

基于企业网络招聘的人才需求文本挖掘 ——以统计学为例

屈文鑫, 王锦权

广东金融学院金融数学与统计学院, 广东 广州

收稿日期: 2024年6月1日; 录用日期: 2024年6月21日; 发布日期: 2024年6月30日

摘要

大数据时代的到来, 带来了数据形式和数据结构的革新, 也给统计学以收集、整理和分析数据为基础的专业创造了有利的发展机会。但随着互联网技术的迅速发展, 如人工智能、金融科技、大数据分析等新兴专业也不断涌现, 这也给传统的统计学专业人才培养和市场需求造成一定的冲击。本文通过数据挖掘方法分析企业网络招聘文本, 理清统计专业人才需求特征, 从而为学生求职和高校教学提供依据。研究分为两个阶段: 数据收集与预处理和深入挖掘招聘信息。结果显示, 企业对应聘者的能力要求分为数据分析、协同沟通和生产管理三类; 岗位类型分为业务类和技术类。

关键词

网络招聘, 统计学, 人才需求, 文本挖掘, LDA模型, 文本聚类

Talent Demand Text Mining Based on Enterprise Online Recruitment

—Taking Statistics as an Example

Wenxin Qu, Jinquan Wang

School of Financial Mathematics and Statistics, Guangdong University of Finance, Guangzhou Guangdong

Received: Jun. 1st, 2024; accepted: Jun. 21st, 2024; published: Jun. 30th, 2024

Abstract

The arrival of the big data era has brought about innovations in data forms and structures, and has also created favorable development opportunities for statistics professionals based on data collection, organization, and analysis. However, with the rapid development of Internet technology,

emerging majors such as artificial intelligence, financial technology and big data analysis are also emerging, which also has a certain impact on the training of traditional statistics professionals and market demand. This article analyzes enterprise online recruitment texts through data mining methods, clarifies the characteristics of professional talent demand in statistics, and provides a basis for student job seeking and university teaching. The research is divided into two stages: Data collection and preprocessing, and in-depth exploration of recruitment information. The results show that the ability requirements of companies for job applicants can be divided into three categories: data analysis, collaborative communication, and production management; job types are divided into business and technical categories.

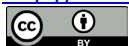
Keywords

Online Recruitment, Statistics, Talent Demand, Text Mining, LDA Model, Text Clustering

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着统计理论和方法的不断发展,市场对统计专业人员需求日益旺盛,使得统计专业人才需求市场具有可观的发展前景。但是,对于统计专业人员而言,应对身处的“内忧外患”局面具有清晰的认知。从目前的情况来看,高校统计人才培养体系存在滞后性和片面性,与市场需求相脱节,致使“就业缺口”问题长期存在[1]。换言之,一方面,学生面临毕业即失业的问题,难以找到对口的岗位;但另一方面,对于企业而言,学生的知识体系和专业素养与招聘需求不符,存在人才难得的问题。除此之外,当前国内经济增速放缓,就业形势不容乐观。

由此可见,探索统计专业人才需求,具有迫切性和必要性。本文以统计专业人才需求为重点,进行相关的工作。随着互联网技术的不断突破,网络招聘应用范围和频率日益提高。当企业具有人才需求时,可借助官网、招聘网站、自媒体等渠道,广泛地发布招聘信息,便于求职者获取相关信息。其中,招聘信息应重点突出岗位对于知识结构、教育背景和综合素养的要求,还应对企业基本信息、发展规划和主营业务简要介绍,便于求职者进行对标和匹配,获取到满意和符合自身特点的工作。综上所述,本文通过广泛搜集统计专业人才招聘信息,深入解析和剖析统计专业人才需求特征,为学生制定求职规划提供依据,也有助于高校制定合适的培养方案。

2. 国内外研究现状

对于人才需求分析的研究现状,国外长期以来都使用网上招聘数据来分析市场人才需求。一方面 Cullen 和 Kennan 等[2]通过对招聘广告的深入解读揭示出市场人才的需求的趋势和走向。另一方面 Smith 和 Ali [3]以程序员行业为例,对相关岗位信息进行系统梳理,准确获知对职业技能和研究工具的需求情况,为高校开展相关课程提供依据。而通过梳理国内的人才需求相关研究可知,国内主要侧重于对在线招聘数据进行分析。孙学军等[4]以物流类专业为研究对象,利用爬虫技术对网络招聘信息进行搜集,并在此基础上构建知识图谱,从而直观和清晰地揭示出创业型企业的人才需求倾向,为相关培训机构制定培训方案提供依据。

对于网络招聘信息的研究现状,国外关注的焦点主要可以分为三个方面。其中 Shenoy [5]从网络招聘

的具体环节出发, 了解招聘流程变更对求职者求职倾向的影响机制。而 Turrell [6]通过对千万级别的招聘信息进行搜集, 构建以职业分类编码为核心的新型算法。Litecky [7]则以 IT 岗位为研究对象, 通过应用系统聚类和 k-means 聚类算法提取出工作技能术语, 了解不同岗位的核心技能需求。国内对在线招聘信息的研究可分为在线招聘信息内容的深度挖掘和招聘信息分析算法。其中, 俞琰[8]通过应用术语抽取算法, 对招聘文本中的信息进行抽取, 实现对不同领域专业技能需求的对比分析。郑思雨[9]对某大型城市按方位划分为四个区域, 并对各个区域的招聘信息进行关联规则分析和 k-means 聚类, 从中寻找岗位要求中隐含的关联信息。

通过梳理相关领域的文献可知, 与国外相比, 我国对于文本挖掘的研究起步较晚, 但仍取得诸多有益成果。在早期阶段, 国内主要侧重于对网络招聘模式优缺点进行探讨, 以定性分析为主, 缺乏具体的实证分析。但是研究人员更加注重对人才需求的分析, 对社会网络分析也尤为关注, 并展开了系统性的研究。因此, 本文以网络招聘文本为切入点, 深入剖析其中对人才的隐形要求, 通过对研究进行的总结和提炼, 为优化人才培养方案提供依据。对于网络招聘信息数据领域, 主要可划分为两类。具体来讲, 第一类, 对已经获取的网络招聘信息进行分析, 以期从中揭示出网络招聘数据与特定社会现象间的具体关联; 第二类, 侧重于寻找适用于招聘数据的获取方法、评估模型和计算算法, 从而提升评估和分析的质量。但是从现有研究来看, 仍未能理清和揭示与职位相关的因素分布。同时, 对于岗位所需技能的评估存在片面性, 缺乏系统和全面的考虑。此外, 对于职位要求缺乏准确和清晰的分类描述方法, 在后续应进一步弥补。

3. 理论基础

3.1. 文本信息向量化

文本信息向量化中的 BOW (词袋模型)和 TF-IDF (词频 - 逆文档频率)是两种常用的方法。

BOW (Bag of Words)是一种常用的文本向量化方法, 它将文本数据表示成能够表达文本语义的向量。具体来说, BOW 方法将文本中的每个词表示为一个向量, 向量的维度等于词典的大小。每个词的向量表示方式通常是该词在词典中的位置或权重。BOW 方法的优点是简单易行, 能够快速地将文本数据转换为数值向量。但是, BOW 方法忽略了词序和语义信息, 因此对于某些任务可能不够准确。为了解决这个问题, 可以采用一些改进的方法, 例如使用 N 元模型等上下文信息来生成词向量。

TF-IDF 是一种常用的文本向量化方法, 全称为词频 - 逆文档频率, 用于评估某个词在一份文件中的重要程度。具体而言, 其中, TF(词频), 顾名思义是指某个词占文档总词数的比例; IDF (逆文档频率)则主要用于评估某个词语普遍重要性。TF-IDF 的公式为:

$$TF-IDF(t, d) = TF(t, d) * IDF(t)$$

其中, $TF(t, d)$ 是指词 t 在文档 d 中出现的频率, $IDF(t)$ 是指词 t 的逆文档频率。TF-IDF 的优点在于能够考虑到一个词在特定文档中的重要程度, 并且能够降低噪音和无关词的影响。因为一个词在文档中若出现频率较高, 则表明与文档的关联性强, 从而具有一定的代表性; 而一个词在多个文件中出现, 说明该词比较通用, 对于特定文档的代表性可能较弱。在实际应用中, TF-IDF 在风险评估、舆情分析和文本分类等应用相当不错。同时, 通过标准化的 TF-IDF 向量, 为后续工作提供基础, 并开展相似度比较、分类、聚类等操作。

3.2. LDA 主题模型

2003 年, Blei [10]等学者在进行文档主题分布研究时, 提出 LDA 主题模型。该模型可通过对文本数

据进行分析, 从中探寻其中暗含的主题结构。按照 LDA 模型的相关理论, 认为文本的是指为某种特殊的概率分布。其中, 每个文档中混合着大量的不同主题。而对于不同的主题, 则包含着大量的单词。根据 LDA 模型生成过程的特点, 可划分为三个阶段, 主要包括: 第一, 确定文档的主题分布; 第二, 确定文档主题; 第三, 选择单词。然后对上述过程进行反推, 最终可以获得到 LDA 模型的参数估计方法。再深入地讲, 可以借助文本数据中观察到的单词, 从而评估出各个主体的单词分布, 再进一步获取到各个文档的主体分布, 以此类推最终推断出文本的主题结构。

LDA 模型的应用非常广泛, 包括情感分析、主题识别、趋势预测等。在商业环境中, LDA 可以帮助企业分析客户反馈、了解市场趋势、预测销售业绩等。同时, 在学术研究领域, LDA 在新闻推荐、电子商务、信息检索等诸多场景均得到不错的应用[11] [12]。

4. 数据收集与数据处理

4.1. 数据获取

本文的主要关注点是中国对统计专业人才的需求。在某种程度上, 网站招聘信息可以反映就业市场对统计专业人员的需求, 现阶段线上招聘已经成为企业最为普遍的招聘途径, 其中以官网和第三方招聘网站为主。对于官方网站, 招聘信息普遍地分散和凌乱, 而对于第三方招聘网站, 对于信息发布具有明确的要求和统一的规格, 便于后续的分析 and 统计, 用于招聘信息搜集更为合适。因此, 本文选择前程无忧作为数据来源, 数据收集截止于 2024 年 1 月 26 日, 以 Python 为工具爬取了岗位详情页信息等相关字段信息。

4.2. 数据预处理

通过搜集到的信息分析发现, 其中存在诸多问题, 具体表现为: 第一, 招聘企业的信息可能重复, 岗位信息部分内容缺失; 第二, 部分内容所取字段过多, 不利于后续开展数据分析, 其中的典型代表为公司情况介绍、地理位置等; 第三, 以“统计”为关键词, 搜索范围过于宽泛, 难以对统计专业人才的实际需求进行精准获知。对于以上出现的问题, 本文采用的解决方案如下: 对于重复值和缺失值进行取舍、补充, 其次对统计专业进行识别, 并将数据标准化处理。

4.3. 分词

对于中文这类非结构数据, 分词是其中的重要环节, 主要内容是将句子转换为单词表示。例如, “熟练使用办公软件”可分为“熟练/使用/办公/软件”。通过进行分词, 从而实现文本向量化, 并在统计分析的基础上构建词频矩阵, 使文本信息转换为数字信息。因此, 分割工作发挥着决定性的作用。因为不同的分割结果, 必然产生不同的文本挖掘结果, 直接影响后续共工作的成败, 应予以重点关注。在本文中, 应用了 Python 的 jieba 库进行分词。

5. 数据分析

5.1. 统计人才需求的主要特征

1、企业画像

从公司名称来看, “科技”关键词更为突出, 占比接近 50%。通过对图 1 分析可知, 科技型企业对统计专业人才需求比较旺盛。因为此类企业普遍涉及信息、生命科学、精密制造、生物医药等高精尖技术领域, 比较侧重于应用数据和信息进行决策。在此基础上, 以“科技”为关键词, 进行拓展分析。经过统计可知, “食品科技”、“医药科技”和“材料科技”最为常见, 表明这些科技型企业主要集中在

食品、医药、材料工程方向。此外,“汽车”、“医疗”、“实业”等关键词出现频率较高,均为5%左右,表明汽车技术、医疗技术和工业企业对统计专业人员也有一定的需求。

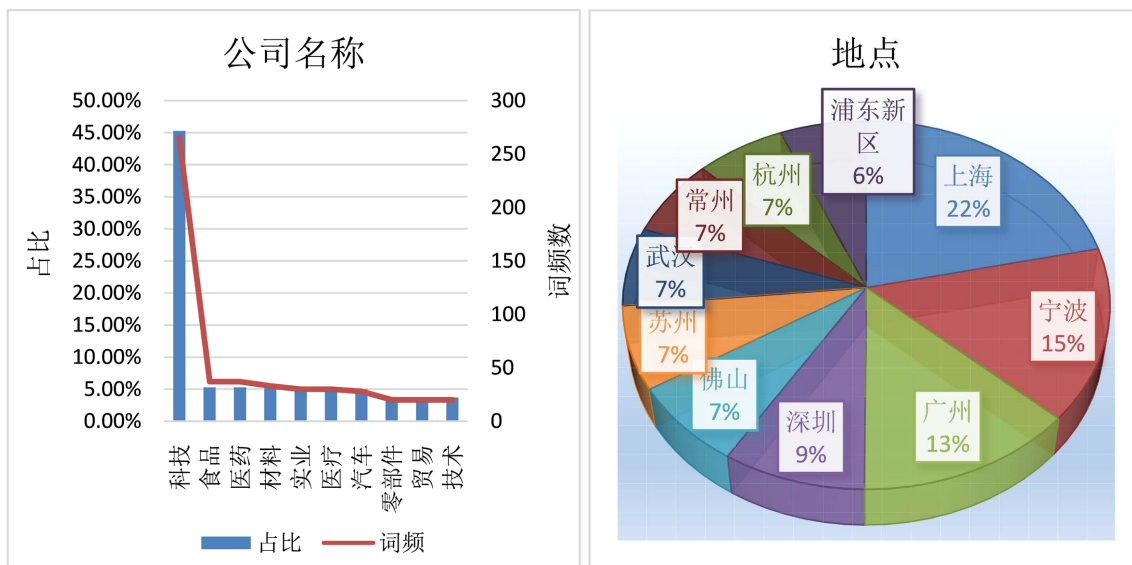


Figure 1. The proportion of word frequency in company name (left) and company location (right)
图 1. 公司名称(左)及公司地点(右)词频的占比情况

从公司所在地来看,对统计专业人才需求占比前三的分别为:上海、宁波、广州,占比分别为22%、15%、13%。通过地域对统计专业人才需求进行统计,其中,华东和华南的需求量最大,占比分别为41.77%和35.16%。通过深入分析发现,其中,上海、苏州、常州等对统计专业人才需求较多的城市,均集中于华东区域。而广州、深圳、佛山等对统计专业人才需求较多的城市,主要集中于华南地区。除此之外,对于华北、西南和华中区域进行统计,占比分别为8.31%、5.98%和10.02%。而对于西北和东北地区,具有统计专业人才需求的企业相当之少,占比不足2%。对于企业而言,选择落户的因素可能包括人才资源丰富、交通便利、市场潜力大、政策扶持等。例如,长三角城市群开设医疗器械相关专业的高校数量最多,共有103家;粤港澳大湾区在专业服务方面存在领先优势;而从交通基础设施来看,长三角地区基础设施建设水平高,主要交通运输指标都明显优于京津冀。这些地区企业重视科技创新,对信息化和自动化等新兴科技比较敏锐,对企业数字化探索转型探索更为深入。因此,对统计专业人才的需求也很旺盛。

2、资历画像

从工作经验来看(图2),其中1年经验出现最为频繁,而3~4年经验、2年经验紧随其后,占比也较高。通过对上述情况分析可知,企业为减少人员培养成本,尽快入手岗位工作,普遍会优先考虑具有工作经验的统计专业人才。从教育水平来看,统计专业人员对大专学位的需求最多,其次是中技,占15%。对博士学位的需求相对较小,仅为0.1%左右。

总之,通过对统计专业人才需求统计分析可知,其中与工作经验和教育具有密切关联。其中,对于尤其是对于1年工作经验和大专学历,需求最为旺盛。而上述规律呈现的特点为:即企业对教育背景的要求不高,但是对经验的要求更高。企业不愿意培养刚毕业的本科生的原因有很多,其中一些原因包括:① 企业培养人才的成本较高。② 技术工种的特点决定了对经验的要求更高。③ 本科生的教育背景可能与企业需求不匹配。

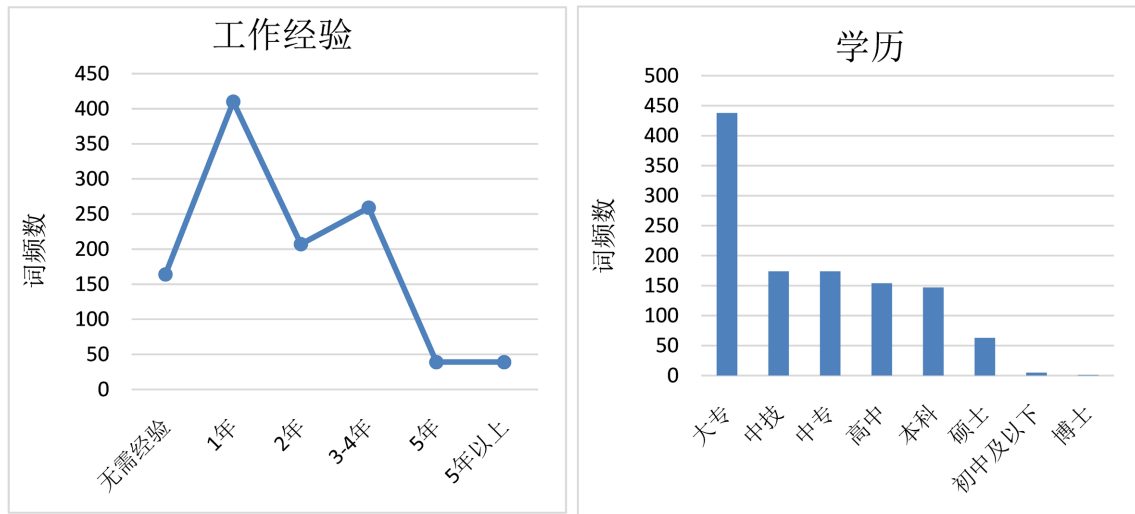


Figure 2. Work experience (left) and educational background (right) word frequency
图 2. 工作经验(左)及学历(右)词频情况

3、岗位画像

从职位名称的单词云图(图 3)可以看出,“统计”、“主管”、“财务”和“统计师”等词出现频率较高。其中,技术岗位主要为质量管理、数据挖掘、数据分析等专业素养要求较高的岗位。而“主管”“仓管”“管理部”等词条,表明需要统计专业人员具有一定的管理能力,主要以财务、HR 等领域为主。除此之外,它还包括生物学、医学、产品、市场和研究等领域的统计工作。根据岗位信息整体分布情况可知,每个月 4~7 千的薪资占据主体,对于 7 千~1.125 万的薪资紧随其后。通过分析发现,统计专业人才月薪主要在 7 千附近范围波动。

通过对上文分析可知,统计专业人才市场需求主要以“统计”、“主管”、“财务”类型为主,侧重于技术类和管理类。而上述岗位除对专业知识要求严格外,对管理能力也提出较高要求,并应该具有一定团队合作和沟通交流能力。而研究中显示,月薪主要分布在 4 千~1 万的区间内,7 千左右是主流。

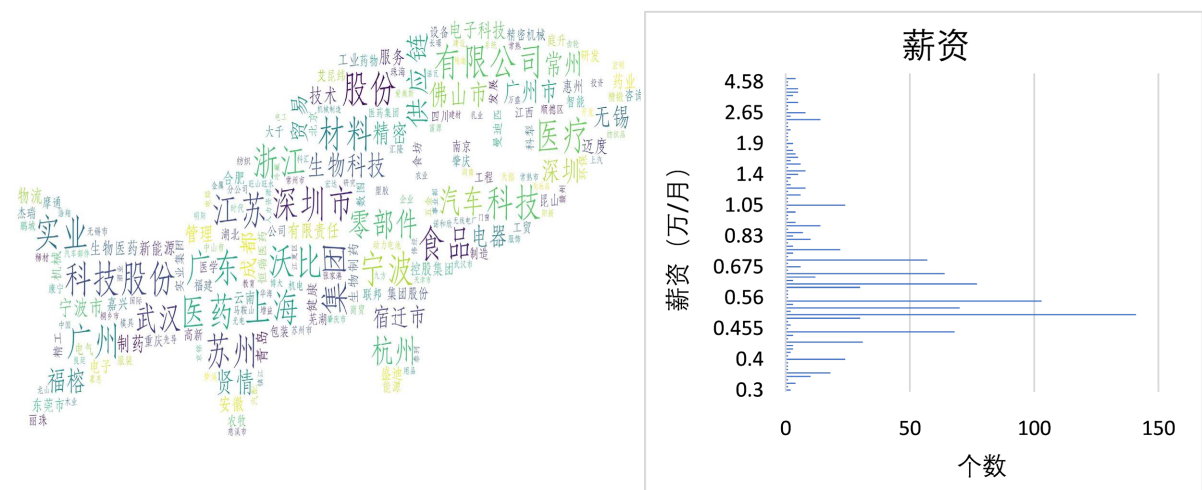


Figure 3. Position name word cloud chart (left) and position salary bar chart (right)
图 3. 岗位名称词云图(左)及岗位薪资条形图(右)

5.2. 岗位要求的主题词提取

由于 LDA 主题模型的无监督性质, 我们无法获得最佳的主题数量。因此, 我们可以使用一致性 (Coherence) 度量来计算主题中高概率单词的语义相似性, 并评估 LDA 主题模型的质量。通过计算一致性公式来测量不同主题的一致性, 一般情况下, 具有较高一致性的主题数量被认为是 LDA 模型的最优主题数量。计算公式为:

$$Coherence(V) = \sum_{(V_i, V_j \in V)} score(V_i, V_j, \epsilon)$$

其中, V 表示描述主题的一系列单词, 以及 ϵ 表示平滑系数。计算不同主题的一致性, 一致性较高的主题数量通常是 LDA 模型的最优主题数量。如图 4 所示, 主题个数为 3~5 时, 一致性得分较高, 为位于 0.6~0.8 之间, 而 6 往后则呈现波动状态, 综上初步确定主题数为 3。

使用 pyLDAvis 这个可视化模块, 我们可以将 LDA 主题模型中的三个主题编号与其对应的主题进行可视化展示, 具体结果如上图所示。pyLDAvis 可在二维坐标系中绘制出文档的主体分布, 具体情况如图 1~3 所示。其中, 面积大小反映出主题在文档中的出现机率, 不同圆圈的距离则反映出主题的相似程度。具体而言, 如图 5 所示, 当 $K = 3$ 时, 三个圆分散在坐标系的三个角上, 其距离相对较远, 这表示主题之间的相似度较低。因此, 此时主题之间的差异化程度较高, 并且该 LDA 模型的表现较好。

综上所述, 根据一致性指标分析和主题模型可视化结果, 我们可以将主题模型中的主题数量参数设置为 3。

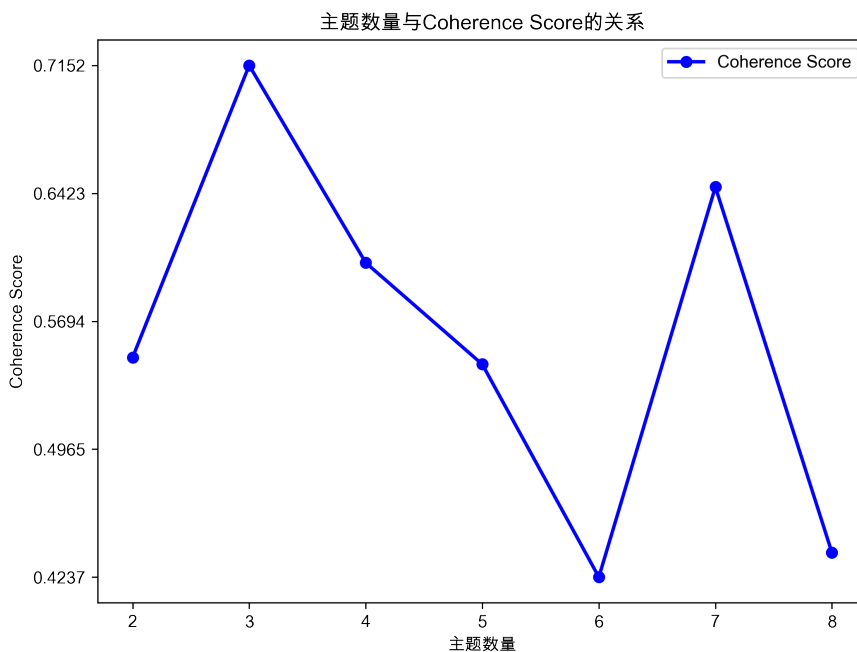


Figure 4. Consistency score for different themes
图 4. 不同主题对应的主题一致性得分

在理解了单词的含义之后, 可以将这三个主题概括如下(表 1):

主题 1 是数据分析, 包括统计软件、算法、数据挖掘、编程、收集数据等。技能要求掌握统计学基础, 如描述性统计、推断性统计等; 熟悉数据处理工具, 如 Excel、SQL、Python 等; 理解数据可视化的基本概念和技术, 如 Tableau、PowerBI 等; 具备基本的机器学习和数据挖掘知识。岗位职责主要负责收

集、整理和分析业务数据, 为决策提供数据支持; 通过数据分析, 发现业务问题, 提出改进方案; 建立和维护数据分析模型, 提高数据分析效率。

主题 2 是协同沟通, 需要承担一些研究任务, 主要面向市场调研、医药研发、金融服务等岗位。技能要求具备良好的沟通能力和团队协作能力、能够有效地表达自己的观点和想法、理解和尊重他人的观点和想法、具备基本的项目管理知识和技能。岗位职责包括与团队成员、其他部门或外部合作伙伴进行有效沟通, 确保信息的准确传递; 协调各方资源, 推动项目的顺利进行; 解决项目过程中出现的问题和冲突。

主题 3 是生产管理, 需要具有执行层面的能力, 例如组织、策划、协调等方面。技能要求掌握生产管理的基本理论和方法, 如生产计划、生产控制、质量管理等; 具备基本的项目管理知识和技能; 理解和掌握生产流程和工艺。岗位职责要求制定和执行生产计划, 确保生产目标的实现; 监控生产过程, 确保产品质量; 优化生产流程, 提高生产效率; 处理生产过程中出现的问题。

综上, 从企业的角度来看, 由 LDA 主题模型可以将应聘者能力提炼为三个主题, 也就是企业对应聘者能力的要求, 分为数据分析能力、协同沟通能力和生产管理能力。

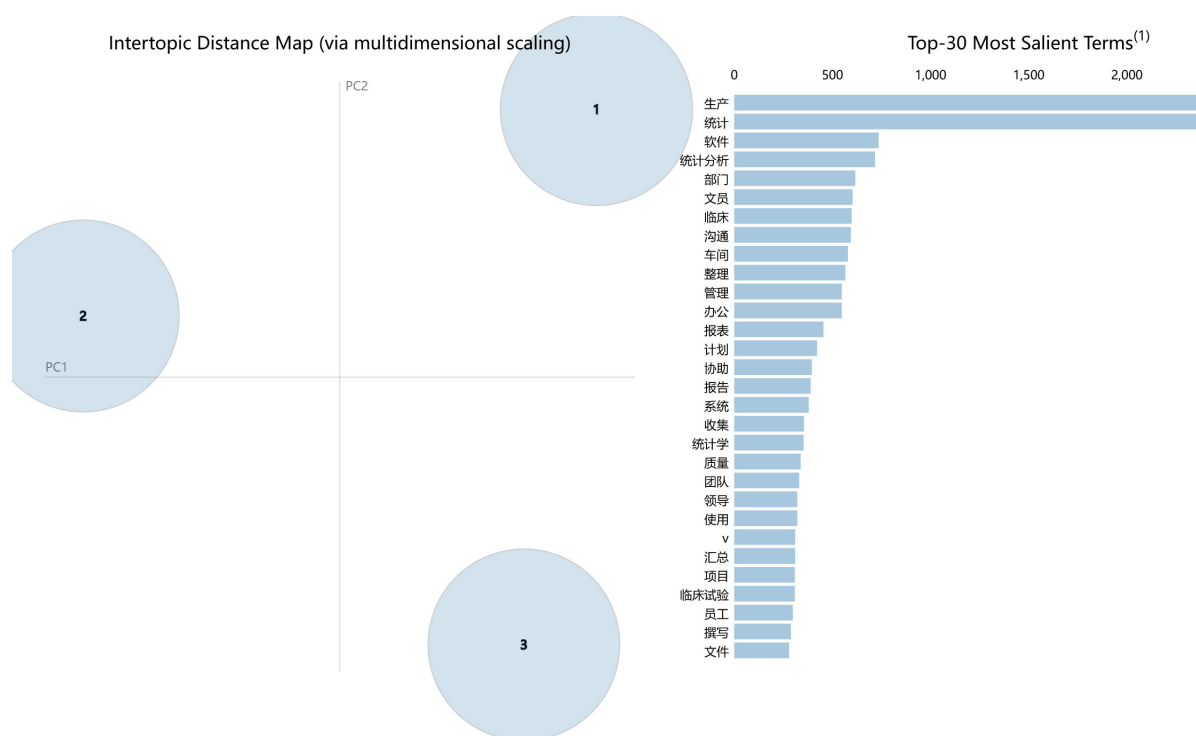


Figure 5. Visualization of topic distance for 3 themes

图 5. 主题数为 3 时的主题距离可视化图

Table 1. Topic feature words for different topics

表 1. 不同话题的主题特征词

主题 1	统计	软件	临床	车间	协助	收集	统计学	团队	临床实验
	SAS	文件	研究	编程	及时	合作	生物	情况	统计员
主题 2	部门	文员	沟通	管理	办公	计划	领导	使用	汇总
	项目	员工	撰写	方案	支持	内容	责任心	录入	精神
主题 3	生产	统计分析	整理	报表	报告	系统	独立	做好	岗位
	一定	财务	产品	考勤	物料	编制	更新	时间	了解

5.3. 岗位的文本聚类

通过岗位需求文本聚类,可从中全面地了解不同类型岗位的需求特征。具体的应用过程如下:首先,使用 TF-IDF 对文本进行向量化,将文本转换为数值表示。然后,基于多个聚类结果的可解释性,选择适当的聚类数量。接下来,应用 T-SNE 算法将文本特征向量进行降维处理,以便于可视化展示。在此基础上,输出聚类结果的可视化图形,并根据作业名称中关键词的频率来命名每个聚类类别。这样可以使得聚类结果更具可解释性和可理解性。

在本文研究中,对于招聘数据中的岗位要求,应用 K-Means 聚类算法进行文本聚类。同时,经过多次尝试之后,在观察每个类别的特征后,发现了一些有噪声的数据。由于噪声数据对聚类结果的显著影响,它们被删除。选择 2 个聚类产生最佳的聚类效果和最强的可解释性(图 6)。

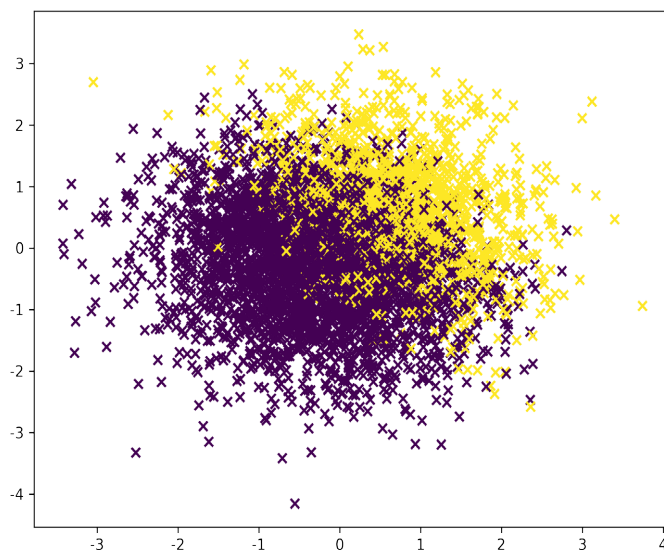


Figure 6. Job clustering results

图 6. 岗位聚类结果

分别对两个岗位类别排名前 10 的聚类词频进行了统计,如下所示:

岗位类别 1: 数据、协作、管理、沟通、报表、系统、监督、文员、职能、办公。

岗位类别 2: 技术、操作、软件、汇总、计算、统计员、开发、表格、SAS、SQL。

从上述两个岗位类别的聚类词频中可以清楚地看出,文本聚类非常有效地帮助我们将工作需求分为两类,并且这两类在工作所需的高频词方面存在显著差异。第一类侧重于数据、协作、管理和通信等关键词。可以看出,这种类型的数据分析职位更倾向于商业,强调管理、沟通、合作等软实力,而不是技术。第二类职位的关键词主要集中在技术、运营、软件总结上,表明这类职位主要倾向于技术。然而,在这两类职位的高频术语中,有“表格”和“报告”。由此可见,无论是从事技术数据分析还是业务数据分析,都需要处理大量的报告,最终输出分析报告。

综上,从求职者角度来看,由 k-means 文本聚类可以将岗位分成业务类和技术类两类。首先,从行业分布来看,业务类岗位分布更广,数据类岗位主要分布在互联网行业。其次,业务类岗位更加强调和注重财务、会计和人力资源等方面的知识储备,特别是行政管理等方面具有较高要求,更加强调办公职能;对于工作技能来讲,办公软件、SAS、SQL 是普遍需要掌握的。对于不同岗位,侧重点有所不同。对于技术类岗位,对于编程语言和大数据工具要求比较严格;对于数据分析师而言,对视觉呈现更为关

注。因此,对于 SAS 和 PERL 应用的相对较为频繁。从综合素质的角度而言,对于沟通协调、团队协作、文字表达、逻辑思维等能力,所有岗位都迫切需要和非常重视。但是,对于不同岗位的侧重点存在差异,例如技术类岗位需要具有准确判断力、执行力强且思维缜密,业务类岗位更强调执行力和抗压能力。

6. 结论与建议

本文通过对“前程无忧”中的企业工作要求进行关键词提取,得到了对统计专业人才的核心诉求,即数据分析、协同沟通和生产管理能力。因此,对于相关专业学生而言,应该有目标地进行能力培养,以便适应市场的需求。不同岗位在行业性质方面存在一定差异。可分为业务类和技术类:业务类岗位更强调办公职能,技术类岗位更强调数据分析。业务类岗位分布更广,数据类岗位主要在互联网行业;业务类岗位强调财务管理、人力资源管理和行政管理,技术类岗位使用多样化编程语言和大数据工具;办公软件、SAS、SQL 是所有职位重要技能,但各有侧重;对于沟通协调、团队协作、文字表达、逻辑思维等能力,都很重要,但技术类岗位更侧重于判断力、执行力强且思维缜密,业务类岗位更强调执行力和抗压能力。

针对当前市场对统计专人员需求趋势,高校首先应当合理地设置专业课程,做到重点突出、特色鲜明和层次丰富,为学生走向工作岗位打下基础,同时也应重视专业技能培养,对于领域的主要研究工具做到熟练掌握,针对 PYTHON、SAS、TABLEAU 等应用频率高的软件,可专门开设培训班,或者组建兴趣小组等,做到紧跟市场需求,优化课程结构;其次要加强通识课程教育,对于团队协作、心理健康、人际关系等领域均要有所涉及,帮助学生实现综合发展,适应未来岗位竞争的需要;再有,现阶段高校的主要精力集中在专业知识教学方面,存在重理论轻实践的倾向,对工作技能培养相对欠缺,使得学生在就业市场的竞争力不强。而就业是学生无法避开的重要问题。因此,应通过多措并举,提升学生的就业能力。

对于统计类专业学生,应认真审视自我,系统梳理自身的优劣势,了解自己的知识结构、技能储备和职业精神等情况,从而评估适合的就业方向;对于业务类岗位可结合目标岗位的需求,提前进行相关技能的打磨,提升求职的成功率。主动参与社会实践,提前适应职场生活,社会实践与职场生活具有诸多相通之处。同时应加强办公软件学习,以便满足工作的基本需要,提高工作的标准化和规范化程度;对于技术类岗位,应注重掌握统计领域的专业软件,对于 Python 或 R 编程语言,应重点予以关注。同时,通过参考和查阅相关领域的文献,了解业内的技术和理论进展情况,有针对地进行学习。BI 和 TABLEAU 等可视化软件,在条件允许的前提下应尽量予以掌握,突出自身的竞争优势。

参考文献

- [1] 卫露楠. 基于文本挖掘的统计专业人才需求分析[D]: [硕士学位论文]. 桂林: 广西师范大学, 2022.
- [2] Cullen, J. (2004) LIS Labour Market Research: Implications for Management Development. *Library Management*, **25**, 138-145. <https://doi.org/10.1108/01435120410522352>
- [3] T Smith, D. and Ali, A. (2014) Analyzing Computer Programming Job Trend Using Web Data Mining. *Issues in Informing Science and Information Technology*, **11**, 203-214. <https://doi.org/10.28945/1989>
- [4] 孙学军, 齐俊景, 阚曲欣. 基于招聘数据挖掘的高职人才培养方案构建研究——以物流类专业为例[J]. 职教论坛, 2019(8): 159-164.
- [5] Shenoy, V. and Aithal, P.S. (2018) Literature Review on Primary Organizational Recruitment Sources. *International Journal of Management, Technology, and Social Sciences*, **3**, 37-58. <https://doi.org/10.47992/ijmts.2581.6012.0035>
- [6] Turrell, A., Speigner, B., Djumalieva, J., et al. (2019) Transforming Naturally Occurring Text Data into Economic Statistics: The Case of Online Job Vacancy Postings. *Social Science Electronic Publishing*.
- [7] Litecky, C., Aken, A., Ahmad, A. and Nelson, H.J. (2010) Mining for Computing Jobs. *IEEE Software*, **27**, 78-85.

<https://doi.org/10.1109/ms.2009.150>

- [8] 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.
- [9] 郑思雨. 网络招聘信息的数据挖掘研究[D]: [硕士学位论文]. 杭州: 杭州电子科技大学, 2020.
- [10] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.
- [11] 董韶琦, 郑静. 基于 LDA 主题模型的杭州亚运会微博话题分析[J]. 统计学与应用, 2023, 12(4): 833-842.
- [12] 唐诗. 基于文本挖掘的豆瓣电影评论的 LDA 主题模型分析——以电影《让子弹飞》为例[J]. 新闻传播科学, 2024, 12(1): 23-28.