Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/64/20241424

Data job requirement characteristics analysis based on large language model and K-means clustering

Jie Zhou[1,2], Yan Chen[1,3], Yiling Guo[1,4], Chong Zhang[1,5,*]

[1]Beijing Language and Culture University, No. 15, Xueyuan Road, Haidian District, Beijing

[2]xk20zj@126.com,
[3]2093362371@qq.com
[4]byxk21gyl@126.com
[5]zhangchong@blcu.edu.cn
*corresponding author

Abstract. This paper aims to analyze the characteristics of data-related job requirements using a Large Language Model (LLM), providing decision support for the construction of data-related courses in universities. Using Zhaopin.com as the data source and employing a key attribute extraction method based on ChatGLM3 along with prompting engineering techniques, this study extracts and clusters key attributes from job descriptions. It analyzes the characteristics of data-related job requirements in terms of competencies, fields of expertise, and tools. The research found that the quality of key attribute extraction using ChatGLM3 improved by 54.1% compared to dictionary-based methods, significantly reducing the workload of manual annotation. Additionally, the analysis revealed commonalities and differences in the demand features of data analysis, data engineering, and data operations roles across six dimensions: soft skills, tools, specialties required, educational background, experience, and salary.

Keywords: Large Language Model, Text Mining, ChatGLM3, Clustering Algorithm, Requirement Analysis.

1. Introduction

With the rapid development of information technology, the digital economy has become a new driving force for global economic growth. Under the influence of policies and industry trends, the integration of data and the real economy is deepening, achieving significant developments in data industries and data applications. In October 2023, the National Bureau of Data officially launched in Beijing, marking a stronger data-driven force for the construction of Digital China and the perfection of the data market [2]. With the national emphasis on data as a production factor and the continuous deepening of enterprise digital transformation, the market demand for data-related job talent is rapidly increasing. The "2023 Industrial Digital Talent Research and Development Report" [3] points out that the current overall shortage of digital composite talents is around 25 to 30 million, and the gap is still widening. For example, in the field of artificial intelligence, employers consider digital skills and application-oriented thinking as primary considerations. However, due to the insufficient depth of integration between industry and education, the long-standing mismatch between supply and demand persists.

DOI: 10.54254/2755-2721/64/20241424

To address this challenge and more clearly define the specific requirements of enterprises for various types of data-related talents, and to guide universities in cultivating talents that better meet market needs, this paper collected 31,718 data-related job postings from Zhaopin.com. Using the ChatGLM3 large language model and text clustering techniques, the study extracted key attributes from the collected recruitment texts. Based on these key attributes, K-means++ clustering method was used to cluster the positions, and further analyze the differences in demand characteristics between different categories of positions.

2. Related Research

Job requirement characteristics analysis, an important branch of the text mining field, has long attracted the attention and research of scholars both domestically and internationally. Since 2023, with the rapid rise of generative artificial intelligence technologies such as ChatGPT, their applications in the field of text mining have continued to expand and deepen. In light of this trend, this paper will comprehensively review the current state of research on the analysis of data-related job requirement characteristics, aiming to provide references for related research and practice.

Gu Yang et al. [4] collected 1,238 job postings for information analysts from the Zhaopin recruitment website, using statistical analysis methods to analyze the current social organization's demand structure for information analysts from four aspects: job responsibilities, skill requirements, background requirements, and organizational needs. Zhang Junfeng et al. [5] gathered 10,796 data from eight recruitment websites concerning data analysts, data mining engineers, and information administrators. They constructed a dictionary of skills and keywords for data-related job postings, analyzing domestic data-related job recruitment characteristics across five dimensions: capabilities, educational background and field, knowledge, tools, and experience. Wei Tingting et al. [6], collected recruitment information about data analysis positions from four recruitment websites, analyzing job characteristics in terms of region, salary, capabilities, specialization, and benefits using a keyword dictionary and K-means clustering. Kang Peng et al. [7] extracted recruitment information about big data analysis positions from the 51Job website, totaling 3,860 posts, employing dictionary-based Chinese word segmentation to analyze knowledge, tools, capabilities, and experience. Wang Yuchen [8] used "data analysis, data mining, data development, data operations" as search terms to collect 19,437 pieces of data from the 51Job website, using keyword extraction methods from a job dictionary and the LDA topic model to analyze demand characteristics across six dimensions: cities, job requirements, education, salary, industry, and skills.

From previous studies, the mainstream method for key attribute extraction involves constructing a custom dictionary, followed by Chinese word segmentation for analysis. However, this method demands a high level of completeness in the dictionary, which can affect the results of subsequent clustering experiments and analyses, and dictionary construction requires substantial time for manual selection and data cleaning. In terms of analysis dimensions, the chosen search terms primarily focus on specific positions like big data or data analyst, which do not fully describe the overall demand characteristics of data-related jobs. Furthermore, the selected attributes for analysis are mainly region, capabilities, specialization, salary, and skills, lacking comprehensiveness in job requirement descriptions. This paper synthesizes the analysis dimensions from previous research, selecting eight data-related terms including big data, data analysis, data development, data governance, data mining, data operations, data statistics, and data architecture as search terms. It analyzes six major aspects: soft skills, tools, fields of expertise, educational background, experience, and salary, providing a more comprehensive description of the demand characteristics of data-related positions. Additionally, this paper aims to explore the application of LLM in extracting key attributes from job descriptions from the perspective of job requirement analysis. It compares the effectiveness of key attribute extraction methods, including a dictionary-based TF-IDF method and an LLM-based method with prompting engineering, using a manually annotated dataset as a seed set.

DOI: 10.54254/2755-2721/64/20241424

3. Key Attribute Extraction of Data-related Job Descriptions Based on Large Models

3.1. Data Collection and Preprocessing

Considering the difficulty and consistency of data scraping, this study selected the Zhaopin recruitment website as the data source, covering the period from August 20 to October 13, 2023. Based on commonly used search terms on the Zhaopin website, key words such as big data, data analysis, data development, data governance, data mining, data operations, data statistics, and data architecture were selected for data extraction.

This study utilized Python web scraping technology and the Octopus Collector to automatically scrape information such as job titles, company names, locations, educational requirements, salary, experience, company type, company size, job descriptions, and industry sectors. By eliminating invalid URLs and repeatedly scraping unsuccessfully retrieved pages, the successful scraping rate was increased to 95%. Preliminary deduplication of recruitment data from 19 cities was conducted using SPSS statistical tools, based on the URL of the recruitment websites, resulting in a total of 31,718 unique recruitment data records after deduplication.

3.2. Key Attribute Extraction Method Based on Dictionary and TF-IDF

Traditional key attribute extraction methods include Term Frequency-Inverse Document Frequency (TF-IDF), TextRank, and topic modeling algorithms such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) [9]. The TF-IDF method, mainly used for feature extraction and feature weight calculation, serves as a parameter for subsequent clustering algorithms. Its principal idea is that if a term appears frequently in a particular type of document but rarely in others, it is considered to have good discriminatory power and is suitable for classification [10].

Therefore, this paper selected the TF-IDF-based key attribute extraction method available in Python's Jieba library as the traditional method. Additionally, to ensure the dictionary included comprehensive content, this study chose a computer and big data-related vocabulary from the Sogou Lexicon to construct a recruitment skill dictionary, adding frequently occurring terms from job requirements to the skill dictionary.

3.3. Key Attribute Extraction Method Based on Large Models

ChatGLM3, a dialogue pre-training model jointly released by Zhipu AI and Tsinghua University's KEG Laboratory, is notable for being an open-source model with fluid dialogue and low deployment thresholds. Therefore, this paper selected ChatGLM3-6B as the experimental model for the large model method.

Prompt engineering, as the process of designing and optimizing text inputs, provides consistent and high-quality responses for specified application goals and models, significantly influencing the outputs with different prompt modes. Current mainstream prompt engineering is categorized into five types: input semantics, output customization, error recognition, prompt improvement, and interaction [11]. This paper simplifies these categories into: basic prompts, refined basic prompts, zero-sample prompts, single-sample prompts, and few-sample prompts, as shown in Table 1:

DOI: 10.54254/2755-2721/64/20241424

Table 1. Example Types of Prompt Words

Prompt type	Input Example
Basic prompt	"The Sun is Shining". Translate to Spanish
Refined basic prompt	"The Sun is Shining". Translate to Spanish more colloquially
Hint prompts	“Role:You are a translator” "The Sun is Shining". Translate to Spanish more colloquially
One-shot prompt	"The Sun is Shining" =>""El Sol está brillando" "lt's a Cold and Windy Day" =>
Few-shots prompt	The player ran the bases => Basebal The player hit an ace => Tennis The player hit a six=> Cricket The player made a slam-dunk=>

Data collected for data-related job descriptions were combined with prompt words as model inputs. Considering the maximum token limits of the model, this paper used a for loop to clear history iteratively and repeated input of prompt words to avoid erroneous outputs.

3.4. Accuracy Testing of Key Attribute Extraction

To validate the accuracy of the model’s extractions, 500 job description data were selected as the test set, and key words were manually annotated for comparison with standard data using text similarity measures. Text similarity calculation methods primarily include string-based, word vector-based, pre-trained model-based, and deep learning-based methods [12]. Given the short text and sparse semantic information characteristics of job key attributes, this study used the string-based calculation method, employing Python’s difflib library SequenceMatcher function to calculate text similarity.

The rate of correctly returned key words as a percentage of the test set was used as an accuracy rate, and text similarity calculations were conducted between the data returned by the model and the manually annotated training set. The change in similarity between the large model extracted data and the traditionally extracted data was used as the primary indicator α; the harmonic mean of accuracy rate and text similarity was used as the second evaluation metric β.

By employing the aforementioned large model key word extraction method and conducting ablation experiments with improved prompt words, the outputs using different prompt words under the same input text were compared with traditional method results for text similarity, as shown in Table 2:

Table 2. Model Effectiveness Comparison

Model type	Number of keywords accuracy	Text similarity	Growth rate of text similarity	harmonic mean of accuracy and similarity,β=0.5
method based on dictionary	1	0.417	0	0.781
Method based on basic prompt	0.95	0.595	0.427	0.849
Method based on Refined basic prompt	0.975	0.598	0.434	0.856
Method based on hint prompt	0.98	0.642	0.541	0.887
Method based on one-shot prompt	0.98	0.619	0.483	0.878
Method based on few-shot prompt	0.982	0.631	0.512	0.884

As seen in Table 2, the experimental effects vary with different types of prompts, with the effects ranking as suggestive prompts > few-sample prompts > single-sample prompts > refined prompts >

DOI: 10.54254/2755-2721/64/20241424

basic prompts. All five key word extraction methods based on the large model outperformed the dictionary-based method, achieving an adjusted F2 value of 88.4%.

In subsequent empirical analysis, this paper utilized key attribute extraction methods based on LLM and suggestive prompt types from the prompt engineering, extracting key words from the initially collected 31,218 job description data.

4. Data Job Requirement Characteristics Analysis

4.1. Job Core Skill Word Clustering Based on k-means++

With the development and expansion of the big data industry, the variety and number of positions are continuously increasing. These positions may involve different business domains, skill requirements, and job responsibilities, making the relationships between positions very complex. To better manage and analyze these positions, it is necessary to cluster them.

K-means, as one of the common clustering algorithms, primarily relies on the Euclidean distance between samples and cluster centers to divide data. K-means++, an improvement over K-means, optimizes the random initialization of cluster centers in K-means [13]. Based on the principle of maximizing the distance between the initial cluster centers, it calculates the distance between a sample point and the first randomly selected cluster center, increasing the probability that points further from the current cluster center will become the next cluster center. Thus, K-means++ enhances clustering effectiveness without the need for random selection of cluster centers, improving computational precision [14]. Then, this paper constructs a TF-IDF text vector matrix between positions using key skills words extracted by the LLM and performs clustering analysis using the K-means++ algorithm. By calculating the average of the within-cluster sum of squares [15], K=4 was chosen as a suitable number of clustering categories. For ease of subsequent empirical analysis, the four categories obtained from clustering were consolidated into three categories: Data Engineering, Data Operations, and Data Analysis.

4.2. Competency Requirements

Table 3. Top 6 Soft Skills Frequency Words

Soft skill	Data Analysis	Data Operation	Data Engineering
Teamwork	35.76%	5.26%	31.51%
communication skills	24.61%	31.09%	12.79%
Stress resistance	11.88%	23.85%	11.42%
logical thinking	10.34%	7.24%	24.20%
responsibility	7.55%	9.87%	5.48%
learning ability	6.58%	13.16%	11.87%

As shown in Table 3, different categories of positions place varying emphasis on soft skills.

Data Analysis positions emphasize teamwork and communication, from data collection to analysis and reporting, with communication integral throughout. Analysts need to convey complex data in simple ways and adopt suggestions to optimize processes. Data Operations key competencies are communication, stress tolerance, and learning. It is necessary to translate data into business language, calmly face challenges, and continuously learn to adapt to changes in job requirements. Data Engineering positions emphasize teamwork and logical thinking abilities to ensure that multiple departments collaborate on data processing and draw accurate conclusions.

4.3. Tool Skills

Referencing the content analysis of big data recruitment ads by Gardiner [16] and others, Stojanović et al. [17], this paper categorizes tool skills into six types, using single-sample prompt inputs with ChatGLM3 for classifying tool skill words:

DOI: 10.54254/2755-2721/64/20241424

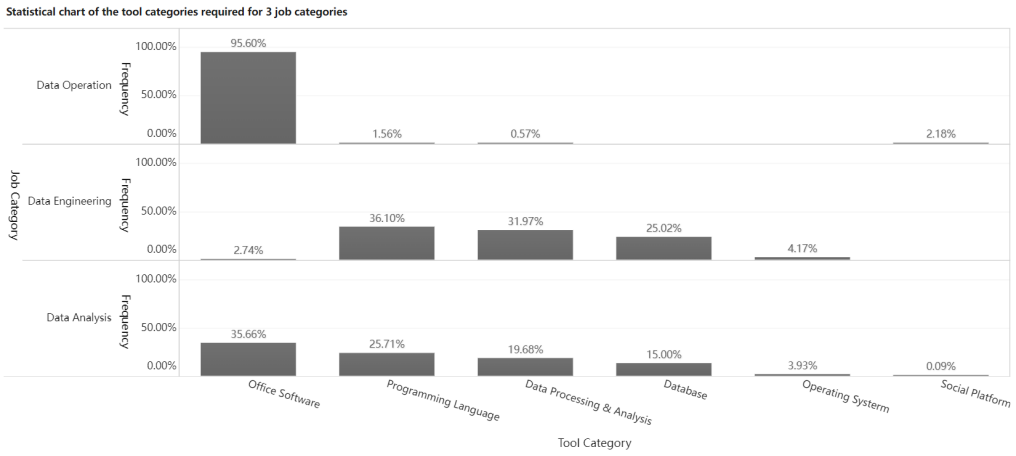

Figure 1. Tool Skills Bar Chart

Different types of positions require different tool skills:

Data Operations need tools like Photoshop and Premiere for visual design and video editing to visually attract users. Data Analysis requires mastery of tools such as Excel, SQL, and Python for data processing, analysis, and visualization, and familiarity with statistical tools like SPSS and SAS for data analysis and modeling to discover data patterns and trends. Data Engineering needs to master database and programming language tools to efficiently process data, use data visualization tools like Matplotlib to convey information, and apply data mining and machine learning technologies to build models that support business decisions and future directions.

4.4. Educational and Professional Requirements

4.4.1. Educational Background

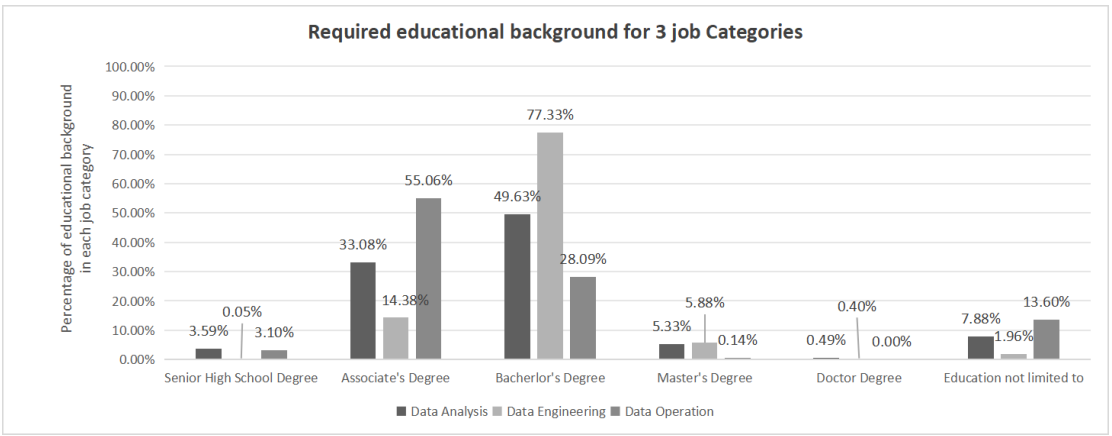

Figure 2. Educational Background Multi-Cluster Bar Chart

For ease of statistical analysis, this study categorizes educational levels into six categories. As shown in Figure 2, recruitment for the three job positions mainly focuses on junior college and bachelor's degree holders, with some positions not restricting the educational level of applicants. This indicates that the market prioritizes professional skills and work experience over academic research abilities for these positions. In data analysis and data engineering positions, the demand for bachelor's degree holders is highest. In data operations positions, due to less stringent requirements for hard skills, junior college and bachelor's degree holders constitute the majority.

DOI: 10.54254/2755-2721/64/20241424

4.4.2. *Specialties Required*

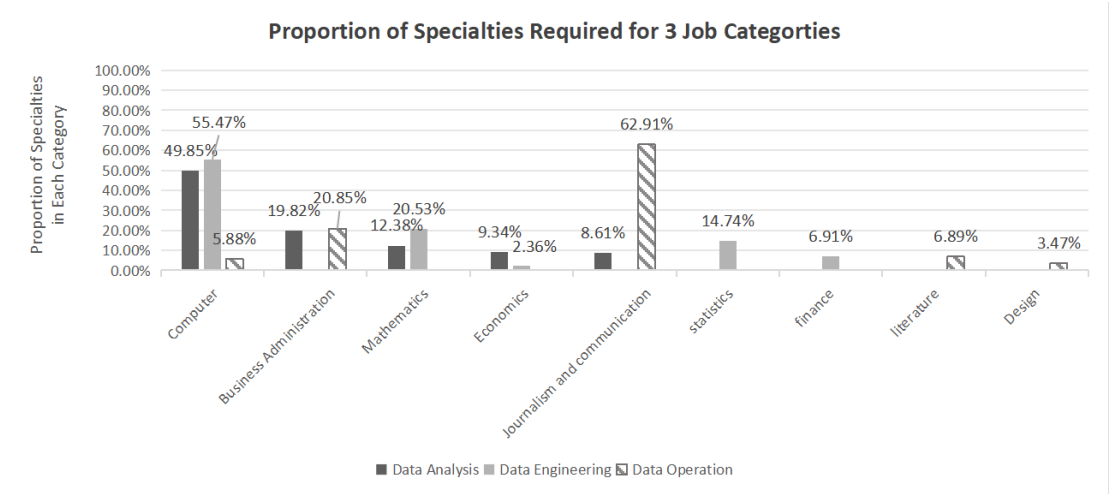

Figure 3. Specialties Required Multi-Cluster Bar Chart

Figure 3 shows that the required fields of study for data operations positions are broader and not concentrated in the science and engineering categories. In data analysis and data engineering positions, companies focus more on abilities in the computer science field, including programming skills, database management, and algorithm design, which are essential core skills for data analysis and processing. Given that data analysis involves extensive use of statistics and probability theory, learning mathematical knowledge is indispensable. This theoretical knowledge provides a solid foundation for data analysis. For data operations positions, companies value the ability to collect and feedback internet information. Currently, the internet is a crucial channel for acquiring market dynamics, user feedback, and competitor information. Operations personnel must have the ability to filter and integrate information efficiently and quickly to support company decision-making. Additionally, they should understand online promotion and public opinion management, use platforms for product promotion, respond swiftly to public sentiment, and maintain the company image.

4.5. *Work Experience*

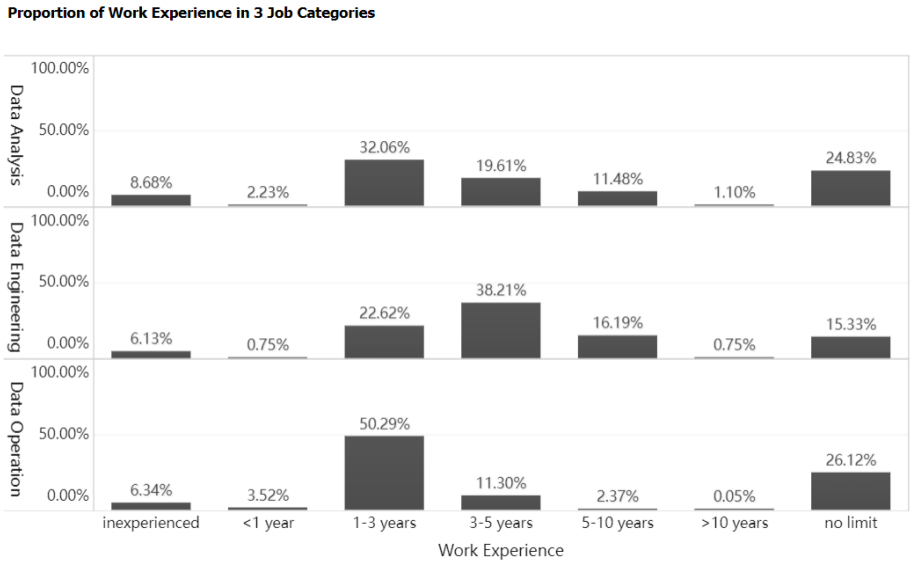

Figure 4. Experience Bar Chart

DOI: 10.54254/2755-2721/64/20241424

As shown in Figure 4, data analysis positions predominantly require 1-3 years of experience or are open to all levels of experience, accounting for more than half of the demand. Applicants are often inexperienced or have less than one year of qualifications, needing internships to gain experience and enhance competitiveness. Data engineering positions favor those with work experience, with the greatest demand for 3-5 years of experience, and very low demand for less than one year. Fresh graduates often lack experience and qualifications, making it challenging for them to assume positions directly. Data operations positions have the highest demand for 1-3 years of experience, with relatively low requirements for qualifications.

4.6. Salary

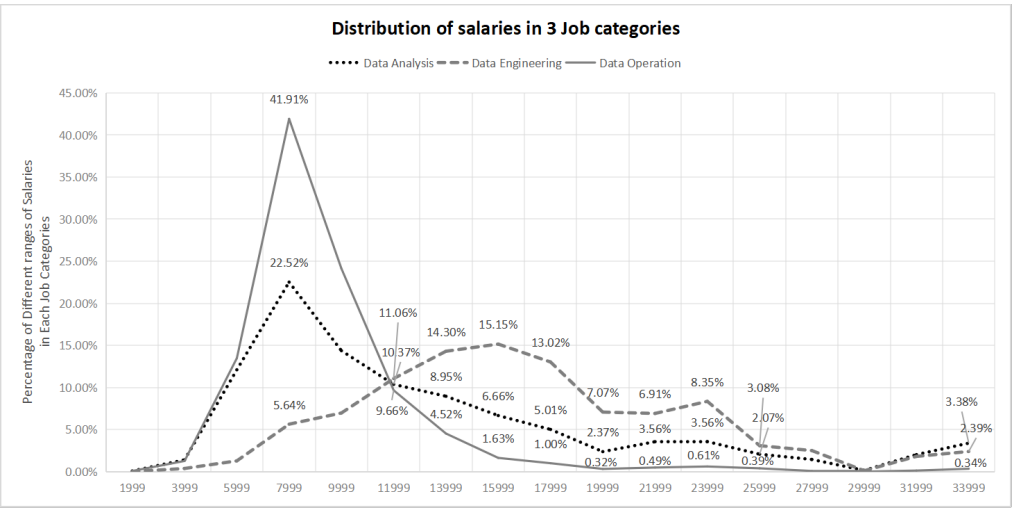

Figure 5. Salary Line Chart

Overall, salaries for data analysis and data operations positions are mostly concentrated around 7000, exhibiting a right-skewed normal distribution, indicating generally low overall pay, while data engineering positions have the highest average salaries. Moreover, data analysis positions offer a considerable chance and range for salary increases. Data engineering positions not only start at higher salaries but also offer greater potential for increases in the future. Data operations positions, however, have fewer opportunities for salary increases.

5. Conclusion

Data-related positions are categorized into three types: Data Analysis, Operations, and Engineering.

While there are similarities among these categories, there are also distinct differences. In terms of skill requirements, all emphasize communication, but each has its focus: Data Analysis and Data Engineering emphasize teamwork, whereas Data Operations prioritize stress resistance and learning ability. In terms of education, bachelor's degrees are common in Analysis and Engineering, while junior colleges are more prevalent in Operations. Regarding work experience and salary, Engineering positions have higher thresholds, requiring 3-5 years of experience with an average salary of 13,000 yuan; Analysis and Operations generally require 1-3 years of experience, with a salary around 8,000 yuan.

Based on these conclusions, this paper proposes the following recommendations for university majors, curriculum development, and student career planning:

(1) In-depth Thinking and Practice in University Majors

Refinement and precise customization of major directions: Universities should further refine data-related major directions, such as Data Analysis, Data Engineering, and Data Operations, to meet diversified market demands.

DOI: 10.54254/2755-2721/64/20241424

Promotion of interdisciplinary integration: Universities should encourage the integration of different majors such as Computer Science, Statistics, and Business Analysis.

Alignment with industry needs and dynamic adjustment of majors: Universities should closely monitor industry trends and understand market changes in demand for data talents.

(2) Innovation and Practice in Course Development

Strengthening the cultivation of core competencies: In curriculum design, universities should focus on cultivating core competencies. By offering courses on communication skills, competitions, or training projects, they can help students improve teamwork, effective communication, stress resistance, and learning abilities.

Introducing hands-on projects to strengthen practical skills: Allowing students to learn data handling, analysis, and application skills in practice, enhancing their ability to solve real-world problems.

Enhancing training in tool software: Universities should strengthen training in computer and statistical tool software. Through organizing practical operations, they can help students master common tools and improve work efficiency.

In summary, universities should focus on major settings, curriculum development, and student career planning in the cultivation of data talents. Through deep reflection and practice, they should continuously optimize talent training programs to adapt to the needs and changes of the data talent market. Additionally, universities should maintain close contact with partners such as enterprises and industry associations to jointly promote the innovation and development of data talent training.

References

[1] "Data Element Value Steadily Unleashed—Excerpt from the Data Element White Paper (2023)." (2024). Business Management, 2024(01), 46-52.

[2] Zhang, L. J. (2023). "The unveiling of the National Data Bureau: How to activate a trillion-yuan market?" China Report, 2023(12), 58-59.

[3] Renrui Human Resources & Deloitte China. (2023, March 17). "Industrial Digital Talent Research and Development Report (2023)." Retrieved from https://www.renruihr.com/web/news/detail?uid=20230317134128622U00000000003063

[4] Gu, Y., & Lü, B. (2018). "Research on the demand structure of domestic information analysts—Based on the mining analysis of recruitment website information." Intelligence Exploration, 2018(07), 41-47.

[5] Zhang, J. F. (2018). "Mining and application research on the demand characteristics of domestic website recruitment positions" [Doctoral dissertation, Anhui University of Finance and Economics].

[6] Wei, T. T., Fang, H. Y., Song, S. L., et al. (2019). "Recruitment characteristic mining of data analysis positions under the background of big data." Modern Computer, 2019(25), 14-17+27.

[7] Kang, P., Zang, J., & Cao, L. H. (2021). "Undergraduate teaching reform and practice under job demand matching—Taking the construction of big data analysis class course system as an example." Business Economy, 2021(04), 194-196. https://doi.org/10.19905/j.cnki.syjj1982.2021.04.069

[8] Wang, Y. C. (2023). "Analysis of talent demand in data-related positions based on text mining" [Doctoral dissertation, Lanzhou University of Finance and Economics]. https://doi.org/10.27732/d.cnki.gnzsx.2022.000296

[9] Zhao, J. L., Zhu, H., & Liu, X. (2019). "Research on the extraction of library knowledge group characteristics based on improved TFIDF." Systems Science and Mathematics, 2019, 39(09), 1450-1461.

[10] Tian, Ziyang, "A Comparative Study of Document Representation Methods" (2019). Electronic Theses and Dissertations. 8183.

[11] White, J., et al., A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. ArXiv, 2023. abs/2302.11382.

[12] Wei, W., Ding, X. X., & Guo, M. X., et al. (2024, February 21). "A review of text similarity calculation methods." Computer Engineering. https://doi.org/10.19678/j.issn.1000-3428.0068086
[13] Guan, S. Z., Tang, Y. B., Cai, Z., et al. (2023). "Ultra-short-term forecasting of solar irradiance based on Kmeans++-Bi-LSTM." Journal of Solar Energy, 2023, 44(12), 170-174. https://doi.org/10.19912/j.0254-0096.tynxb.2022-1294
[14] Yue, S., & Yong, Q. L. (2023). "An optimized K-means algorithm based on determining initial cluster centers." Digital Technology and Applications, 2023, 41(11), 140-142. https://doi.org/10.19695/j.cnki.cn12-1369.2023.11.44
[15] Debao, D., Y. Ma and Z. Min, Analysis of big data job requirements based on K-means text clustering in China. PLoS ONE, 2021. 16.
[16] Gardiner, A., et al., Skill Requirements in Big Data: A Content Analysis of Job Advertisements. Journal of Computer Information Systems, 2018. 58: p. 374 - 384.
[17] Stojanović, D., et al., Employer Requirements for Graduate Competencies in Applied Artificial Intelligence. 2023 16th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS), 2023: p. 299-303.