基于文本挖掘的企业招聘数据分析

——以数据分析岗位为例

章群英，史培瑶，陈鹏鑫

（嘉兴学院商学院，浙江 嘉兴 314001）

摘　要　在大数据时代背景下，企业招聘需求信息成为求职者的求职关键，能读懂企业招聘需求信息，合理利用信息，是求职者需要具备的能力。为帮助求职者理解企业招聘信息，本文以数据分析岗位为例，探索就业市场中企业招聘单位对数据分析岗位的用人需求，挖掘就业市场对于相关领域人才的技能要求。利用Python构建了词云图和LDA主题模型以对企业需求进行探测分析。结果表明，对于数据分析岗位，数据挖掘、建模、编写算法等专业技术能力是企业招聘的首要关注点，其次为求职者的个人特质、沟通能力及商业知识等。

关键词　词云图；LDA；文本挖掘；企业招聘；就业指导

中图分类号：TP3　　**文献标识码**：A　　**文章编号**：1007-0745(2023)05-0073-03

高校招生数量的增多以及企业招聘要求的提高，使就业形势日益严峻，同时，企业招聘需求信息大多以文本形式展现，具有海量和非结构化等特点，因此对于求职者来说，在海量的岗位信息中搜寻自己想要的关键信息是较为困难的。为了真实深入地了解企业的需求变化，辅助求职者更高效地就业，帮助求职方与招聘方实现双赢，提供给求职者具有参考价值的信息，本文收集BOSS直聘、拉勾网以及应届生求职网站中数据分析岗位的企业招聘需求数据，利用Python自编程序创建词云图、构建LDA主题模型，对企业招聘需求进行探测分析，为求职者提供就业指导。

1 文献综述

文本挖掘是近年来一个新兴研究领域，主要是从大量的、无结构的文本信息中发现潜在的、可能的数据模式、内在联系、规律、发展趋势等，抽取有效、新颖、有用、可理解的、散布在文本文件中的有价值知识，并且利用这些知识更好地组织信息的过程[1]。近年来，文本挖掘被应用于许多领域，对招聘信息的文本挖掘也逐渐兴起，如潘保国等人利用文本挖掘，爬取并分析招聘网站中与大数据相关的岗位，从不同方面对招聘信息的调研结果进行分析，提取出不同地区对相关岗位的需求量及用人单位对求职者的要求[2]。袁莉通过对互联网企业招聘数据进行文本挖掘，从不同维度的数据进行分析，解答互联网岗位招聘的相关问题，为求职者高质量就业提供决策指导[3]。吕宏红则从新角度出发，利用文本挖掘技术对人力资源市场进行分析，理清求职者、企业和高校子系统之间的关系，探讨人力资源市场系统与环境之间的信息交换，更加有效地为求职者提供求职建议，同时也为我国人力资源市场的高效运转提供了全新的实践视角[4]。郝素利、王瑞芳利用文本挖掘获取招聘信息，对会计人才的整体需求进行分析，并依据分析结果给出了会计人才培养的具体建议[5]。这些都体现了文本挖掘技术能提取出更丰富、更有效的信息，在解决就业招聘问题方面具有强大的生命力。

2 研究过程

2.1 数据来源和研究方法

本文选定国内规模较大的BOSS直聘、拉勾网以及应届生求职网进行研究，通过获取1940余条关于数据分析岗位的招聘需求信息，其中包括岗位名称、岗位描述和岗位要求，并对数据样本进行了文本分词、去除停用词、短句删除等预处理操作。主要利用Python软件进行词云图的制作和LDA模型的构建。在制作词云图的过程中，利用事先导入的WordCloud库进行初

★基金项目：本文为2022年度嘉兴学院大学生研究训练（SRT）计划项目“基于多源数据的大学生就业辅助系统的设计与实现——以浙江省高校为例”（编号：8517221223）研究成果。

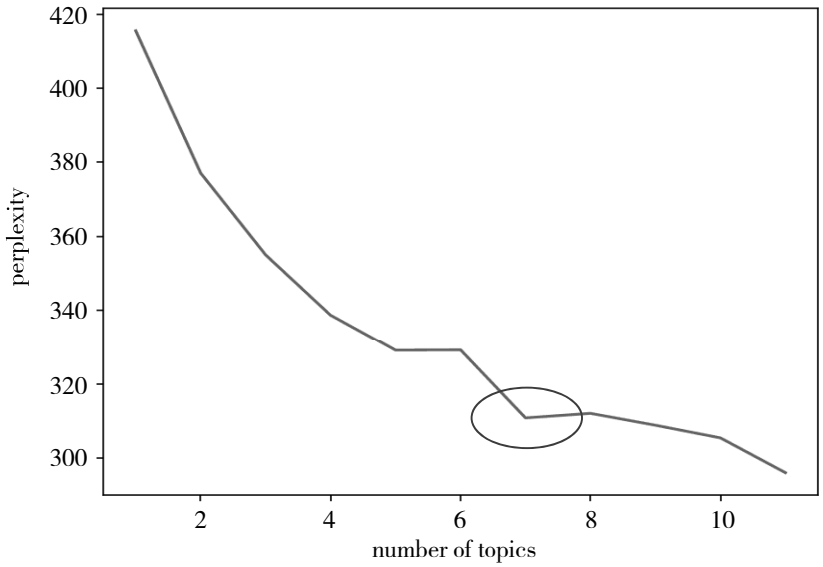

图1 主题数与困惑度的折线图

图2 词云图

始参数设置，包括字体、背景颜色、最大单词量、停用词等，并在过程中根据词云图的效果进行参数的调整，以得到最优效果。在LDA模型的构建过程中，采用了手肘法和计算困惑度（Perplexity）来确定LDA主题模型的最佳主题数，绘制的主题数与困惑度的折线图如图1所示，在图1中，随着K（Number of Topic）值的增大，困惑度（Perplexity）逐渐减小。根据手肘法，并且当K约为7的时候，存在一个显著的拐点：当K属于（1，7）时，曲线急剧下降；当K属于（7，11）时，曲线基本趋于平稳；故拐点7即为K的最佳值，因此在本文所选定的数据集中，LDA主题的生成数量选定为7时效果最佳。

2.2 词云图分析结果

本文对采集到的关于数据分析岗位的招聘要求数据进行数据预处理，并利用Python中的WordCloud生成更具视觉冲击力的词云图，如图2所示。

经分析可发现，词云图体现的企业对求职者的需求大致可分为专业知识与特长、工作态度以及人际关系的处理能力三方面。在专业知识与特长方面，图中的“建模”“数据挖掘”“开发”“设计”等字样较为突出，且关于该方面的字样在图中出现的频率较高，所占篇幅较大，这表明专业知识与特长是企业对数据

分析岗位的求职者的主要关注点，且需要求职者掌握多种技能。在工作态度方面，“负责”“学习”“独立”等字样较为突出。这表明，对于数据分析岗位，企业需要的是负责独立、具有上进心的人才。同时也告诉该岗位的求职者，需要培养自身的独立思考能力，要有足够的耐心和深入业务的思考来进行决策支撑。在人际关系的处理能力方面，“沟通”“合作”“协作”等词语较为突出。这表明，作为一名合格的求职者，仅有优秀的个人能力是不够的，还需要学会与团队合作，与他人沟通。

综上所述，对于数据分析岗位，求职者的专业知识与特长是企业的首要注重点，优秀的专业知识与特长是求职者竞争力的根本，其次严谨的工作态度以及良好的人际关系的处理能力也能提升求职者的竞争力，实现自身竞争力的最大化。

2.3 LDA 主题模型分析结果

为进一步探究企业的招聘需求，本文利用评论数据构建 LDA 主题模型，通过主题词直观地分析招聘需求。通过 LDA 主题建模生成 7 个主题，其中每个主题的前 5 个高频词取出，如表 1 所示。

表 1 主题 - 特征词分布表

主题	主要关键词 (概率)
0	模型、建模、算法、数据挖掘、机器
1	运营、工具、数学、决策、方案
2	技术、设计、团队、数据库、系统
3	责任心、团队、合作、精神、抗压
4	项目、行业、业务、专业、流程
5	管理、报告、市场、销售、运营
6	产品、设计、测试、用户、文档

经表 1 分析发现，企业对于数据分析这一岗位的招聘需求主要分为四个方面：

在求职者专业能力方面，LDA 模型中出现了“模型”“算法”“数据挖掘”“技术”“数据库”等主题，说明对于求职者的专业能力，企业希望求职者精通数据分析技术、拥有较强且多样的编程能力。

在业务理解方面，LDA 模型中出现了“业务”“流程”“项目”“运营”“产品”等主题，说明企业希望求职者在数据分析前能明确项目的业务流程、产品的特性，理解项目的运营方式，再进行全面的数据分析。

在个人特质方面，LDA 模型中出现了“责任心”“团队”“抗压”“合作”等主题，可见企业更倾向于富有责任心、抗压能力强大的求职者。

在其他方面，LDA 模型中出现了“市场”“销售”“用户”等主题，说明数据分析岗位可能还需要求职者拥有一定的商业知识，能够理解市场规律，察觉用户需求的求职者会具有更大的优势。

综上所述，企业在选择求职者时，首先关注求职者的专业能力是否合格，是否能够满足企业的实际需要，其中数据库、数据挖掘等数据分析技能以及算法、模型等编程技能是其关注的重点；其次，求职者的业务理解能力以及其个人特质也是影响企业选择的重要因素。

3 总结与建议

本文对 BOSS 直聘、拉勾网以及应届生求职网中关于数据分析岗位的企业招聘需求数据进行了搜集，经数据预处理后，利用 Python 自编程序得到了词云图和关于招聘需求的 7 个评论主题，同时对得到的数据结果进行综合分析，结果表明，对于数据分析岗位，企业招聘时最为注重求职者的专业技术能力，其次为求职者的个人特质，业务理解能力，沟通能力及其他能力。

从词云图以及 LDA 模型中可以看出，专业知识是求职者的核心竞争力，求职者对于 SQL、Python 等数据分析和编程技术以及数据分析相关工具的使用方法，掌握的程度越深、种类越多，在求职时就越有优势；个人特质是成功求职的重要条件，优秀的独立思考能力、业务理解能力以及沟通能力等都能在不同程度上为求职者加分，增大成功的概率；其他方面，对于数据分析岗位的求职者而言，掌握一定的商业知识，懂得市场规律能获得更大的优势。

参考文献：

[1] 郑双怡 . 文本挖掘及其在知识管理中的应用 [J]. 中南民族大学学报 : 人文社会科学版 ,2005,25(04):127-130.
[2] 潘保国 , 黄永杰 , 张慧敏 , 等 . 基于招聘网站的数据科学与大数据技术专业人才需求的文本挖掘 [J]. 湖北工程学院学报 ,2022,42(06):94-98.
[3] 袁莉 . 基于文本挖掘的互联网企业岗位对比研究 [J]. 现代信息科技 ,2022,06(07):112-115,119.
[4] 吕宏玉 . 雇主网络口碑的情报挖掘研究 [D]. 南京 : 南京大学 ,2021.
[5] 郝素利 , 王瑞芳 . 基于 Web 文本挖掘的会计人才需求分析 [J]. 中国管理信息化 ,2022,25(19):165-173.