IC-015
Exploring Data Scientist Roles and Requirements via Topic Modeling : A Case Study of Thailand

Sompong Promsa-ad[1, *]
[1]Faculty of Economics and Business Administration, Thaksin University, Songkhla, Thailand
[*]Corresponding author's email: psompong@tsu.ac.th

ABSTRACT

The growing demand for data scientists in the expanding field of data science has led to an evolving role for these professionals, presenting challenges in defining their responsibilities. This study investigates the changing roles and requirements of data scientists in Thailand's job market. We analyzed 108 job postings for data scientist positions in Thailand from March and April 2024, sourced from prominent job sites including Indeed, Glassdoor, LinkedIn, Jobsdb, and Jobtopgun. Using Latent Dirichlet Allocation (LDA) topic modeling, we identified nine themes of roles and two themes of qualifications for data scientists. Our findings suggest that the new generation of data scientists is expected to collaborate closely with team members and stay updated on emerging technologies, particularly Large Language Models and Artificial Intelligence. With the market favoring data scientists with formal training, there are significant opportunities for higher education institutions to develop curricula focusing on hands-on projects and analytical skills development.

Keywords: data scientist, roles, qualifications, topic modeling

Introduction

Data science is an emerging field that encompasses various disciplines, including programming, statistics, and domain knowledge. The significance of data science is rapidly increasing as the proliferation of digital activities continues to generate vast amounts of data (Salloum et al., 2021). Although there is currently no universally agreed-upon definition, the term could be described as the tool to extract value or insights from data (Ahmad & Hamid, 2023). The data science process typically starts with identifying the business problem, acquiring data, preparing data, conducting exploratory data analysis, building data models, visualizing and communicating findings, and deploying and maintaining solutions (Inkeaw, 2021). The rise of data science field leads to a strong demand for a data scientist, responsible for identifying problems, leveraging available data, and transforming them into opportunities for value creation (Vicario & Coleman, 2020). This profession is also considered as one of the sexiest jobs of the 21st century (Harvard Business Review, 2012).

Currently, however, the role of a data scientist has become ambiguous due to multiple reasons. For example, job roles often lack clarity as data scientists are expected to perform a wide range of tasks, including statistics, programming, and domain-specific knowledge (Gutman & Goldmeier, 2021). In addition, difficulty often arises in distinguishing the roles of data scientists from related concepts due to their overlapping responsibilities. For example, confusion frequently occurs between data science and data engineering. The latter involves designing, constructing, and maintaining systems that enable efficient data collection, storage, and processing, essential for effective data analysis (Pierson, 2021). Moreover, the field of data science is currently experiencing a surge in Generative AI, which can effectively perform various tasks, especially for unstructured data tasks such as sentiment analysis and text classification. (Hassani & Silva, 2023).

Literature attempting to provide clarity on these emerging job roles is still limited. Michalczyk et al. (2021) characterized six roles in data science: the business user, the data analyst, the data scientist, the data engineer, the software architect, and the software developer. The role of a data scientist relies heavily on skills in statistics, analytical decision-making, business intelligence, big data technology, and computer science conceptual understanding, with related work experience being a preferred qualification. Goretzki et al. (2023) examined the professional identity development of data scientists and the challenges they face in applying it. Their findings, based on semi-structured interviews, identified three main components of data scientists' occupational identity: a scientific mindset, an affinity for complex data tasks, and a problem-solving approach.

In Thailand, the significance of data science is pivotal in driving the country's digital transformation initiative, with a persistent demand projected for data scientists across various sectors in the foreseeable future (Deloitte, 2020). However, the study of the responsibilities and qualifications required for the job, or other aspects, is also in its infancy stage. In their study focusing on identifying critical technical and soft skills necessary for software development roles, Hiranrat & Harncharnchai (2018) grouped data scientist and data analyst positions together due to their perceived similarity in responsibilities. Moreover, the study's sample size for data scientist positions was limited to only 31, potentially affecting the inclusivity of the results.

Given the rapidly evolving nature of data scientist roles and the limited existing research in Thailand, this study aims to bridge the gap by examining the roles and requirements of data scientists in the country using topic modeling. The hypothesis is that the current traits of desired data scientists differ from those specified in the past. In addition, the topic modeling proves to be a valuable instrument for uncovering concealed insights within job listings, enabling the identification of distinct professional profiles based on their skills and competencies. For example, Culasso et al. (2023) utilized Latent Dirichlet Allocation (LDA) topic modeling to discern the duties and prerequisites associated with the Chief Digital Officer (CDO) role. It's worth to highlight that this method finds widespread application in exploring emerging job roles, as demonstrated by Chunmian et al. (2022) in their examination of the demand for blockchain talent, and by Madding et al. (2020) in their analysis of the demand for technology professionals.

Purposes

To identify roles and requirements of data scientists in Thailand's job market.

Research tools

Data source

The sample comprised 108 job postings for data scientist positions in Thailand from March and April 2024. Data were sourced from leading job sites, including Indeed, Glassdoor, LinkedIn, Jobsdb, and Jobtopgun. Only job postings directly related to the position of data scientist were included in the study. The collected data were divided into two sets: one for role/responsibilities and another for qualifications/requirements. These data were stored in an Excel file for processing in the next stage.

Research Process

In Latent Dirichlet Allocation (LDA) topic modeling, it is assumed that each document is a mixture of topics, and each topic is a probability distribution over words in the vocabulary. To perform LDA, raw data were processed following these steps.

1) Data Preprocessing

In the data preprocessing steps, the input text is cleaned to enhance the quality of input for the Latent Dirichlet Allocation (LDA) model, improving the accuracy and interpretability of the results. The process begins by tokenizing the text using NLTK's tokenize function, which splits the text into individual words. Punctuation is then removed to prevent interference with analysis. Common stop words, such as 'is', 'the', and 'and', are eliminated to reduce noise. Finally, each token is lemmatized using NLTK's WordNetLemmatizer, reducing word dimensionality and capturing core meanings.

2) Data Preparation

To prepare the input for LDA model, a dictionary is created from the tokenized text data. This dictionary maps each unique token (word) to a unique numerical ID. In the next step, each document's word frequencies are counted and transformed into a bag-of-words representation using the dictionary created in the first step. This representation captures how frequently each word appears in each document. The resulting corpus is a list of lists, where each inner list contains tuples representing word IDs and their frequencies in a particular document.

3) Building the LDA model

The next step involves implementing an LDA model using the specified parameters and the preprocessed data with the Gensim library. This step aims to identify topics and their associated top words, along with their probabilities. The determination of the optimal number of topics in the model is guided by assessing Coherence score, which evaluates the semantic coherence among highly ranked words within each topic. This metric aids in selecting the number of topics that yield the most cohesive and interpretable results.

Results

Before performing the LDA model, coherence scores were calculated for different numbers of topics, and the optimal model is determined by the highest coherence score. For the LDA model of roles/responsibility data, the highest score was 0.38 when the number of topics was nine. For the LDA model of qualifications/requirements, the highest score was 0.46 when the number of topics was two.

Table1: Distribution of topics along with the associated keywords for each topic representing the roles or responsibilities of data scientists

Topic	**Top ten keywords**
Topic 1	network (0.043), model (0.032), chain (0.025), supply (0.025), scenario (0.020), optimization (0.016), define (0.014), run (0.014), business (0.011), standard (0.009)
Topic 2	risk (0.001), credit (0.001), business (0.001), engine (0.001), including (0.001), system (0.001), portfolio (0.001), rule (0.001), strategy (0.001), requirement (0.001)
Topic 3	data (0.028), business (0.027), team (0.015), risk (0.009), analytic (0.009), analytics (0.008), model (0.007), technique (0.007), customer (0.007), support (0.007)
Topic 4	data (0.019), model (0.016), ai (0.014), business (0.011), insight (0.011), create (0.011), valuable (0.011), application (0.011), llm (0.008), collaborate (0.008)
Topic 5	data (0.054), business (0.031), team (0.016), project (0.015), management (0.013), analytics (0.010), product (0.010), information (0.010), learning (0.010), develop (0.009)
Topic 6	data (0.039), model (0.021), team (0.009), solution (0.010), machine (0.009), learning (0.009), scientist (0.009), engineer (0.008), customer (0.007), build (0.007)
Topic 7	data (0.039), business (0.015), model (0.013), team (0.011), project (0.011), new (0.010), client (0.009), develop (0.009), machine (0.008), analytics (0.008)
Topic 8	data (0.043), model (0.031), business (0.025), learning (0.014), machine (0.013), analysis (0.012), solution (0.011), visualization (0.009), stakeholder (0.008), complex (0.008)
Topic 9	data (0.043), model (0.023), business (0.018), learning (0.014), team (0.014), work (0.013), machine (0.013), solution (0.012), develop (0.012), implement (0.009)

Table 1 illustrated topics and associated keywords representing the roles or responsibilities of data scientists. The first topic focuses on supply chain management optimization, including scenario planning and defining standard procedures. The second topic involves risk management and financial aspects, such as setting portfolios and strategies. The third topic centers around business analytics, emphasizing the analysis of business data and the utilization of various techniques and models. The fourth topic pertains to collaborative work in utilizing AI applications for business insights, with mentions of 'ai,' 'insight,' 'application,' and 'valuable,' along with the use of large language models (LLM).

The fifth topic focuses on data-driven project management and analytics, with prominent keywords like 'data,' 'business,' 'project,' 'management,' and 'analytics'. The sixth topic pertains to the role of a machine learning engineer, involving activities centered around building and deploying machine learning solutions or systems. The presence of the word 'new' in the seventh role suggests a focus on developing innovative solutions, exploring emerging ideas, or addressing evolving challenges within the field. In a business or project development context, this may indicate involvement in new initiatives, projects, or ventures. This could encompass launching new products, services, or business lines, or undertaking innovative projects to drive growth and gain a competitive advantage. The presence of 'visualization,' 'stakeholder,' and 'complex' in the eighth topic suggests a focus on effectively communicating complex ideas or insights to relevant stakeholders through visualization. The last topic signifies collaboration within a team environment, reflecting a collective effort to achieve common goals or objectives related to data-driven projects or initiatives, as indicated by the inclusion of terms such as 'work' and 'team'.

Table 2: Distribution of topics along with the associated keywords for each topic representing the qualifications or requirements of data scientists

Topic	**Top ten keywords**
Topic 1	experience (0.033), data (0.032), science (0.017), learning (0.012), skill (0.011), python (0.011), machine (0.009), strong (0.009), computer (0.009), knowledge (0.009)
Topic 2	experience (0.031), data (0.028), skill (0.015), learning (0.015), computer (0.013), strong (0.012), machine (0.012), science (0.011), degree (0.010), analysis (0.009)

Table 2 illustrated topics and associated keywords representing the qualifications or requirements of data scientists. The first topic focuses on technical skills and qualifications, with terms such as 'experience', 'data', 'science', and 'python', highlighting the importance of expertise in data science methodologies and programming, particularly in Python. Additional terms like 'learning', 'skill', 'strong', 'computer', and 'knowledge' underscore the significance of robust technical skills and domain expertise. In the second topic, the presence of 'degree' and 'analysis' suggests an emphasis on educational qualifications and analytical skills, often crucial for data science roles.

Discussion

The variation in the number of themes between job qualifications and responsibilities may stem from the diverse nature of the data analyzed. Job responsibilities often encompass a wide array of tasks, leading to a greater diversity of themes as each employer prioritizes different aspects of the role. In contrast, job qualifications tend to be more standardized across postings, resulting in fewer distinct themes as employers typically list similar requirements, such as programming skills and statistical knowledge. Notably, the presence of themes like 'supply chain' and 'finance' underscores the strong current demand for data scientists in these industries.

The findings indicate an evolution in the role of data scientists. Specifically, they are no longer expected to perform all tasks individually but rather to work collaboratively as part of a team. Their role now involves collaborating with others to develop models and derive valuable insights or solutions from data. This trend contrasts with the past, where data scientists were expected to handle a wide range of tasks (Gutman & Goldmeier, 2021). In addition, with the current surge in Generative AI (Hassani & Silva, 2023), data scientists are expected to utilize Large Language Models and other forms of Artificial Intelligence to enhance their data-driven solutions. On the qualifications/requirements side, the presence of the theme 'degree' indicates the importance of formal education for those aspiring to enter the data science profession.

Conclusions

The rise of the data science field has led to a strong demand for data scientists, but their role has become increasingly ambiguous due to various factors. This study aims to identify the evolving roles and requirements of data scientists in Thailand's job market, guiding educational institutions and governmental agencies in tailoring curricula and policies to nurture highly skilled professionals. LDA results revealed nine themes of roles and two themes of qualifications for data scientists.

The findings indicated that while traditional expectations such as creating data science or machine learning models and communicating insights to stakeholders persist, data scientists are now expected to collaborate closely with team members and adeptly apply artificial intelligence, particularly Large Language Models. In preferred qualifications for data scientists, the job market prioritizes expertise in data science methodologies and programming, especially in Python, along with strong analytical skills. Despite the abundance of online and short course training options for data scientists, a solid formal education background, such as a degree in computer science, is considered advantageous.

Recommendations

Based on the findings about roles and qualifications, this study recommends an approach to train desired data science professionals. First, although various training courses are available, there is an opportunity for higher education to offer a program in data science, as the industry believes that formal education with degrees will equip data scientists with strong fundamental knowledge and skills. Second, the curriculum should focus on hands-on projects to provide students with practical experience and help them build portfolios. Third, the curriculum should emphasize fostering analytical skills and collaborative abilities. Lastly, students should be equipped with rapidly emerging knowledge in applying artificial intelligence in various job settings, particularly with Large Language Models.

For future research, increasing the sample size in LDA can yield more robust, accurate, and generalizable results, enhancing the utility of topic modeling analysis. Additionally, comparing the roles and requirements of related data job positions, such as data analysts or data engineers, would be an intriguing and promising research avenue.

References

Ahmad, N., & Hamid, A. (2023). Will Data Science Outrun the Data Scientist?. *Computer*, 56(2), 121-128.

Chunmian, G. E., Haoyue, S. H. I., Jiang, J., & Xiaoying, X. U. (2022). Investigating the demand for blockchain talents in the recruitment market: evidence from topic modeling analysis on job postings. *Information & Management*, 59(7), 103513.

Culasso, F., Gavurova, B., Crocco, E., & Giacosa, E. (2023). Empirical identification of the chief digital officer role: A latent Dirichlet allocation approach. *Journal of Business Research*, 154, 113301.

Deloitte. (2020). The Thailand digital transformation survey report 2020. https://www2.deloitte.com/content/dam/Deloitte/th/Documents/technology/th-tech-the-thailand-digital-transformation-report.pdf.

Earnhart, C. L. (2018). Evaluating an on-line education module for autism screening in pediatric primary care in Arizona [Doctoral dissertation, University of Arizona]. ProQuest Nursing & Allied Health Database. https://search.proquest.com/docview/2160956827?accountid=34902

Harvard Business Review. (2012). Data Scientist: The Sexiest Job of the 21st Century. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-

Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. Big data and cognitive computing, 7(2), 62.

Hiranrat, C., & Harncharnchai, A. (2018). Using text mining to discover skills demanded in software development jobs in Thailand. In Proceedings of the 2nd international conference on education and multimedia technology, 112-116.

Goretzki, L., Messner, M., & Wurm, M. (2023). Magicians, unicorns or data cleaners? Exploring the identity narratives and work experiences of data scientists. Accounting, Auditing & Accountability Journal, 36(9), 253-280.

Gutman, A. J., & Goldmeier, J. (2021). Becoming a data head: How to think, speak, and understand data science, statistics, and machine learning. John Wiley & Sons.

Inkeaw, P. (2021). Introduction to Data Science. Chiang Mai University.

Madding, C., Ansari, A., Ballenger, C., & Thota, A. (2020). Topic Modeling to Understand Technology Talent. *SMU Data Science Review*, 3(2), 16.

Michalczyk, Sven; Nadj, Mario; Maedche, Alexander; and Gröger, Christoph (2021). Demystifying Job Roles in Data Science: A Text Mining Approach. ECIS 2021 Research Papers. 115. https://aisel.aisnet.org/ecis2021_rp/115

Pierson, L. (2021). Data science for dummies. John Wiley & Sons.

Plows, J. F., Stanley, J. L., Baker, P. N., Reynolds, C. M., & Vickers, M. H. (2018). The pathophysiology of gestational diabetes mellitus. *International journal of molecular sciences, 19*(11), 3342. https://doi.org/10.3390/ijms19113342

Salloum, M., Jeske, D., Ma, W., Papalexakis, V., Shelton, C., Tsotras, V., & Zhou, S. (2021). Developing an interdisciplinary data science program. In Proceedings of the 52nd ACM Technical Symposium on Computer Science Education, 509-515.

Vicario, G., & Coleman, S. (2020). A review of data science in business and industry and a future view. *Applied Stochastic Models in Business and Industry*, 36(1), 6-18.