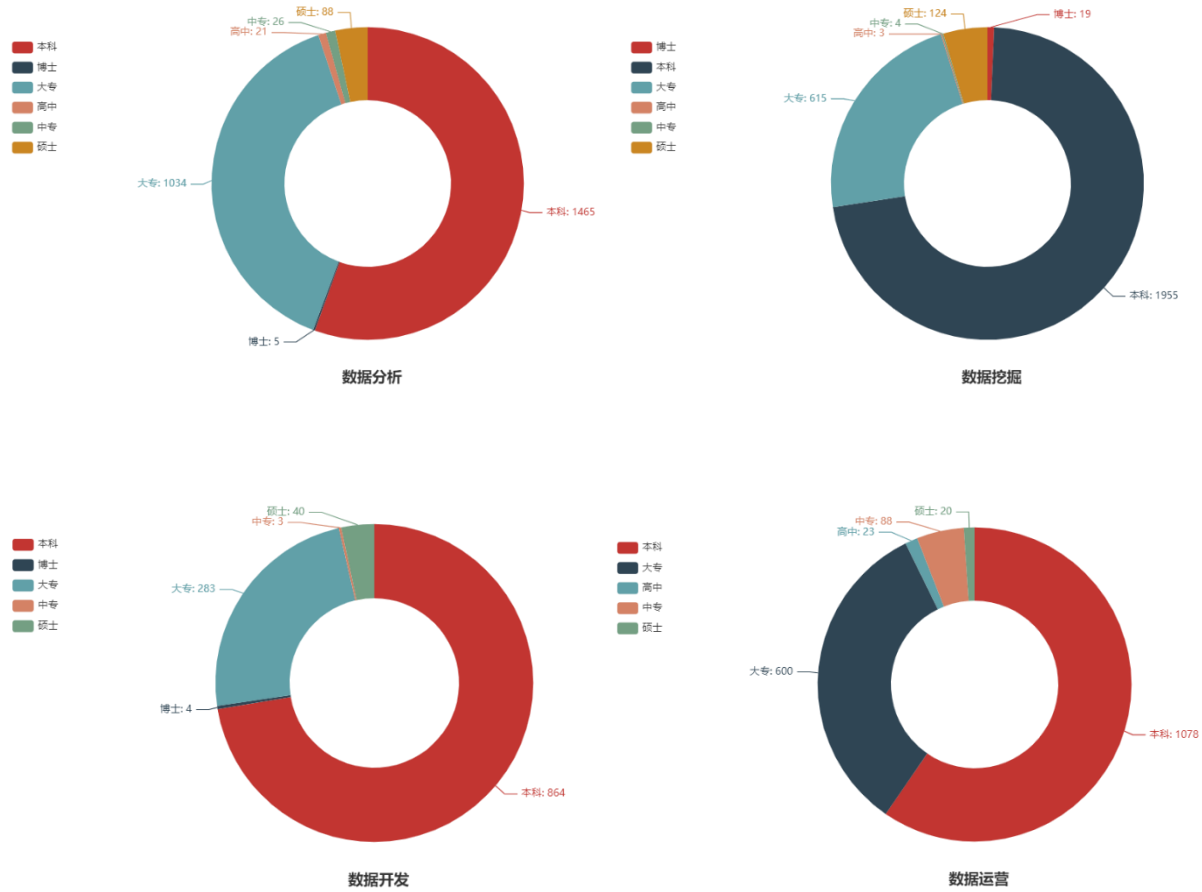

图 4.3 数据类岗位学历要求分布

4.4 岗位薪资分布

薪资是求职者较为关心的因素，本部分主要研究数据类岗位的薪资状况，主要对数据类岗位薪资与工作经验要求、学历要求的变动情况进行分析。本文对薪资字段的处理方法在数据预处理阶段已经说明，主要是将薪资不同的度量单位统一为“xx-xx 万/月”，并对薪资区间的上下限求平均值代表该区间的平均状态。

首先，对这四类岗位的薪资与工作经验要求进行说明。如图 4.4 所示，四类岗位薪资变化曲线的峰值均出现在“10 年以上经验”的岗位要求上，从“1 年经验”到“8-9 年”经验的薪资状况一直处于上升状态，说明岗位的薪资会随着工作经验年限不断增加。此外，图 4.4 中每类岗位折线图的纵轴刻度代表的是月薪，观察刻度值可以发现这四类岗位薪资状况由高到低为：数据挖掘、数据开发、数据分析、数据运营。究其原因，不难发现这种薪资走势状况与岗位所需技术性有一定关系，数据挖掘对数据分析能力和编程技术都有要求，数据开发更加具备技术岗位的特征，数据分析和数据运营在技术能力方面的要求则较低一些。

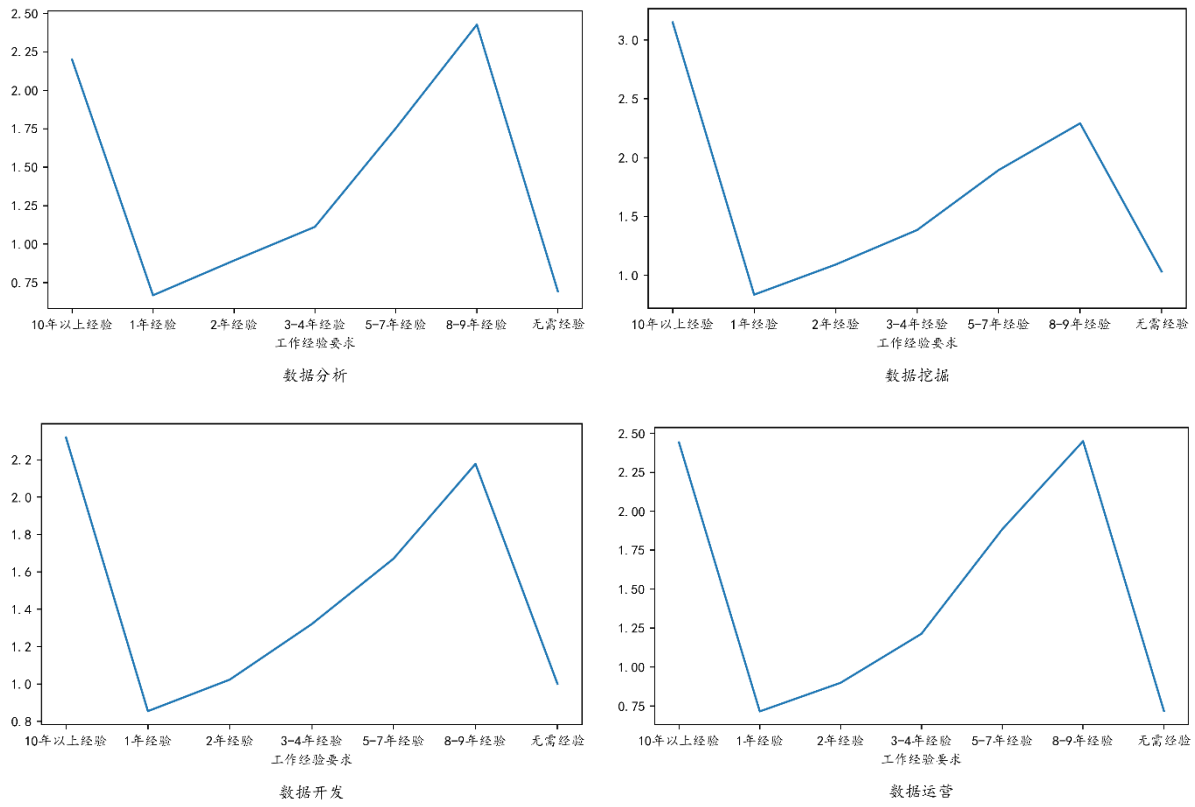

图 4.4 数据类岗位薪资与经验要求走势分析

接下来，对岗位薪资与学历要求状况进行分析，综合上述关于学历要求分布状况，可以发现数据分析岗位薪资的走势的两个极大值点对应的是博士和硕士，数据挖掘岗位薪资前两位是博士和硕士，数据开发岗位薪资前两位是博士和硕士，数据运营从大专到硕士的薪资状况是直线上升。如图 4.5 所示，不难发现，岗位薪资走势与学位高低状况是呈正向关系的，高学历要求对应岗位的薪资状况会更高些。

以本科学历为例，对比这四个数据类岗位的薪资状况，发现数据挖掘岗位的平均本科薪资是最高的，其次是数据开发和数据分析，数据运营岗位的平均本科薪资是最低的，这与我们上述分析结果保持一致。

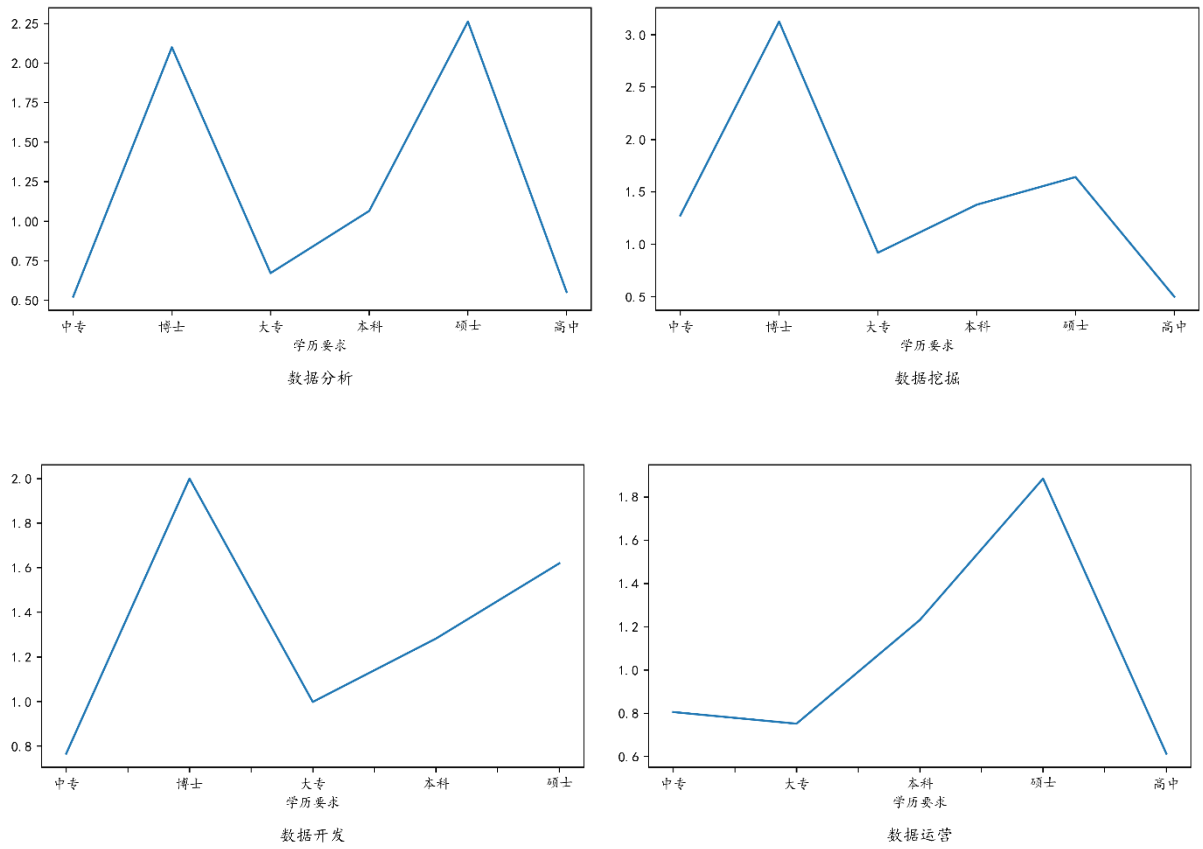

图 4.5　数据类岗位薪资与学历要求走势分析

5 基于 LDA 模型的数据类岗位招聘需求分析

5.1 基于 LDA 模型的招聘信息主题提取

通过对招聘信息的文本挖掘，有助于求职者从海量招聘信息中提取出重要信息，寻找到与自身实力相匹配的岗位，指导求职者提升个人实力。本章对清洗整理后的岗位文本数据使用 LDA 模型对招聘信息主题词进行提取，再通过 Word2Vec 识别与 LDA 模型主题词相似度较高的词汇，得到数据分析、数据挖掘、数据开发、数据运营等 4 类岗位招聘信息的关键词集。

5.1.1 数据清洗与分词

作为文本挖掘的第一步，在构建 LDA 模型之前，首先按照 2.2 节的文本预处理方法对招聘文本进行清洗与分词。原始招聘文本信息中包含大量与岗位需求无关的词语，如换行符、空格、标点符号等无效与非法字符，此处举例进行说明。

（1）“\n 岗位职责：1、监控、分析用户运营数据，针对业务问题/需求，分析产品、机构、客户等数据，建立统计模型，根据运营数据提出产品构想、策略及计划”。

（2）（空格字符）“\n 1、负责沟通、收集和分析客户再业务监管和分析方面的运营管理需求，研究用户需求，拟定产品发展策略，并不断根据策略和市场需求，进行业务分析，能够独立完成各类报表统计制作和分析任务...”。

（3）“Job Description 职位描述：这个职位的主要职责是简历主数据，根据业务需要从数据表中分析有价值的信息，为业务发展提出基于数据结论的建议...”。

（4）“1、CDH 集群维护与管理：负责 CDH 集群日常维护与管理，并且根据实际业务需要优化 Hive、kudu 等各种参数的配置；2、有 Spark 或 flink 实时或离线数据处理的经验。能够熟练使用 sql；职能类别：大数据开发工程师，关键字：数据开发数据仓库；微信分享。

（5）“岗位职责：1. 深入理解业务需求、市场挑战和客户痛点，通过数据分析和跟踪，输出整体项目进度报告；2. 与产品开发团队、数据科学家团队协作，分解工作，解决对应业务问题任职要求；1. 统招本科以上学历，2 年以上工作经验，数学、统计等相关专业优先；勤奋踏实、高度责任心及工作热情”。

显然，由于存在多样的非法字符，无法对原始的岗位需求文本直接进行分析。因此，首先通过 python 的 re 模块，根据正则表达式原理，使用 re.sub 方法对文本字符进行匹配，清洗掉文本内容中的换行符、空格等标点符号和非法字符，只保留文本中的汉字、数字和英文字母。 以数据分析岗位中某条招聘文本信息进行展示，说明文本清洗后的效果。

原始文本：[岗位职责：1、监控、分析用户运营数据，针对业务问题/需求，分析产品、机构、客户等数据，建立统计模型，根据运营数据提出产品构想、策略及计划；2、负责挖掘并分析行业的现状及需求、负责研究市场及竞品，进行分析对比，提供产品策略和运营建议；3、定期或结合重要项目输出数据分析报告，包括用户分层、成长和激励体系数据，给出明确的优化方案，提升用户体验，提升新用户转化率、老用户的活跃率和活跃度运营，降低用户转换成本；4、日常数据监控，及时反馈数据异常，发现并报告问题，对源数据进行收集、整理、存档，完善部门数据支撑平台。任职资格 1、本科及以上学历；2、数据各种数据处理软件和分析工具，有结合内容营销的数据运营能力，有电商、APP 数据分析经验优先考虑；3、有良好的沟通能力，优秀的逻辑思维能力，归纳总结能力。备注：前期属于派遣编制，优秀员工可转银行行编。\n 职能类别：数据分析师 \n 关键字：数据分析大数据处理数据挖掘统计数据分析师 sql \n 微信分享]。

清洗后的文本：[岗位职责 1 监控分析用户运营数据针对业务问题需求分析产品机构客户等数据建立统计根据运营数据提出产品构想策略及计划 2 负责挖掘并分析行业的现状及需求负责研究市场及竞品进行分析对比提供产品策略和运营建议 3 定期或结合重要项目输出分析报告包括用户分层成长和激励体系数据给出明确的优化方案提升用户体验提升新用户转化率老用户的活跃率和活跃度运营降低用户转换成本 4 日常数据监控及时反馈数据异常发现并报告问题对源数据进行收集整理存档完善部门数据支撑平台任职资格 1 本科及以上学历 2 熟悉各种数据处理软件和分析工具有结合内容营销的数据运营能力有电商 APP 数据分析经验优先考虑 3 有良好的沟通能力优秀的逻辑思维能力归纳总结能力备注前期属于派遣编制优秀员工可转银行行编职能类别数据分析师关键字数据分析大数据数据处理数据挖掘统计数据分析师 sql 微信分享]。

接下来，对于只包含中文、英文和数字的文本字符进行分词处理。首先，将表 3.1 中的专业技能词汇添加到 jieba 分词包中，再调用 jieba 分词包中的 jieba.lcut 方法对文本内容进行分词。仍对上述例子进行说明，分词后的文本内容为[‘岗位职责’，‘1’，‘监控’，‘分析’，‘用户’，‘运营’，‘数据’，‘针对’，‘业务’，‘问题’，‘需求’，‘分析’，

‘产品’，‘机构’，‘客户’，‘等’，‘数据’，‘建立’，‘统计’，...]，可以看出此时招聘文本已经被分割成若干词项。显然，分词后的文本中包括“岗位职责”和“1”等对于研究岗位需求内容意义不大的词汇，因此将诸如“岗位职责”、“问题”、“针对”、“微信分享”等与岗位信息无关的词加入停用词表中，对分词后的文本进行停用词处理。停用词处理之后的文本内容形如[‘监控’，‘分析’，‘用户’，‘运营’，‘数据’，‘业务’，‘需求’，‘分析’，‘产品’，‘客户’，...]。

最后，对于分词和停用词处理之后的文本，使用 bigram 算法将高频词汇连接为同一词汇。文本分词最终效果为：[‘监控’，‘分析’，‘用户’，‘运营’，‘数据’，‘业务’，‘需求’，‘分析’，‘产品’，‘机构’，‘客户’，‘统计’，‘策略’，‘计划’，‘挖掘’，‘指标’，‘用户分层’，‘指标体系’，‘优化’，‘转化率’，‘活跃度’，‘反馈’，‘收集整理’，‘电商’，‘数据运营’，‘APP’，‘沟通能力’，‘逻辑思维’，‘归纳能力’，‘优秀员工’，‘数据分析’，‘数据处理’，‘数据挖掘’，‘数据库’，‘统计数据’，‘sql’]。

经过文本清洗、分词和停用词处理之后的文本，从原始的一条招聘信息变成了若干个具备特定含义的词汇，含义更加明显，可供后续分析。

5.1.2 文本向量化表示

LDA 主题模型的输入主要包括分词词典与向量化表示的文本，因此需要通过分词后的每条招聘文本构建语料库，并将每条招聘文本进行文本向量化表示。通过 corpora.Dictionary 构造词典，并使用词袋模型将分词后的文本信息向量化表示。

以上述清洗后的文本为例：[‘监控’，‘分析’，‘用户’，‘运营’，‘数据’，‘业务’，‘需求’，‘分析’，‘产品’，‘机构’，‘客户’，‘统计’，‘策略’，‘计划’，‘挖掘’，‘指标’，‘用户分层’，‘指标体系’，‘优化’，‘转化率’，‘活跃度’，‘反馈’，‘收集整理’，‘电商’，‘数据运营’，‘APP’，‘沟通能力’，‘逻辑思维’，‘归纳能力’，‘优秀员工’，‘数据分析’，‘数据处理’，‘数据挖掘’，‘数据库’，‘统计数据’，‘sql’]，文本向量化效果为[(0, 1), (1, 1), (2, 1), (3, 1) (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 1), (11, 1)， (12, 1), (13, 1), (14, 1), (15, 1), (16, 1), (17, 1), (18, 1), (19, 1), (20, 1), (21, 1), (22, 1), (23, 1), (24, 1), (25, 1), (26, 1), (27, 1), (28, 1), (29, 1), (30, 1), (31, 1), (32, 1)， 33, 1), (34, 1), (35, 1), (36, 1)]。文本向量化效果表示某个词在该段招聘信息中出现了多少次，如（0，1）表示编号为 0 的词在该段文本中出现了 1 次，其余词汇同理。

通过匹配词典内容与词汇编号，可以输出单词内容与对应的出现频次。此处以数据分析和数据挖掘的某条招聘信息文本向量化表示效果举例说明。

数据分析岗位某条招聘信息文本向量化效果：[(‘APP’，1)，(‘sql’，1)，(‘业务’，1)，(‘产品’，1)，(‘监控’，2)，(‘分析’，2)，(‘策略’，1)，(‘计划’，1)，(‘用户分层’，1)，(‘沟通能力’，2)，(‘转化率’，1)，(‘逻辑思维’，1)，(‘归纳能力’，1)，(‘数据分析’，2)，(‘数据挖掘’，1)，(‘统计数据’，1)，(‘数据库’，1)，(‘指标体系’，1)，(‘统计学’，2)，(‘跟进’，1)，(‘推广’，1)，(‘数据咨询’，1)，(‘反馈’，1)，(‘业务增长’，1)，(‘本科及以上学历’，1)，(‘两年’，1)，(‘福利待遇’，1)，(‘五险一金’，1）]。

数据挖掘岗位某条招聘信息文本向量化效果：[（‘ETL’，1)，(‘PowerBI’，1)，(‘Python’，2)，(‘sql’，2)，(‘业务需求’，1)，(‘交易’，1)，(‘产品开发’，2)，(‘优化’，2)，(‘信息’，1)，(‘市场’，1)，(‘报表’，1)，(‘探索’，1)，(‘数据挖掘’，2)，(‘数据可视化’，2)，(‘浓厚兴趣’，1)，(‘海量’，1)，(‘数据清洗’，1)，(‘数据源’，1)，(‘算法’，2)，(‘统计学’，1)，(‘维护’，1)，(‘推荐系统’，1)，(‘运维’，1)，(‘NLP’，1)，(‘三年’，1)，(‘大数据’，2)，(‘下午茶’，1）]。

5.1.3 LDA 主题模型的构建

文本向量化表示完毕后，对数据类岗位招聘信息进行 LDA 主题建模。通过 gensim 文本处理库中的 gensim.models.ldamodel.LdaModel 方法构建 LDA 主题模型，主要参数是指定语料库 corpus 和主题数 num_topics。对于构建好的 LDA 主题模型，我们可以使用 pyLDAvis[42]将文档-主题分布绘制在坐标系中进行可视化展示，圆的面积表示主题在文档中的概率，圆之间的距离表示主题之间的相似度，当圆圈之间的距离比较远时，说明该主题区分较为明显；当圆圈之间距离很近，甚至发生重叠时，说明主题区分不够明显，并非最佳的主题数目。主题-词项分布通过条形图表示，条形图长度反映了词项的频率。当未选中主题时，条形图显示全部文档中的词项及其频率；当选中某一主题时（如红色圆形区域），条形图反映的是该主题的主题-词项分布。图 5.1 为数据分析岗位招聘需求主题的可视化效果。

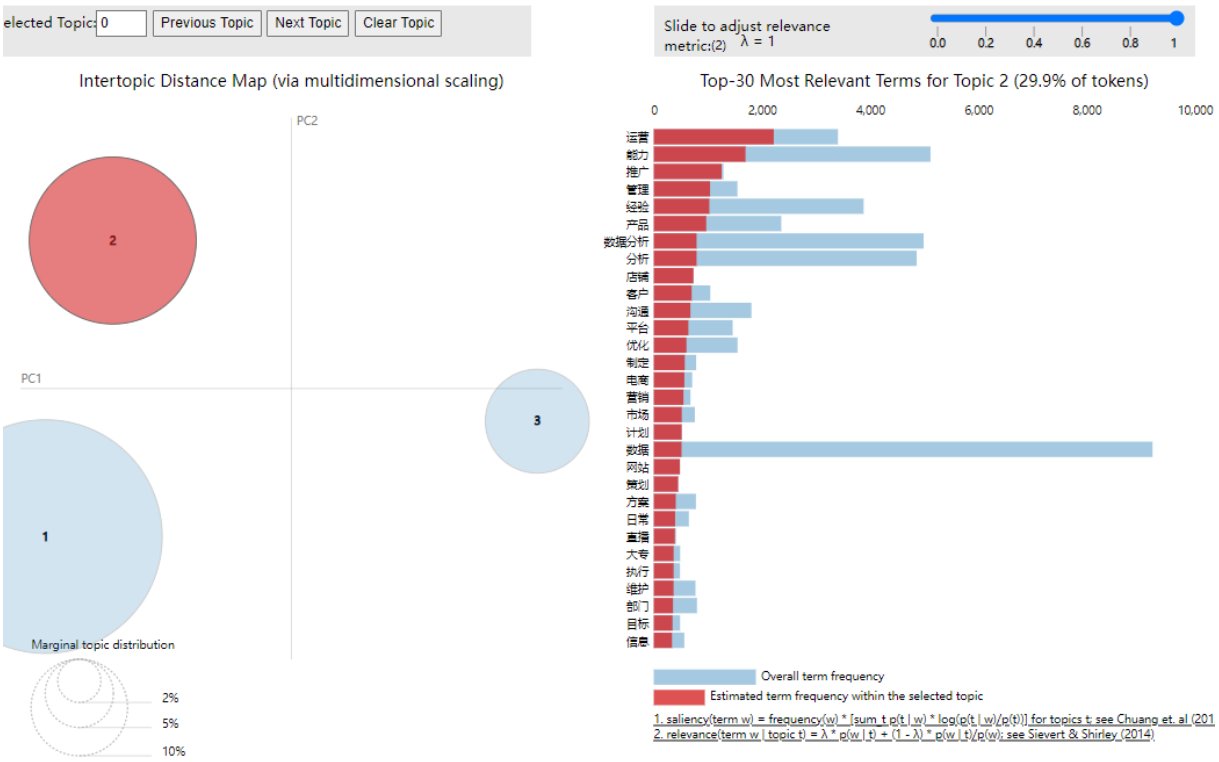

图 5.1 岗位主题模型可视化（数据分析 K=3）

LDA 主题模型抽取语料完成主题词的构建，对数据分析岗位以 K=3 建立 LDA 主题模型时，可以生成 3 个主题词，每个主题词可由若干个词语加权组成，接下来对最初提取的 3 个主题词进行说明。K=3 时，生成的第一个主题词可表示为：“0.031*‘运营’+0.024*‘能力’+0.018*‘推广’+0.015*‘管理’+0.014*‘经验’+0.014*‘产品’+0.011*‘数据分析’+0.011*‘指标体系’+0.010*‘店铺’+0.010*‘客户’；第二个主题词可表示为：“0.062*‘数据’+0.030*‘数据库’+0.029*‘运营’+0.024*‘沟通能力’+0.023*‘业务’+0.020*‘相关经验’+0.011*‘开发’+0.010*‘项目’+0.010*‘需求’+0.009*‘产品’；第三个主题词可表示为“0.017*‘福利待遇’+0.016*‘员工’+0.006*‘产品竞品’+0.006*‘假期’+0.006*‘旅游’+0.005*‘底薪’+0.005*‘外贸’+0.005*‘亚马逊’+0.005*‘电商’。

可以通过 LDA 抽取出的主题词构成情况对每类主题词有初步的认识，这些主题词具备一定的代表性。但由于没有确定较为合适的主题数，并且这些主题词是通过文档中出现频次较高的词汇的线性加总构成，在可解释性方面略有欠缺，具体每个主题内词汇分布情况及主题反映的含义，则需要在确定合适的主题数之后对主题的关键词进行分析。接下来，首先对四种岗位分别对应的最优主题数 K 进行讨论。

5.2 最优主题数的确定

LDA 主题模型的可调节参数比较少，关键参数是主题数目 K。主题数 K 的确定方法可从 2.3.3 节介绍的主题困惑度与一致性得分进行讨论，另一种确定技巧是对业务数据的理解。对数据本身良好的理解，可以帮助我们确定较优的主题数目 K。

借助主题困惑度和一致性得分的计算方法，尝试对数据分析、数据挖掘、数据开发和数据运营这四类岗位的文本招聘信息在不同主题数 K 条件下的效果进行评估，以此筛选出这四类岗位使用 LDA 主题模型时效果较优的主题数 K。具体过程如下：对于不同岗位的招聘文本数据集，使用 gensim.models.ldamodel.LdaModel 方法构建 LDA 主题模型时，构建各自岗位的语料库 corpus，主题数取值为从 1 到 10，并将 random_state 参数设为固定值使得抽取的主题词相同。对于每类岗位构建的 10 个 LDA 主题模型，分别计算各自对应的主题困惑度与一致性得分，并将不同主题数 K 对应的一致性得分波动情况进行可视化展示，见表 5.1 与图 5.2。

表 5.1 不同主题数 K 对各数据类岗位的影响

岗位	主题数 K	主题困惑度	一致性得分	岗位	主题数 K	主题困惑度	一致性得分
数据分析	1	-7.4009	0.4671	数据开发	1	-6.8452	0.5439
	2	-7.3167	0.4361		2	-6.7648	0.5352
	3	-7.2369	0.4438		3	-6.7645	0.5324
	4	-7.3203	0.4821		4	-6.7688	0.5254
	5	-7.2547	0.4941		5	-6.7780	0.5315
	6	-7.2919	0.4858		6	-6.7917	0.5323
	7	-7.3432	0.4671		7	-6.8123	0.5275
	8	-7.3996	0.4715		8	-6.8734	0.533
	9	-7.5321	0.4566		9	-6.9875	0.5304
	10	-7.7338	0.4560		10	-7.1648	0.5322
数据挖掘	1	-7.2371	0.4817	数据运营	1	-7.4876	0.4783
	2	-7.0648	0.4926		2	-6.9606	0.4861
	3	-6.9836	0.4965		3	-6.9295	0.4971
	4	-6.9755	0.5049		4	-6.8812	0.5034
	5	-6.9652	0.4792		5	-6.8894	0.4760
	6	-6.9854	0.4787		6	-6.8807	0.4990
	7	-7.0520	0.4698		7	-6.9648	0.4894
	8	-7.1278	0.4900		8	-7.0160	0.4781
	9	-7.2479	0.4880		9	-7.1355	0.4826
	10	-7.4495	0.4830		10	-7.2931	0.4777

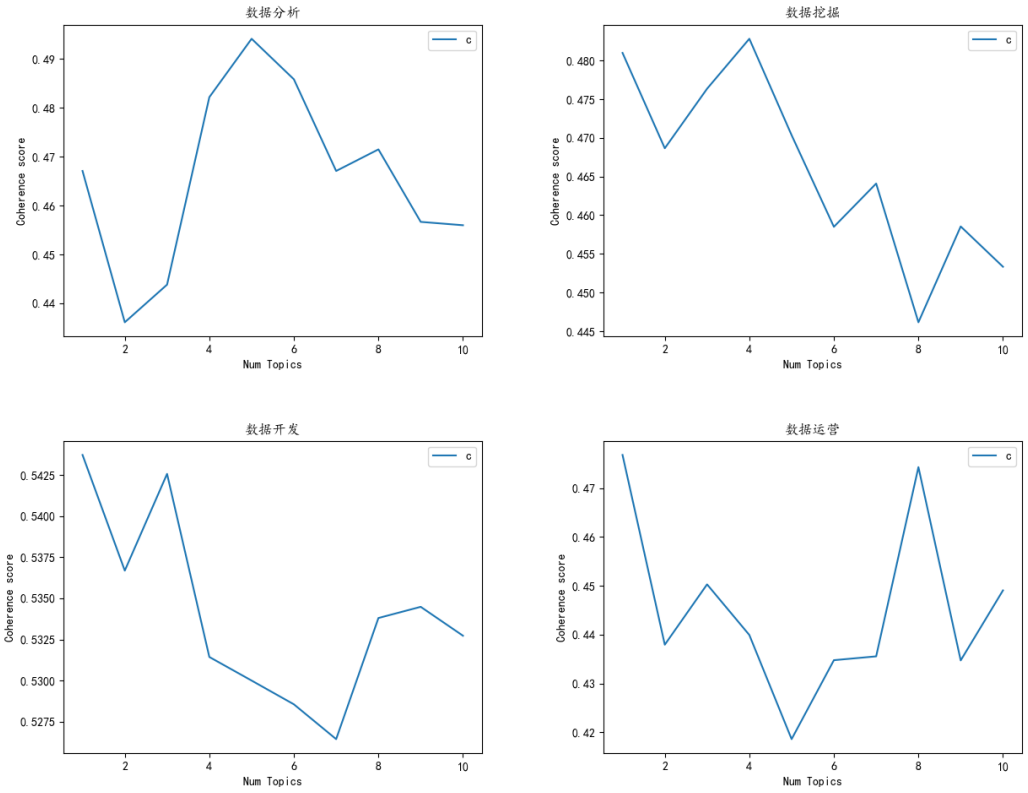

图 5.2 不同主题数 K 对应主题模型一致性得分

主题困惑度取值越小，一致性得分取值越大时，表明该主题数 K 对应的 LDA 模型对文本主题的挖掘效果更好。基于这两项指标的特点，观察图 5.8 中一致性得分走势，不难发现，这四类岗位所对应的最优的主题数取值范围都集中在 3-5 之间。其中，“数据分析”的一致性得分对应的极大值点在 K=5 处；“数据挖掘”的一致性得分对应的极大值点在 K=4 处；“数据开发”的一致性得分对应的极大值点在 K=3 处，数据运营的一致性得分对应的极大值点在 K=3 处。

此处我们以“数据分析”岗位信息进行说明，探讨最优主题数 K 的选取效果与确定方法。已知图 5.1 为 K=3 时数据分析岗位主题可视化效果，下图 5.3 与 5.4 分别为 K=4 与 K=5 时的主题可视化效果。将 K=3、4、5 时对应的可视化效果与可解释效果进行对比，筛选并确定最优的主题数 K。

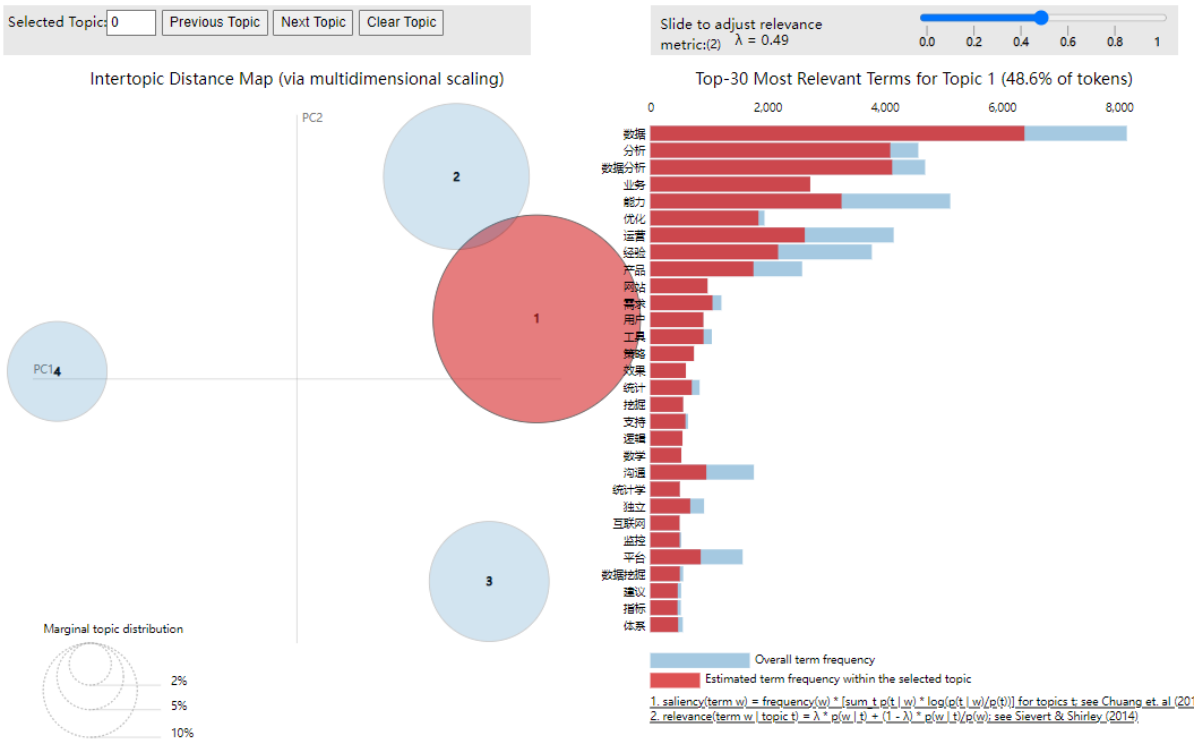

图 5.3 岗位主题模型可视化（数据分析 K=4）

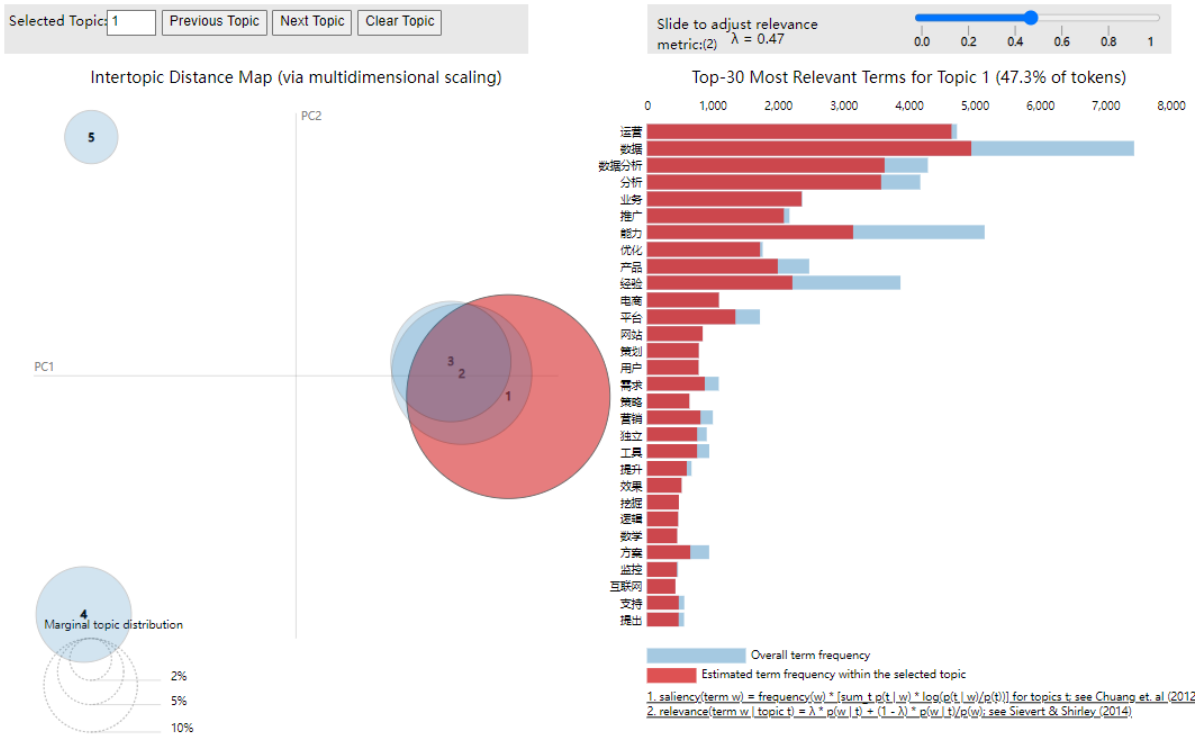

图 5.4 岗位主题模型可视化（数据分析 K=5）

不难发现，虽然“数据分析”岗位当主题数 K=5 时一致性得分最高，但主题 1、2、3 发生重叠，说明 K=5 时主题区分效果不佳。当 K=4 时，主题划分效果较为明确，并且

主题并未出现明显的重叠情况，因此选择 K=4 作为“数据分析”岗位需求文本信息的最佳主题数。

同理，对其他三种岗位使用同样方法进行对比，确定这四类岗位对应的最优主题数如下：数据分析（K=4）、数据挖掘（K=4）、数据开发（K=3）、数据运营（K=3）。

接下来，对数据分析、数据挖掘、数据开发、数据运营这四类岗位招聘需求的文本主题分别进行描述与分析。在 jieba 分词效果的基础上，对 LDA 主题模型结果中偶尔出现的“擅长”、“使用”等对于描述岗位职责内容意义不大的词语进行人工剔除，并分别对各岗位提取出的主题词进行整理，对岗位中每个需求主题选取 11 个主题词进行说明，见表 5.2。

表 5.2 数据类职位招聘需求主题词分布

岗位	主题	关键词
数据分析	1	数据分析 业务 优化 经验 数据敏感 策略 逻辑 沟通 统计学 互联网 指标体系
	2	店铺 管理 运营 客户 商品 营销 电商 文案 跟进 维护 推广
	3	数据库 sql python 数据仓库 SAS R Orcale 编程 Hive 机器学习 SPSS
	4	福利待遇 旅游 补贴 提成 薪酬 奖金 享受 节假日 五险一金 下午茶 团建
数据挖掘	1	大数据 数据分析 数据挖掘 优化 方案 平台 体系 软件 需求 沟通 统计
	2	算法 技术 工程师 机器学习 数据库 数据仓库 建模 架构 实施 编写 深度学习
	3	服务 保障 迭代 流程 管理 维护 方案 硕士及以上学历 模型 推荐 NLP
	4	客户 推广 平台 网络 策划 拓展 体验 产品 协助 开发 预算
数据开发	1	开发 分析 技术 平台 数据仓库 ETL Hadoop Hive Spark Java 架构
	2	数据库 维护 系统 集群 Oracle DBA MySQL 脚本 Linux 配置 运维
	3	福利 员工 假期 补贴 五险一金 周末双休 团建 年终奖 法定节假日 薪酬 机会
数据运营	1	运营 产品 经验 业务 需求 沟通 管理 市场 策略 方案 提升
	2	管理 招商 审核 收集 客户 运维 门店 区域 计划 商圈 标注
	3	推广 品牌 电商 用户增长 客户 策划 资源 清晰 市场营销 数据敏感 竞品分析

5.3 基于 Word2Vec 模型的主题词扩展

在使用 LDA 主题模型计算出每类主题词的基础上，本部分尝试使用 Word2Vec 算法对主题词进行扩展。Word2Vec 模型可以理解为基于神经网络的词向量模型，使用目标词汇的上下文训练词向量，因此 Word2Vec 中数值形式的词向量包含着一定程度的语义信息，可通过计算向量间的余弦值来表示词项之间的关联程度，对传统的词袋模型存在的问题进行改善，并且保留优良的运行速度。

使用 LDA 主题模型求得的数据类岗位的招聘需求主题词中，部分主题提取的主题词不够详细，如“经验”等词对于描述工作经验情况略显不足，没有将“一年”、“两年”等词进行关联；对于技能要求的主题词中，如“机器学习”、“深度学习”等词对于该领域更为细节的专业词汇没有涉及。为了能够提取出更为详细的关键词，对岗位需求进一步分析，使用 Word2Vec 模型识别与主题词相似程度较高的词。

本文采用 Python 加载 Gensim 库实现 Word2Vec 模型，通过 Skip-gram 结构训练各数据类岗位招聘数据的词向量，计算招聘文本信息中词项与 LDA 主题词的相似度，进而识别出基于需求主题的招聘关键词，对 LDA 主题模型的关键词进行扩展。本文以数据挖掘岗位为例，对招聘关键词的部分新词识别结果进行展示，如表 5.3 所示。通过 Word2Vec 模型能够更加准确识别岗位需求的关键词，并且通过补充词汇的汇总和归纳，有利于后续岗位特点的呈现。

表 5.3 LDA 初始主题词扩展（数据挖掘）

初始主题词	扩展主题词与初始主题词的相似度			
算法	推荐算法 0.98	排序算法 0.96	召回率 0.95	分类 0.94
技术	大数据 0.99	计算机视觉 0.97	分布式计算 0.96	流数据 0.95
软件	Java 0.99	Python 0.99	C++ 0.97	Matlab 0.94
工程师	数据安全 0.99	CV 0.99	NLP 0.98	信息安全 0.97
数据仓库	ETL 0.99	Hadoop 0.98	Hive 0.97	Oracle 0.96
机器学习	分类 0.99	回归 0.99	聚类 0.97	决策树 0.96
平台	云计算 0.99	CDH 0.99	ODBC 0.98	BI 0.97
架构	流式架构 0.99	Kappa 0.99	离线计算 0.97	实时计算 0.96
深度学习	LSTM 0.99	Bert 0.99	目标检测 0.98	CNN 0.97

5.4 数据类岗位招聘主题词分析

5.4.1 数据分析岗位

对数据类四种岗位的招聘信息使用 LDA 模型得到各主题词分布，并基于 Word2Vec 模型对主题词进行扩展，并对扩展后的主题词进行进一步的分析与可视化。“数据分析”岗位共提取出 4 类主题词，第一类主要内容包括“数据分析”和“业务”、“数据敏感”、“指标体系”；第二类主要内容包括“店铺”、“管理”、“电商”、“营销”；第三类主要包括“数据库”、“sql”、“python”和“Hive”；第四类主要包括“福利待遇”、“补贴”、“下午茶”等等。数据分析岗位词云如图 5.5 所示。

图 5.5 岗位需求主题词词云（数据分析）

对数据分析岗位的主题词进行深层分析，能够发现第一类主题词主要描述“数据分析”相关工作的主要内容和所需技能，如业务理解能力、沟通能力、逻辑思维能力和数据敏感性等都是一名数据分析师应该具备的基本素养。第二类主题词主要与电子商务行业相关，表明电商行业有关数据分析的岗位就业缺口较大，更需要数据类相关人才。具体分析出现这种情况的部分原因是本次采集数据中涉及上海、杭州的数据分析岗位比较多，而江浙沪地区又是电商行业发展较早且较为成熟的区域，因此反映出数据分析岗位在电商行业的需求较大。显然，第三类主题词描述的是数据分析师应掌握的工作技能要

求，如使用数据库 sql、Oracle、Hive 的能力，使用编程语言 R、python 的能力。此外，数据分析师岗位对传统的统计软件 SAS、SPSS 等操作也有要求，对更高级的数据分析师的要求包括掌握数据仓库、Hive 的使用等等。第四类关键词包括岗位的福利待遇等信息，与数据分析师技能要求没有直接的关系。

5.4.2 数据挖掘岗位

对数据挖掘岗位的主题词进行深层分析，发现第一类主题词包括“大数据”、“数据分析”、“数据挖掘”等描述岗位背景的词项；第二类主题词包括“算法”、“技术”、“深度学习”等描述岗位技能的词项；第三类包括“推荐算法”、“NLP”、“CV”、“硕士及以上学历”等有关岗位学历要求及职位信息等更具技术性词项；第四类包括“数据仓库”、“数据平台”、“协助”、“开发”等词项，这类词都与大数据平台的设计与使用有关。数据挖掘岗位词云如图 5.6 所示。

图 5.6 岗位需求主题词词云（数据挖掘）

通过二者对比发现，相较于数据分析，数据挖掘更像是一种工程师的岗位角色，对求职者在数据技术领域提出更高的要求。相较于传统的统计分析，数据挖掘任务面向的是实时产生海量数据的大数据平台，数据更新方式更快、数据存量更大，因此需要具备更高的计算机水平才能胜任岗位。近些年，机器学习与深度学习算法发展较为迅速，数

据挖掘工程师大多需要具备机器学习与深度学习算法的相关知识，在建模方面需要具备较强的动手能力，而深度学习的应用领域主要面向自然语言处理（NLP）、推荐系统、计算机视觉和数据安全等场景。

5.4.3 数据开发岗位

对数据开发岗位的主题词进行深层分析，可以发现数据开发岗位可以归为真正意义上的技术岗位。第一类主题词包括“开发”、“ETL”、“Hadoop”、“Hive”、“Spark”等描述数据开发岗位必备的技能与应用平台；第二类主题词包括“数据库”、“维护”、“集群”、“配置”等数据开发岗位主要的工作职责，主要是负责大数据处理框架的研发工作，设计并开发分布式存储、数据分析的架构等等。第三类主题词包括岗位的福利待遇等信息，这类主题词内容与数据分析岗位的第四主题词类似，都是对岗位福利待遇等信息进行说明，对于分析岗位需求价值不大。数据开发岗位需求词云如图 5.7 所示。

图 5.7 岗位需求主题词词云（数据开发）

对比上述三类岗位主题词，不难发现，数据开发岗位主要为数据类工作提供运行环境与平台，包括大数据处理框架和数据仓库的建设。特别的，数据开发岗位对求职者的数据库技术提出更为明确的要求，需要求职者熟悉 Oracle、MySQL 等数据库，能够熟练掌握 sql，并且希望求职者能够掌握 sql 调优等能力。此外，数据开发岗位需要求职者

熟悉 Hadoop 生态系统, 如 HDFS、Hive、MapReduce、Hbase 等等。

5.4.4 数据运营岗位

对数据运营岗位的主题词进行深层分析，发现第一类主题词包括“产品”、“业务”、“EXCEL”、“数据报表”、“BI”、“PPT”等，描述的数据运营岗位需要的基本技能；第二类主题词包括“管理”、“收集”、“数据整理”、“标注”、“监控”等，描述的是数据运营岗位的工作内容；第三类主题词包括“推广”、“策划”、“数据敏感”、“宣传”、“用户增长”等，明确数据运营岗位的业务场景。数据运营岗位需求词云如图 5.8 所示。

图 5.8 岗位需求主题词词云（数据开发）

不难发现，相较于数据开发岗位，数据运营岗位对数据技术层面要求较低，与业务贴合程度更高，需要求职者以数据为切入点，驱动业务增长。数据运营岗位从业者主要通过数据层面理解业务，通过监控业务数据指标体系，将数据与业务紧密联系，实现数据价值与商业赋能。数据运营岗位是通过数据技术手段，收集并归纳总结各种数据，整理数据中所蕴涵的信息，需要负责监控并分析客户的业务、运营、产品等相关数据，并对业务发展提出可行建议，运营与用户、产品、增长等概念紧密联系。

5.4.5 数据类岗位对比

通过分别对数据分析、数据挖掘、数据开发、数据运营岗位需求的关键词进行描述，能够对四类岗位的、工作内容、主要职责、技能需求等有一定程度的认识，本部分对四类岗位的侧重点进行对比，对四类岗位的异同点进行阐述。从工作内容的导向进行对比，可以将岗位的主要工作内容分为指引业务增长和数据技术建设与保障两部分。

以指引业务增长为导向来看，数据分析与数据运营岗位的主要职责属于此范围，这两种岗位以数据为媒介，以分析方法为途径，以业务增长为目的，主要侧重于通过分析数据实现业务增长，使用数据为业务赋能，这点在电商行业尤为凸显。二者相比较而言，数据分析岗位需要了解业务知识，但更加侧重理解数据分析方法，从数据库中提取并分析数据，并对业务提供基于数据层面的意见与建议；而数据运营更加贴合业务知识，对于数据分析方法不需要了解太深，更加侧重业务知识与商业技能的掌握，如产品运营、店铺管理、竞品分析、用户增长等等实际工作的业务能力。涉及数据分析软件的使用时，数据分析岗位需要掌握“sql”、“Python”、“Hive”等数据库及数据分析软件；数据运营岗位则需要掌握“EXCEL”、“PowerBI”、“Tableau”等表格处理、报表制作与数据可视化等软件即可，能够使用数据库技术更佳。

以数据技术建设与保障为导向来看，数据挖掘与数据开发岗位更加侧重于此。数据挖掘岗位侧重于从海量数据中提取有价值的信息，主要涉及数据挖掘算法的设计，并需要熟悉数据平台，辅助数据开发岗位运维数据平台。数据挖掘需要对计算架构有一定程度的了解，掌握主流的数据挖掘算法理论知识，主要围绕自然语言处理、推荐系统、计算机视觉等应用领域，技能需求方面需要掌握精通一门编程语言，如 Python、Matlab、Java、C++。数据开发岗位更加侧重于数据计算架构与数据平台的开发，为其他的数据类岗位提供计算环境与平台，可归纳为技术岗位。需要熟悉主流的数据平台搭建与计算方法，主要分为实时计算与离线计算，还需要精通数据仓库的开发与使用，如 Hadoop、MongoDB、Oracle、CDH 等等，能够熟练使用 Shell 更佳。

从对业务理解的需求程度来看，由高到低依次为：数据运营、数据分析、数据挖掘、数据开发；从对技术层面的需求程度来看，由高到低依次为：数据开发、数据挖掘、数据分析、数据运营。

6 结论建议与展望

6.1 结论

本文通过网络爬虫技术获取了前程无忧招聘平台中数据类岗位的招聘信息，对国内数据类相关工作招聘市场的现状、四种主要的数据类相关岗位基本情况与差异性进行对比分析。

对本文所采集的“数据分析”、“数据挖掘”、“数据开发”、“数据运营”四类相关岗位信息进行挖掘，得出以下结论：从岗位需求的城市分布来看，数据类岗位缺口在我国东部和南部地区较大，华东地区整体上对四种数据类岗位的需求都较大，特别是数据分析和数据运营岗位，可能与上海、杭州等地发展电商行业起步较早有关，华南地区对数据挖掘和数据开发岗位需求较大，主要以广州、深圳为代表，作为国家首都的北京，是我国北部城市中对数据类岗位需求最大的城市，其他地区如成都、长沙、武汉和西安对于数据类岗位也有较大需求；从岗位经验要求来看，至少需要 1 年以上工作经验的比例在数据分析、数据挖掘、数据开发、数据运营岗位的比例分别为 91.55%、90.92%、93.80%和 87.51%，可见数据类岗位对求职者的工作经验要求较高，数据类工作更加看重工作经验；从岗位学历要求分布来看，本科学历在数据分析、数据挖掘、数据开发、数据运营岗位的比例分别为 59.03%、77.17%、76.05%和 60.7%，可见本科学历是从事数据类岗位学历的敲门砖，且对于数据挖掘和数据开发岗位而言不具备竞争优势，数据运营岗位中没有出现学历要求为博士的岗位；从岗位薪资与工作经验要求的走势分析来看，参加相关工作的经验时长与薪资呈明显的正向增长关系，且在“3-4 年经验”、“5-7 年经验”、“8-9 年经验”、“10 年以上经验”等节点处增速明显；从岗位薪资与学历要求来看，高学历要求岗位薪资往往更高，且只考虑学历因素时，从本科学历到硕士学历涨幅明显。

在对数据类岗位基本信息分析的基础上，重点对岗位要求的文本数据进行挖掘与分析，得到四种数据类岗位的技能侧重点与需求差异。首先使用 LDA 主题模型对四种岗位的招聘信息文本进行主题词提取，基于主题困惑度、一致性得分以及 LDA 主题模型可视化效果确定四种岗位最优的主题数目，分别为 K=4（数据分析、数据挖掘）与 K=3（数据开发、数据运营），并整理各主题内出现频次较高的关键词，发现包括描述岗位职责、职位场景、主要所需技能、福利待遇等主题内容，以数据分析岗位为例，分别为“数据敏感”“指标体系”、“店铺”、“电商”、“数据库”及“福利待遇”等主题词。在 LDA

主题模型提取主题词的基础上，使用 Word2Vec 模型对主题词进行深层挖掘，将与各主题词相关程度较高的词项提取出来，并对四种岗位最终关键词中的高频词汇绘制词云。通过词云可视化过程，发现如下结论：从岗位需求的侧重点来看，数据分析岗位强调指标体系的搭建与数据处理能力，能够熟练使用 sql 语句等数据库技术，并将数据处理技术与业务结合起来；数据挖掘岗位更加注重数据分析算法的设计，并且对大数据平台的使用能力要求更高；数据开发岗位强调大数据平台与架构的研发能力，对数据库及海量数据处理方式的要求更高，对业务知识要求较低；数据运营岗位对业务要求知识较高，通过使用 BI 报表等平台与产品、运营结合起来，将数据信息与业务增长紧密结合，实现通过数据为商业赋能。从对技术需求的层面来看，数据挖掘和数据开发岗位要求更高，数据挖掘主要应用在计算机视觉、自然语言处理、推荐系统等等；从对业务理解的层面来看，数据运营和数据分析岗位的要求更高，并在电商行业应用范围较广。

6.2 建议

大数据时代为数据类人才提供了较为良好的就业机会，数据类岗位近些年可谓炙手可热，市场上存在大量需求。尽管数据类专业毕业生人数逐年增多，但是还是会出现企业找不到优秀人才、人才难以寻求到满意岗位的情况。以下基于本文研究内容与结果，对求职者提出一些建议。

对于立志从事于数据类岗位的求职者来说，可以注意关注以下几个方面[43-47]：

（1）明确想要从事的数据类职位类型。在本文中，对数据分析、数据挖掘、数据开发、数据运营类岗位的岗位信息与任职要求等进行分析，可以将数据分析和数据运营岗位归于业务型数据类人才，将数据挖掘和数据开发归于技术型数据类人才。虽然都和数据相关，但岗位侧重点是不同的，求职者应该在工作之前充分了解自己的专业背景和技能优势，明确自己更想成为哪种数据类人才。

（2）提升自身实际动手能力和软实力。在主题模型提取出的关键词中，发现数据类岗位都对求职者的数据分析能力提出一定要求，不过差异在于使用软件的不同，从 EXCEL 到 sql、R、Python 等等，只有通过平常实际的动手操作才能提升这些数据分析工具的使用能力。其中，数据库技术是这些岗位要求中都提到的，主要包括 MySQL、Hive 和 Hadoop 等等，求职者应在日常提升数据库的使用水平，毕竟实际工作中的数据往往不能直接被我们所用，需要进一步的检索和整理。

（3）积极参加实习，在数据工作的实践中提升数据处理技能，积累业务知识。通过分词岗位需求，不难发现，诸多岗位对求职者的工作经验要求较高，可以说数据类工作本身就是侧重实践的岗位，只有通过大量实际工作积累才能提升实际的动手能力，因此在校生需要积极参加实习，尽早接触数据类工作的实际内容。在岗位地区选择上，可以考虑华东地区的上海、杭州、南京、宁波、苏州等城市，实习工作机会较多，武汉、成都等城市也值得一试。

6.3 展望

本文在写作过程中还存在诸多不足，在未来的研究中值得进一步补充修正：

（1）数据获取的是 10 月份数据，属于秋招末期，获取的数据量较少，不能很充分的把握企业对应届生的岗位要求，所以在以后的研究中爬取数据时可以选择春招或秋招进行的关键节点，这样获取到的岗位信息会更加充分。除了对正式岗位招聘信息的研究，还应通过某些数据标签将实习岗位提取出来，有针对性的对数据类的实习岗位需求进行挖掘研究。此外，获取数据只通过前程无忧招聘平台，样本范围不够丰富，在以后的研究中可以考虑爬取的招聘网站更加多元化。

（2）在对岗位数据清洗时，所采取的方法是对某些关键字段缺失的数据直接删除，这种方法虽然可操作性强，但是会删除掉部分有价值的信息，使得最终分析使用的数据量与最初爬虫得到数据量差距较大。在后续研究中，应当提升数据清洗和数据规范化的技能，更大程度利用有价值的信息。

（3）专业技能词典的内容可以更加完善。各数据类岗位所对应的软件及技能名称不尽相同，特别是数据挖掘和数据开发岗位涉及到的模型、算法与数据平台等术语内容，由于更新较快且专业性较强，造成本次文本分析涉及到的分词与停用词处理环节不够完善。在以后的工作学习中，应当留意对数据类岗位技能名词的积累，构建更加完备的数据专业技能词典，以期在将来相关研究中取得更好的效果。

参考文献

[1]FELDMAN R, DAGAN I. Knowledge Discovery in Textual Databases (KDT)[J]. Proc of Kdd, 1995.

[2]BEIL F, ESTER M, XU X. Frequent term-based text clustering[J]. Proceedings of Int.conf.on Knowledge Discovery & Data Mining, 2002.

[3]STEVENS K, KEGELMEYER P, ANDRZEJEWSKI D, et al. Exploring Topic Coherence over many models and many topics: Conference on Empirical Methods in Natural Language Processing, 2012[C].

[4]GOSWAMI S, SHISHODIA M S. A fuzzy based approach to stylometric analysis of blogger's age and gender: 2012 12th International Conference on Hybrid Intelligent Systems (HIS), 2013[C].

[5]HU K, LIU H, HAO T. A Knowledge Selective Adversarial Network for Link Prediction in Knowledge Graph, 2019[C].

[6]杨亚楠, 赵文辉, 张健, 等. 基于多视图协同的政策文本可视化研究[J]. 数据分析与知识发现, 2019,3(06):30-41.

[7]沈健, 胡洁, 马进, 等. 基于文本挖掘的生物领域实例获取[J]. 上海交通大学学报, 2018,52(08):954-960.

[8]沈艳, 陈赟, 黄卓. 文本大数据分析在经济学和金融学中的应用:一个文献综述[J]. 经济学（季刊）, 2019,18(4):1153-1186.

[9]倪志恒. 基于文本挖掘的虚假评论识别[D]. 兰州财经大学, 2021.

[10]戴德宝, 兰玉森, 范体军, 等. 基于文本挖掘和机器学习的股指预测与决策研究[J]. 中国软科学, 2019(04):166-175.

[11]李伦珑. 基于LDA模型的中国古典诗词在不同历史时期的主题发现[D]. 兰州财经大学, 2021.

[12]杨文清. 基于文本挖掘的投资者情绪对股票收益率的影响[D]. 哈尔滨工业大学, 2021.

[13]李梦杰, 刘建国, 郭强, 等. 基于文本挖掘的互联网教育课程主题发现与聚类研究[J]. 上海理工大学学报, 2018,40(03):259-266.

[14]谭春辉, 熊梦媛. 基于LDA模型的国内外数据挖掘研究热点主题演化对比分析[J].

情报科学, 2021,39(4):174-185.

[15]周云泽, 闵超. 基于LDA模型与共享语义空间的新兴技术识别——以自动驾驶汽车为例[J]. 数据分析与知识发现,2021:1-16.

[16]黄琳, 王丽亚, 明新国. 基于改进的LDA模型的产品服务需求识别[J]. 工业工程与管理, 2022:1-14.

[17]SHENOY V, AITHAL S. Literature Review on Primary Organizational Recruitment Sources[J]. Social Science Electronic Publishing, 2018.

[18]AKEN A, LITECKY C, AHMAD A, et al. Mining for Computing Jobs[J]. IEEE Software, 2010,27(1):78-85.

[19]TURRELL A, SPEIGNER B, DJUMALIEVA J, et al. Transforming Naturally Occurring Text Data into Economic Statistics: The Case of Online Job Vacancy Postings[J]. Social Science Electronic Publishing, 2019.

[20]SODHI M S, SON B. Content Analysis of O.R. Job Advertisements to Infer Required Skills[J]. SSRN Electronic Journal, 2010.

[21]俞琰, 陈磊, 姜金德, 等. 融合论文关键词知识的专利术语抽取方法[J]. 图书情报工作, 2020,64(14):104-111.

[22]韩春光, 许艳丽. 高职学生就业质量因素影响调查分析——基于北京地区数据[J]. 中国职业技术教育, 2018(33):56-64.

[23]杨文泽. 网络招聘薪资对劳动力流动的影响[D]. 河北大学, 2019.

[24]詹川. 基于文本挖掘的专业人才技能需求分析 ——以电子商务专业为例[J]. 图书馆论坛, 2017,37(5):116-123.

[25]张俊峰. 国内网站招聘岗位需求特征挖掘及其应用研究[D]. 安徽财经大学, 2017.

[26]刘睿伦, 叶文豪, 高瑞卿, 等. 基于大数据岗位需求的文本聚类研究[J]. 数据知识与分析发现, 2017.

[27]谭云鹤. 基于招聘网站数据处理类岗位的人才需求分析[D]. 天津财经大学, 2019.

[28]刘畅. 数据类岗位招聘需求信息研究[D]. 兰州财经大学, 2019.

[29]朱爱璐. 基于文本挖掘的数据分析岗位人才需求分析[D]. 江西财经大学, 2020.

[30]黄崑, 王凯飞, 王珊珊, 等. 数据类岗位招聘需求调查及对图情学科人才培养的启示[J]. 图书情报知识, 2016(6):42-53.

[31]杨静. 基于文本挖掘的网络招聘信息分析[D]. 山东师范大学, 2019.

[32]郑思雨. 网络招聘信息的数据挖掘研究[D]. 杭州电子科技大学, 2020.

[33]郭欢欢. 基于大数据方法的精准招聘研究[D]. 北京工业大学, 2020.

[34]吴汉龙, 梁嘉鹏, 余泽汇. 基于特征评分算法的网络招聘信息分析与研究[J]. 2021.

[35]李寿清. 基于机器学习的网络招聘薪资影响因素研究[D]. 长江大学, 2020.

[36]俞琰, 陈磊, 赵乃瑄. 基于网络招聘文本挖掘的课程知识模型自动构建研究[J]. 图书情报工作, 2019,63(10):134-142.

[37]MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. 2013.

[38]赵凯, 王鸿源. LDA最优主题数选取方法研究:以CNKI文献为例[J]. 统计与决策, 2020,36(16):175-179.

[39]BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation[J]. The Annals of Applied Statistics, 2001.

[40]HOFFMAN M D, BLEI D M, BACH F R. Online Learning for Latent Dirichlet Allocation: Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada., 2010[C].

[41]唐家渝, 刘知远, 孙茂松. 文本可视化研究综述[J]. 2013,25(3).

[42]SIEVERT C, SHIRLEY K E. LDAvis: A method for visualizing and interpreting topics: Workshop on Interactive Language Learning, Visualization, and Interfaces at the Association for Computational Linguistics., 2014[C].

[43]涛张, 黄海军. 大数据背景下应用统计专业硕士人才培养模式研究[J]. 教育教学论坛, 2017(29):178-179.

[44]阮敬. 大数据背景下的应用统计专业硕士人才培养模式研究[J]. 统计与管理, 2015.

[45]聂淑媛. 数据科学的发展与人才培养研究[J]. 统计与信息论坛, 2019,34(1):117-122.

[46]侯锡林, 李天柱, 马佳, 等. 大数据分析师的能力分析及其复合培养模式研究[J]. 高等工程教育研究, 2017(3):149-153.

[47]陈振冲, 贺田田. 数据科学人才的需求与培养[J]. 大数据, 2016,2(5):95-106.

后 记

写下这篇后记的时候，雨夜过后的杭州城外一副郁郁葱葱的景象，生机勃勃，而我硕士阶段的学习生活就要步入尾声。青春是一本太仓促的书，求学时光最终化为几万字的硕士论文和对生活的无限感恩。

首先，感谢我的导师黄恒君教授，是他带我进入数据挖掘的大门。他严谨的治学态度、深刻的思想见解和严格的学术要求深深影响着我，在论文写作的过程中，老师对我进行了无私的指导和帮助。学生稚气率性，难免遇挫，老师总将自己的人生思考和处世之道倾囊相授。师恩永铭，学生唯有勤学善思、敬业奉献来回报老师的谆谆教诲。

其次，我能够顺利完成论文，离不开家人的大力支持。小时候不懂事，总是对家人的高标准和严要求心怀不满，却不得不照做，长大后才觉得这是一笔宝贵的财富。家人在我身后默默的支持与鼓励，让我不必回头、一直向前，祝愿你们身体健康，万事顺遂。祝愿侄女“小苹果”能够快乐成长，好好读书，叔叔也将是你未来成长坚强的后盾。

此外，我要感谢在兰州读书时的同窗，虽然大家已奔赴国内各地，但共同度过的美好时光将深深烙印在我的脑海中。感谢师姐杨梦玲、师弟师妹们对我学习上的帮助；感谢同门杨帆、杨丽娜、景鹏志、张俞茜，一起上讨论班的日子很开心；感谢朱立群、杜文豪、杨越、曹义来、戴加州等男子汉们给我的鼓励和帮助；尤其要感谢倪志恒和李宪慧、王帅和张瑞萍，没有他们的帮助和鼓励，我的求学生涯将更加艰难，祝愿他们工作顺利，生活幸福。

最后，感谢兰州财经大学统计学院的老师们，是你们带我进入统计学的大门，让我在数据分析知识的海洋里遨游，西北人民真实坚韧的美德与“兰财统计”四个字将伴随我的一生。希望疫情结束后还能回到兰州，与老师同学们吃碗牛肉面，品尝手抓羊肉的美味，把酒言欢。

满招损，谦受益。26 岁的我才刚刚踏入校园之外，生活中还有很多困难值得我去挑战，也有很多快乐幸福等待我去体验。未来会发生什么还不清楚，无论遇上冷风雨还是暖阳光，希望我能以一颗平常心勇敢面对，泰然处之。

再次道声感谢！兰州喂~ 兰州哦！