

# DATA-RELATED JOB ANALYSIS USING K-MEANS ALGORITHM

CHANG QITING

UNIVERSITI TEKNOLOGI MALAYSIA



**UNIVERSITI TEKNOLOGI MALAYSIA****DECLARATION OF THESIS / POSTGRADUATE PROJECT REPORT AND  
COPYRIGHT**

Author's full name : CHANG QITING

Date of Birth : 25 OCTOBER 1998

Title : DATA-RELATED JOB ANALYSIS USING K-MEANS  
ALGORITHM

Academic Session :

I declare that this thesis is classified as:

☐**CONFIDENTIAL**(Contains confidential information under the  
Official Secret Act 1972)\*☐**RESTRICTED**(Contains restricted information as specified by  
the organization where research was done)\*☒**OPEN ACCESS**I agree that my thesis to be published as online  
open access (full text)

1. I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:
2. The thesis is the property of Universiti Teknologi Malaysia
3. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
4. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

\_\_\_\_\_  
**SIGNATURE OF STUDENT**\_\_\_\_\_  
**SIGNATURE OF SUPERVISOR**\_\_\_\_\_  
**MATRIC NUMBER**\_\_\_\_\_  
**NAME OF SUPERVISOR**

Date: 5 JAN 2024

Date: 5 JAN 2024

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction



“I hereby declare that I have read this thesis and in my  
opinion this thesis is sufficient in term of scope and quality for the  
award of the degree of Master of Data Science”

Signature : \_\_\_\_\_  
Name of Supervisor I : Dr Chan Weng Howe  
Date : 5 JANUARY 2024



## **BAHAGIAN A - Pengesahan Kerjasama\***

Adalah disahkan bahawa projek penyelidikan tesis ini telah dilaksanakan melalui kerjasama antara \_\_\_\_\_ dengan \_\_\_\_\_

Disahkan oleh:

Tandatangan :

Tarikh :

Nama :

Jawatan :

(Cop rasmi)

*\* Jika penyediaan tesis atau projek melibatkan kerjasama.*

---

---

## **BAHAGIAN B - Untuk Kegunaan Pejabat Sekolah Pengajian Siswazah**

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar :

Nama dan Alamat Pemeriksa Dalam :

Nama Penyelia Lain (jika ada) :

Disahkan oleh Timbalan Pendaftar di SPS:

Tandatangan :

Tarikh : 5 JAN 2024

Nama :





# DATA-RELATED JOB ANALYSIS USING K-MEANS ALGORITHM

CHANG QITING

A thesis submitted in fulfilment of the  
requirements for the award of the degree of  
Master of Data Science

Faculty of Computing  
Universiti Teknologi Malaysia

January 2024



## DECLARATION

I declare that this thesis entitled “*DATA-RELATED JOB ANALYSIS USING K-MEANS ALGORITHM*” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : .....  
Name : CHANG QITING  
Date : 5 JANUARY 2024

## **DEDICATION**

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have been that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished if it is done one step at a time.

## **ACKNOWLEDGEMENT**

In preparing this thesis, I am very much indebted to my Supervisor Dr Chan Weng Howe and many kind fellow postgraduate student as I have received much useful advice from them.

## **ABSTRACT**

As the global demand for data scientists and data engineers keeps growing, failure to fill job vacancies adequately and timely results in organizational costs and data talent waste, while data-related job seekers are facing the obstacle that hard to identify helpful information from the massive job postings. This thesis aims to identify the skills groups and the most required skills of data-related jobs, by developing the clustering analysis for data-related job skills in Malaysia based on the job postings from JobStreet.com. The collected dataset of job postings is from 18 February 2023 to 18 May 2023, has 15504 records and 6 features including job title, location, job description, career level, experience and job type. The thesis will be discussed in 5 phases whereby phase 1 will focus on problem formulation and the significance of research via literature reviews of relevant studies. Phase 2 will be the data collection to apply the web scraper and gather the information on data-related job postings. Phase 3 is where data pre-processing will be conducted, the collected textual data will be cleaned, extracted by an Artificial Intelligence (AI) model and vectorized as the numerical value in the stage. Phase 4 will conduct the K-means model on the processed data to perform clustering. Phase 5 will discuss the research outcomes, the performance of the model will be measured by the silhouette coefficient and Davies Bouldin index (DBI), where a higher silhouette coefficient between -1 and 1 indicates that the model has more coherent clusters and a lower DBI value indicates a better clustering performance. The insight of the clustering result will be illustrated by the dashboard, which will indicate the groups of job skills and related job information, to help job seekers identify the trend and manifesting technologies of data-related jobs in Malaysia.

## ABSTRAK

Memandangkan permintaan global untuk saintis data dan jurutera data terus berkembang, kegagalan untuk mengisi kekosongan kerja dengan secukupnya dan tepat pada masanya mengakibatkan kos organisasi dan pembaziran bakat data, manakala pencari kerja yang berkaitan dengan data menghadapi halangan yang sukar untuk mengenal pasti maklumat berguna daripada kerja besar-besaran. pengesanan. Tesis ini bertujuan untuk mengenal pasti kumpulan kemahiran dan kemahiran yang paling diperlukan dalam industri berkaitan data, dengan membangunkan analisis pengelompokan untuk kemahiran kerja berkaitan data di Malaysia berdasarkan siaran pekerjaan daripada JobStreet.com. Set data penyiaran kerja yang dikumpul adalah dari 18 Februari 2023 hingga 18 Mei 2023, mempunyai 15504 rekod dan 6 ciri termasuk jawatan, lokasi, perihalan kerja, tahap kerjaya, pengalaman dan jenis pekerjaan. Tesis ini akan dibincangkan dalam 5 fasa di mana fasa 1 akan memberi tumpuan kepada perumusan masalah dan kepentingan penyelidikan melalui tinjauan literatur kajian yang berkaitan. Fasa 2 akan menjadi pengumpulan data untuk menggunakan pengikis web dan mengumpulkan maklumat tentang penyiaran kerja berkaitan data. Fasa 3 ialah di mana pra-pemprosesan data akan dijalankan, data tekstual yang dikumpul akan dibersihkan, diubah oleh model Kepintaran Buatan (AI) dan divektorkan sebagai nilai berangka dalam peringkat. Fasa 4 akan menjalankan model K-means pada data yang diproses untuk melaksanakan pengelompokan. Fasa 5 akan membincangkan hasil penyelidikan dan prestasi model akan diukur dengan pekali siluet dan indeks Davies Bouldin (DBI), di mana pekali siluet yang lebih tinggi antara -1 dan 1 menunjukkan bahawa model mempunyai kelompok yang lebih koheren dan DBI yang lebih rendah. nilai menunjukkan prestasi pengelompokan yang lebih baik. Wawasan hasil pengelompokan akan diilustrasikan oleh papan pemuka, yang akan menunjukkan kumpulan kemahiran pekerjaan dan maklumat pekerjaan yang berkaitan, untuk membantu pencari kerja mengenal pasti arah aliran dan mewujudkan teknologi dalam industri berkaitan data di Malaysia.

## TABLE OF CONTENTS

	<b>TITLE</b>	<b>PAGE</b>
	<b>DECLARATION</b>	<b>iii</b>
	<b>DEDICATION</b>	<b>iv</b>
	<b>ACKNOWLEDGEMENT</b>	<b>v</b>
	<b>ABSTRACT</b>	<b>vi</b>
	<b>ABSTRAK</b>	<b>vii</b>
	<b>TABLE OF CONTENTS</b>	<b>viii</b>
	<b>LIST OF TABLES</b>	<b>xi</b>
	<b>LIST OF FIGURES</b>	<b>xii</b>
	<b>LIST OF ABBREVIATIONS</b>	<b>xiv</b>
	<b>LIST OF APPENDICES</b>	<b>xv</b>
<b>CHAPTER 1</b>	<b>INTRODUCTION</b>	<b>16</b>
1.1	Introduction	16
1.2	Problem Background	17
1.3	Problem Statement	19
1.4	Research Questions	19
1.5	Aims and Objectives	19
1.6	Scope	20
1.7	Significance of Research	20
<b>CHAPTER 2</b>	<b>LITERATURE REVIEW</b>	<b>21</b>
2.1	Data Talent in Labor Market	21
2.2	Difficulties in Recruitment	22
2.2.1	Difficulties for Employers	22
2.2.2	Difficulties for Job Seekers	23
2.3	Job Postings	23
2.4	Content Analysis in Job Posting	24
2.5	Text Mining in Job Posting	25



2.5.1	Pre-processing of Job Posting	25
2.5.2	Text Vectorization	27
2.5.3	Machine Learning in Job Analysis	30
2.6	Challenges and Opportunities	33
<b>CHAPTER 3</b>	<b>RESEARCH METHODOLOGY</b>	<b>35</b>
3.1	Research Operational Framework	35
3.1.1	Phase 1: Problem Formulation	37
3.1.2	Phase 2: Data Collection	37
3.1.3	Phase 3: Data Pre-processing	38
3.1.4	Phase 4: Modeling	39
3.1.5	Phase 5: Outcomes Discussion	40
3.2	Performance Measure	40
3.3	Summary	41
<b>CHAPTER 4</b>	<b>RESEARCH DESIGN AND IMPLEMENTATION</b>	<b>42</b>
4.1	Data Preparation	42
4.1.1	Data Collection	44
4.1.2	Data Cleaning	50
4.1.3	Data Extraction	51
4.1.4	Exploratory Data Analysis	54
4.1.5	Data Vectorization	59
4.2	K-means Model Development	60
4.2.1	Principal Component Analysis	62
4.2.2	Elbow Method	63
4.2.3	Generate K-means based on selected k	64
4.2.4	Model Evaluation	64
4.3	Summary	64
<b>CHAPTER 5</b>	<b>OUTCOMES DISCUSSION</b>	<b>65</b>
5.1	Introduction	65
5.2	Performance Evaluation	65
5.3	Result Analysis	65

5.3.1	Result of Initial Clusters	68
5.3.2	Validation of Words in Each Cluster	68
5.3.3	Result of Final Clusters	70
5.4	Insight by Dashboard	73
5.5	Summary	75
<b>CHAPTER 6</b>	<b>CONCLUSION</b>	<b>76</b>
6.1	Introduction	76
6.2	Achievements of Project Objectives	76
6.3	Project Limitation	76
6.4	Future Work	77
<b>REFERENCES</b>		<b>79</b>

## LIST OF TABLES

TABLE NO.	TITLE	PAGE
Table 2.1	Vectorization Methods Comparison	29
Table 4.1	Description of Data Source	44
Table 4.2	Feature Description	48
Table 4.3	Data Extraction Example	54
Table 5.2	Skill Validation	69
Table 5.3	Most common skills in cluster	72

## LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
Figure 2.1	One-Hot Encoding Example	27
Figure 2.2	Word2Vec Example	29
Figure 2.3	Jothi's Research Workflow	32
Figure 3.1	Operational Framework of Job Posting Analysis	36
Figure 3.2	Data Collection	38
Figure 3.3	Data Pre-processing Stages	38
Figure 3.4	K-means Clustering	39
Figure 4.1	Data Preparation Flow	43
Figure 4.2	JobStreet.com Website	45
Figure 4.3	Search Results	46
Figure 4.4	Selection of Job Features	47
Figure 4.5	First 5 Samples of Raw Dataset	49
Figure 4.6	Elementary Analysis Result	49
Figure 4.7	Data Cleaning Flow	50
Figure 4.8	Data Cleaning Result	51
Figure 4.9	Data Extraction Flow	52
Figure 4.10	API Usage	53
Figure 4.11	Job Titles Distribution	55
Figure 4.12	Top Frequent Job Locations	56
Figure 4.13	Career Level Distribution	57
Figure 4.14	Required Experience Distribution	57
Figure 4.15	Job Type Distribution	58
Figure 4.16	Required Skills Example	58
Figure 4.17	Skill Phrases Overview	59
Figure 4.18	Vector Example	60

Figure 4.19	K-means Model Development	61
Figure 4.20	PCA Performance (k=3)	63
Figure 4.21	Elbow Method	63
Figure 5.1	Data Points Distribution	65
Figure 5.2	First 15 Samples of Clustering Distribution	67
Figure 5.3	Initial Observation of Skillset Distribution	68
Figure 5.4	Skillset Distribution of Each Cluster (validated)	70
Figure 5.5	Insight of Job Information (Cluster 2)	74
Figure 6.1	Generalized Analysis Example	78

## LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
API	-	Application Programming Interface
CSV	-	Comma-Separated Values
DBI	-	Davies Bouldin index
GPT-3	-	Generative Pre-Trained Transformer-3
HTML	-	HyperText Markup Language
idf	-	inverse document frequency
IoT	-	Internet of Things
JSON	-	JavaScript Object Notation
KNN	-	k-nearest neighbours
LDA	-	Latent Dirichlet Allocation
LLMs	-	Large Language Models
ML	-	Machine Learning
NLP	-	Natural Language Processing
NMF	-	Nonnegative Matrix Factorization
PCA	-	Principal Component Analysis
SVD	-	Singular Value Decomposition
tf	-	term frequency

## LIST OF APPENDICES

APPENDIX	TITLE	PAGE
Appendix A	Gantt Chart	84
Appendix B	Code	85
Appendix C	Dashboard of Clusters	86

## **CHAPTER 1**

### **INTRODUCTION**

#### **1.1 Introduction**

Today, the vigorous development and implementation of new technologies such as Big Data analytics, the Internet of Things (IoT), and Artificial Intelligence (AI), have a significant impact on various fields of society and industries (Grüger & Schneider, 2019). For instance, the ability to apply specific technical and analytical approaches to optimize the value of high-volume and high-velocity data assets has a prominent position on the business managers' agenda. And it has been proved that for high technology-intensive firms, the employee's experience and skills determine the success of the firms (De Mauro et al., 2018). Therefore, to improve competitiveness quickly, firms are racing to acquire data talents. In the past decade, technology companies have demanded a large number of data-related workers with skill-mix requirements. However, the labour market absents adequate countermeasures to fit this impact. Companies are still in great demand for data analysts and data scientists, while for job seekers of data specialists, there is a gap between job positions and their knowledge. Not only due to the nature that data scientist is a term that roughly describes a complex set of interrelated skills (De Mauro et al., 2018), but also the vague and subjective description in the job postings. Especially for graduates who major in data science and lack practical experience in companies, how to locate themselves and find a matched job becomes an essential obstacle. Possible approaches are to manually search online job postings or review job advertisements offered by the recommendation system, in either case, it will be very time-consuming and tedious (Grüger & Schneider, 2019). This thesis introduces a solution to investigate the requirements of jobs in Malaysian data-related industries, which based on the scraped job information from Malaysian recruitment websites with the implement of AI technology and unsupervised Machine Learning (ML). This thesis will provide a reference of data-related jobs for job seekers in Malaysia, which could assist them in



locating themselves and rationalizing the acquisition of skill sets. In addition, this may contribute to the development of the data industries in cultivating a high-quality workforce.

## **1.2 Problem Background**

With the rapid progress in computer software, hardware and network, the quantity of available data is surging as the structured or unstructured format day by day, and the powerful tool such as Cloud service allows users to store massive amounts of data. The raw data collected is mostly of poor quality, if lacks effective analysis this data is rarely transformed into useful knowledge for decision-making within organizations. Thus, the growing importance of data analytics in modern business has left many companies ill-equipped with data talent (Almgerbi et al., 2021). For instance, the usefulness of personal data has attracted both online and offline businesses, which desire data talent to conduct an in-depth analysis and exploit the hidden value behind personal information (Oh et al., 2019). A data scientist is regarded as the most promising occupation in the United States and other European regions in the past few years (Ramzan et al., 2021).

For a long period, job seekers through various platforms such as recruitment websites, social media or company recruitment portals to search for jobs. Advanced internet connectivity brings a lot of opportunities for both employers and job seekers (Sridevi & Suganthi, 2022). However, reviewing and identifying relevant profiles among numerous pieces of information is a big challenge not only for the human resource department but also for job applicants. This process costs labour and time with difficult to fill the job vacancy adequately and timely. The same situation exists in data-related fields, and it is even more thorny. Since data talent is a term that generally illustrates a set of interrelated skills from computer science, statistics and business domains, job requirements for data talent often consist of any relevant fields that may utilize data analytics in the company. For instance, in a study of the role of data scientist in the organization, this position is described as a profession that is capable to solve most analytical necessities individually (De Mauro et al., 2018). This simplistic and subjective recruiting vision reveals the ignorance of the diversity of specific skills required for collecting, organizing and transforming data. In this case,

job seekers rarely get information from job postings and hard to match themselves with job vacancies.

Based on previous studies, various methods have been applied to address this issue. The statistical content analysis has been applied to analyse a large number of job postings online, identify “job families” by the correlation among job postings and finally match appropriate skill sets (De Mauro et al., 2018). This thesis has contributed to job analysis, but a large workforce is in need during the research process, additionally, the results are based on experts' assessment, which indicates the cognitive level of the experts is an unstable factor in the research. ML analysis of text mining algorithm was also conducted on the survey of job requirements, to dig the information about knowledge fields and skill sets in the position description text (Mirjana et al., 2020). Supervised ML such as k-nearest neighbours (KNN) algorithm was used to classify recruitment data by categorization on job description text (Xiao et al., 2019), and unsupervised ML method such as topic modelling was conducted to discover the latent patterns of the textual contents in the online job posting (Gurcan & Cagiltay, 2019). Compared with supervised ML methods, the unsupervised algorithm is not mandatory for manual labelling or predefining the dataset, which reduces the bias from labels and human influence. Besides, since there is no strict restriction for unsupervised techniques to explore specific features in the dataset, it can find insights beyond the predetermined categories. In 2018, Large Language Models (LLMs) trained on large data sets came into prominence. Subsequently, different LLMs have been released and the most notable one is the Generative Pre-Trained Transformer (GPT) model from OpenAI. As an AI technique, the GPT model shows significant advantages due to its powerful language model that combines ML and Natural Language Processing (NLP) techniques, which can comprehend, interpret, and proficiently generate text that closely resembles human language. With the rise of GPT models, ML is enabled to learn from large amounts of textual data and recognize complex patterns in natural language without explicit programming of rules (Saka et al., 2024).

### **1.3 Problem Statement**

In a nutshell, with the rapid progress in technology companies have demanded a large number of data-related workers with skill-mix requirements that cover interrelated skills from computer science, statistics and business domains. However, due to the lack of clear skill requirements for data-related jobs, it is hard for job seekers to locate themselves among numerous, diverse and subjective job postings and match with suitable jobs.

### **1.4 Research Questions**

Hereby, this thesis has the innovation of obtaining recruitment data from JobStreet.com in Malaysia with a focus on data-related jobs, will perform text mining by AI technology and conduct job skills clustering by K-means model, to provide an overview of the required skills for data-related jobs in Malaysia.

There are few questions needed to answered throughout the thesis:

- i. What are the skills can be extracted from the job descriptions?
- ii. How does the unsupervised ML method process job skills?
- iii. What are the insights from grouped job skills?

### **1.5 Aims and Objectives**

The aim of this thesis is to extract insights from data-related job descriptions and identify the most required skills groups of data in Malaysian.

There are four objectives in this thesis:

- i. To establish a dataset of data-related job postings through scraping from website JobStreet.com using Python web scraper.
- ii. To extract features from job descriptions by implementing the GPT model.
- iii. To develop a K-means model for data-related skills clustering.

- iv. To analyse the clustering outcomes for extracting the most demanded data-related job skillsets and display the findings through a dashboard.

## **1.6 Scope**

There are some scopes in this thesis:

- i. Data is collected from the Malaysian online recruitment website JobStreet.com.
- ii. Job postings published from 18/02/2023 to 18/05/2023 are scraped by Python.
- iii. The keyword “data” is used for searching relevant job postings.
- iv. The unsupervised ML model of K-means is implemented for clustering job postings.
- v. Assessment matrices are silhouette coefficient and Davies Bouldin index.
- vi. Dashboard is constructed by PowerBI.

## **1.7 Significance of Research**

With the growing demand for data scientists and data engineers around the world, being unable to fill the job vacancy adequately and timely leads to the cost for organizations and a waste of data talent. Through the implementation of AI and ML clustering methods, this thesis is capable to decompose the complex set of interconnected skills for job seekers to easily understand data-related job requirements. The trends and manifesting technologies in the Malaysian data-related labour market can be identified in a smart way without a large amount of workforce assistance. By knowing the demands on employees, job seekers are able to locate themselves in the desired direction and have a cognition towards the required skill sets. There is great significance that effectively and scientifically promoting data talent to provide guidance on the development of the data industries.

## **CHAPTER 2**

### **LITERATURE REVIEW**

#### **2.1 Data Talent in Labor Market**

Today, data is constantly being generated with the development of big data in the industry and digital market, data as an asset is beneficial for all kinds of organizations to gain insights and promote their products or help managers with decision-making. For instance, people on Facebook share more than 100 terabytes per day, Google users conduct 1.2 trillion searches every day (Ramzan et al., 2021). Under the context of the ongoing evolution of technology, companies rely more on data in their day-to-day business to achieve goals and stay ahead of the competition. Data science is considered a fourth scientific paradigm alongside theoretical, empirical and computational sciences (Ramzan et al., 2021). Therefore, data analytics becomes increasingly valuable for companies and there is a growing demand for data talent.

According to Emerging Jobs Report from LinkedIn (Linkedin.com, 2017), the quantity of data scientists has increased by more than 650% from 2012 to 2017 in the United States. According to a survey conducted by IBM in 2017 (Markow et al., 2017), the job openings for data scientists would grow 15% during three years in the United States, and demand for both data scientists and data engineers would grow 39%. According to the U.S. Bureau of Labor Statistics (Chen, 2017), from 2016 to 2026, the technical and healthcare practitioners' occupations group is projected to create more than 1.3 million jobs, business and financial operations occupations group is projected to create more than 700 thousand jobs, computer and mathematical group is projected to create more than 600 thousand jobs. Companies and institutions related to these major occupational groups have significant demands on data talent to process and manage various types of data.

Due to the projects assigned by different institutions, the role of a data talent changes according to distinct working circumstances (Ramzan et al., 2021). In healthcare, data talents work on medical projects; physical data talents are often specialized in physics as well; in the computer science industry, data talents are those who have programming skills such as ML or AI; business data talents are required a good sense of business value and ability to perform business intelligence. Data analysts equipped with sector-specific knowledge are required by firms operating in a range of sectors. Furthermore, the recently emerged areas such as blockchain and cybersecurity technology also require data expertise, which is also often of a heterogeneous nature (Smaldone et al., 2022).

## **2.2 Difficulties in Recruitment**

In the aforementioned context, having active knowledge of data science is increasingly essential for job seekers. However, this is a field in perpetual development, there is rare consensus about what qualifies an applicant in the recruitment stage as a potential asset to organizations.

### **2.2.1 Difficulties for Employers**

The emerging job roles for data talent are making a quick entry onto the list of job vacancies in the hand of corporate recruiters. However, companies often rely on their subjective interpretation of corporation needs, then post the vague description of skills and requirements for data analytic jobs (De Mauro et al., 2018). The reason for this phenomenon is that business managers and human resource professionals lack the literature of a structured framework, which disables them to investigate the data-related requirements in distinct business fields. In other words, they lack a comprehensive understanding of constitutions in contemporary careers around data when designing job descriptions and assessing the adequacy of candidates, which affect the efficiency of filling job vacancies and recruiters may qualify unmatched candidates with the actual job needs.

### **2.2.2 Difficulties for Job Seekers**

The fact mentioned above that data scientist is the most promising occupation motivates not only graduates with data-related backgrounds, but also people with different educational backgrounds to follow this industrial trend. However, there is an absence of an overview of the practical skills or knowledge required in organizations for these people who desire to pursue data science professions (Ramzan et al., 2021). In the case that the scope of job descriptions covers skills generally from collecting, organizing, and transforming data into any relevant fields that may utilize data analytics in the company, job seekers are more difficult to gain useful information and match themselves with the job. Besides, it is a very time-consuming and tedious activity for job seekers to review a large number of job description and identify the most likely opportunities.

### **2.3 Job Postings**

In the early stages of information systems development, skill sets were divided into four traditional career paths: databases, networking, systems analysis, and programming (Lyu & Liu, 2021). Many researchers have pointed out that these skill sets are insufficient to keep up with growing market demands, and surveys of managers have tended to focus on general knowledge. Hence, surveys of job advertisements focusing on technical skills have become common. For medium of job advertisement, job information was initially posted in newspaper classified advertisements. In the last two decades, online job advertisements have evolved due to no strict space restrictions and allow more detailed job descriptions, which significantly contribute to job analysis.

To achieve the goal that employers engage with well-qualified talents and job seekers are able to find suitable employment, there is a lot of research on contextualizing the parameters of hard and soft skills from online job postings, mapping the constructs of employability with job categories and skill set taxonomy, and relating these skills to the role of data talent (Smaldone et al., 2022). These

analyses tend to fine-tune the scope of respective expectation from employers and employees, so the employability of data talent can be improved.

## **2.4 Content Analysis in Job Posting**

Kim et, al (2013) analyzed the 173 job postings from October 2011 to April 2012 on several online sources in North America, which focused on aspects of position title, degree, experience, institution, skills and duty. A qualitative analysis tool NVivo was used to code the job descriptions. Then, compare the frequency and proportion of terms that appeared in job descriptions. The research outcomes presented a great level of detail in the desired features in terms of skills, knowledge, duties and abilities.

To better understand the complex landscape of data professionals, Thielen and Neeser (2020) collected 180 job positions of data professionals from 2013 to 2018, involving experience, educational level and skills. Then, the 177 data professional job postings corresponding to the 180 job positions were analyzed independently. The research provided a descriptive analysis of the data set, including frequency, percentage, and median value for the notable terms in postings.

Lyu and Liu (2021) gathered a sample of 250 job posts on Monster.com from February 2005 to February 2006, use statistical method to determine the most required technical skills in the labor market. For most job postings, the text content under the heading “required skills” often reflects the information that needs to be analyzed. Some job postings divide the skills and technical knowledge, while both sets of listings were included in the sample. When ranking the skill listings, paired-sample tests were conducted for the difference in means. The P-value was used to identify the next lower-ranked skill.

In summary, traditional content analysis is mostly relying on statistics and human experts, which has limitations not only in the small size of the dataset but also the insufficient capability to discover latent information within the textual data of job postings.



## **2.5 Text Mining in Job Posting**

A few decades ago, the development of text mining in the analysis of job profiles was rare. For instance, there were only 4% of the studies conducted text mining among the 70 research in 2012 (Mirjana et al., 2020). However, text analysis has evolved considerably over time and the usage of text mining in job analysis is increasing. Compare with manual or statistical content analysis, text mining is attributable to the growing availability of technology resources, it is good at dealing with excessive information without the help of a large labor, which is more efficient and inexpensive. Therefore, text mining approach has been widely used for analyzing the knowledge stored on websites.

Text mining techniques has been used for various purpose, for instance, to analyze online textual reviews for investigating customer preference (Xu et al., 2017), to utilize the posts on social media for designing product (Jeong et al., 2019) and to analyze big data in financial sector (Bach et al., 2019). In the field of job posting analysis, text mining has been used for classifying online job advertisements or extracting useful information from job descriptions to gain insight. For instance, text mining can detect the arising skills in organizations and foster employment advancement accordingly (Mirjana et al., 2020).

Text mining can extract meaningful information from textual data by algorithms of average linkage, latent semantic analysis and so forth (Mirjana et al., 2020). It involves data pre-processing and algorithm implementation. During the pre-processing phase, there are various steps to clean textual data. Afterward, algorithms are conducted to extract words from the cleaned data, and ML techniques can be used for further phrases and topics research.

### **2.5.1 Pre-processing of Job Posting**

In the analysis of web-based textual data, the data pre-processing stage is very important to improve the quality of research outcomes and must be undertaken. It includes steps such as finding text sources from websites, converting textual data into

document warehouses, text tokenization, stop words removal, word stemming and lemmatization based on the existing knowledge bases or semantic networks and so on (Smaldone et al., 2022).

Fatih and Nergiz (2019) applied several sequential steps of data pre-processing in the study of knowledge domains and skill sets. In the first step, tokenization was conducted to divide the textual content into words and each word was represented by a vector. Secondly, meaningless content like private tags, website links, punctuation and stop words were removed. Word stemming has not been applied in this study due to there was technical jargon in the dataset, stemming may cause the loss of information in the content. Thirdly, with word vectors as the result of pre-processing, the job advertisements were eliminated from 13877 to 10432 unique words. At last, based on the “bag of words” approach these word vectors were combined to generate a document-term matrix for the quantitative analysis. For instance, this study generated 2638 rows and 10432 columns which stand for 2638 job advertisement documents and 10432 terms, while the weighting process in the matrix is according to the frequency of the word terms.

Almaleh et al. (2019) used the online job postings from Gresper and curriculum dataset from the courses by departments to analyze the gap between market skills and university curriculum. They applied data pre-processing of sampling and feature extraction. Referring to the study purpose, job description and job title from a total of 2550 job postings were selected as target features and loaded into R to perform feature extraction. Similarly, for the structured curriculum dataset, the course title and course description were loaded into R. Then, the text was tokenized into a single word, the capital letters were converted into lower case, stop words and words that appear rarely in the document were removed.

Mirjana et al. (2020) conducted lemmatization and exclusion on data pre-processing for job advertisement analysis and extracted the most frequent words and phrases subsequently. Firstly, lemmatization transformed the plural words into singular and verbs into the present tense. Exclusion expelled the words that occurred often but were less relevant for the research purpose, this process was based on the exclusion list of standard English words offered by Wordstat and selected additional

words by the researchers. Then, the words that occurred in more than 10 advertisements and phrases that consisted of no more than 5 words and occurred in more than 50 advertisements were extracted.

### 2.5.2 Text Vectorization

The spoken and written language is commonly stored in the form of textual data to conduct studies such as text similarity, sentiment analysis, classification and clustering. To be successfully feed into the computer's algorithms like ML, the vocabulary, grammatical rules and other language components in text need to be transformed into the numerical format since computers only able to process numbers of zeros and ones. In the NLP processing, transferring the textual data into numerical is vectorization (Uymaz & Metin, 2022).

In vector space modeling, one-hot encoding is one of the earliest methods which builds a binary vector, for each word in the vocabulary list of unique words based on its occurrence. The vector size is  $n$  where  $n$  is the length of the vocabulary list, the  $j$ th word has 1 in the  $j$ th position of the vector and the rest numbers of the vector are zero, like the example shown in Figure 2.1.

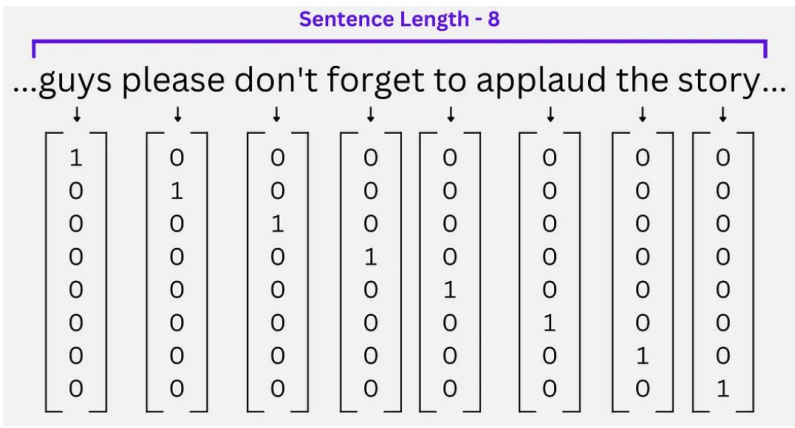


Figure 2.1 One-Hot Encoding Example

At present, term frequency (tf) and inverse document frequency (idf) are commonly used in vector space modelling. On the basis of one-hot encoding, tf-idf are able to assign the weights to the word based on its occurrence frequency. The value of

term frequency is obtained by the occurrence frequency in the document, and the value of inverse term frequency is obtained by dividing the term frequency in the document by the number of total documents containing the term. If a term appears several times in a few documents, the weight will be higher than the term appears in all documents.

Let  $t$  for terms,  $d$  for document,  $D$  for collection of documents,  $tf$  is defined as:

$$tf(t, D_i) = \frac{count(t)}{|D_i|} \quad (2.1)$$

The  $idf$  is defined as:

$$idf(t, D) = \log \frac{|D|}{1 + |\{d \in D : t \in d\}|} \quad (2.2)$$

The mathematical formula for  $tfidf$  is defined as:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \quad (2.3)$$

One-hot encoding and  $tf-idf$  are not considering the semantic relationships between the terms, and require large memory since the length of vectors is equal to the size of the vocabulary. In some analyses that worked with high-dimensional features like clustering, these methods may suffer from inefficient memory.

In recent studies, Word2Vec as the model that can capture semantic and contextual information is being popular. Word2Vec consists of word embeddings to build up the numerical vectors in a fixed length (Sabri et al., 2022). Word embeddings can detect the similarity between words, which represent each term as a unique vector and enable terms with similar contexts to have similar representations as shown in Figure 2.2. (Uymaz & Metin, 2022).

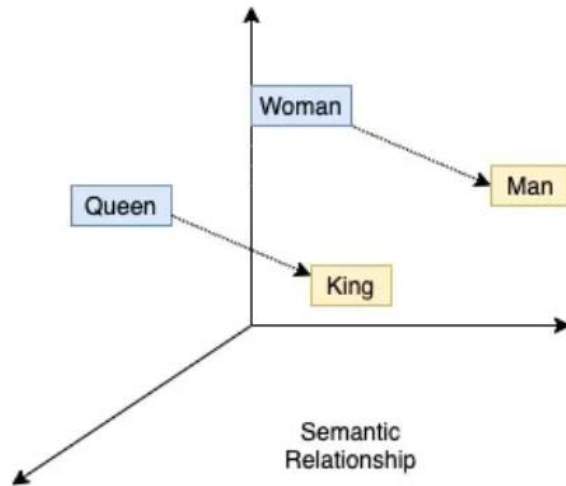


Figure 2.2 Word2Vec Example

Table 2.1 presents the pros and cons of different common vectorization methods. Although Word2Vec has the strength of considering semantic context and requires less memory, this thesis uses tf-idf as the vectorization method to process the skill phrases effectively.

Table 2.1 Vectorization Methods Comparison

	Pros	Cons
<b>One-Hot Encoding</b>	Simple	Not consider the semantic relationships, require large memory
<b>tf-idf</b>	Consider word occurrence frequency	Not consider the semantic relationships between the terms, require large memory
<b>Word2Vec</b>	Vectors in a fixed length, consider semantic context	Computational expensive, unable to process phrases

### **2.5.3 Machine Learning in Job Analysis**

The machine learning algorithms are suitable for training the model on nonlinear and dynamic data, due to the advantages of dealing with nonlinear and high dimensional relationships (Hayati et al., 2020). Textual data is nonlinear and not structured, so the most classic modern text mining techniques are composed of machine learning algorithms, either supervised or unsupervised.

There are some studies applying supervised ML methods. Almaleh et al. (2019) aimed to classify the job posting and curricula through the label of their common factors. They have conducted a supervised algorithm to train the textual dataset and predict the label as the output. R Studio and the keyword-search function have been used for labeling the dataset. Specifically, the Naïve Bayes classifier was used to apply the classification due to its efficiency. Xiao et al. (2020) proposed a heuristic KNN to classify the online recruitment information, which combines the KNN algorithm and heuristic search. This approach measures using fuzzy distance measurement, is able to adjust the weight of features and solve the unstable and inaccurate issues of the traditional KNN algorithm in text mining for categorizing talent.

Topic modelling and text clustering are the most used unsupervised machine learning technique in natural language processing. Topic modeling understands textual data by using NLP, which will tag the document by the closest resembles. Topic modeling discovers the topics in context by semantic clusters generated by the most used words in documents, at the meantime, the probability distribution is calculated as well by the Latent Dirichlet Allocation (LDA). LDA is commonly used in text mining for clustering the words in the document through a probability processing, since LDA is unsupervised, it can effectively work with the random text corpus. Gurcan et al. (2019) applied topic modeling with LDA to discover the pattern of the skills, knowledge and tools for software engineering jobs in the big data industry. Text clustering has high efficiency in large-scale data, while different from topic modeling that discovers the latent topics among the documents, clustering groups documents by the fit number based on a similarity metric. In other words, topic modeling based on the probability of the word's occurrence, clustering based on the similarity of the

words. Since the thesis aims to group data-related skills and find the most demanded skill set within the group, this thesis will use clustering as the ML technique.

There are various types of algorithms for clustering, including partitional clustering, hierarchical clustering and density-based clustering. Considering hierarchical clustering and density-based clustering don't match the aim of the thesis, only partitional clustering is accepted. Typically, there are two types of clustering methods which are soft clustering and hard clustering. Soft clustering assigns each data a certain probability belonging to each of the partitions. Hard clustering obtains the disjoint partitions of the data so each data is solely assigned to one partition, the most famous hard clustering algorithm is the Euclidean K-means algorithm due to its simplicity and efficiency. K-means clustering in semantic analysis can mine the information from multiple aspects. Debao et al. (2021) use K-means text clustering to process data of job names. They combined NLP and clustering algorithms to improve the similarities within the cluster and reduce the similarities between clusters. By the optimal number of clusters, high-similarity groups of big data positions were divided. Usabiaga et al. (2022) used a hierarchical method of clustering in Polish online job analysis, their method merged the two highest-similar attributes (columns or rows) into a new attribute, and subsequently form a larger cluster gradually. The approach allowed them to obtain a bicluster map of occupation groups, and the structure of each bicluster described the skills of the occupation with the task content. Mirjana et al. (2019) applied clustering in text analysis of job advertisements. They used an average-linkage hierarchical clustering algorithm to extract phrases and create a matrix of similarity, so the averages of the distances between each phrase among clusters can be calculated. Jothi et, al (2023) developed a machine learning model for grouping the learning difficulty into high and low levels using K-means clustering. Cluster evaluation measures such as Silhouette analysis, Elbow Method, and Davies Bouldin Index are used as shown in Figure 2.3. This research flow is used as the benchmark for the flow of the K-means implementation in the thesis.

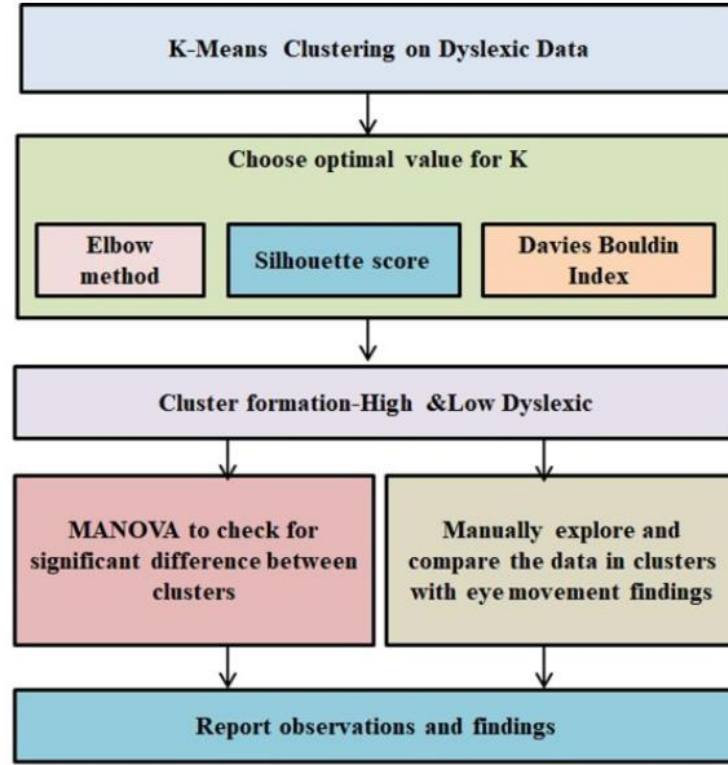


Figure 2.3 Jothi's Research Workflow

The clustering algorithms have used a range of distortion functions based on Bregman divergence to achieve the simplicity and scalability of the algorithm during the clustering. For instance, squared Euclidean distance and relative entropy have been widely applied in centroid-based parametric clustering approaches (Banerjee et al., 2005).

For the Bregman divergence in Equation (2.4), let  $\phi: S \rightarrow \mathbb{R}$  be a strictly convex function defined on a convex set  $S \subset \mathbb{R}^d$  such that  $\phi$  is differentiable on  $\text{ri}(S) \neq \emptyset$ . So  $d_\phi: S \times \text{ri}(S) \rightarrow [0, \infty)$  is defined as:

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla \phi(y) \rangle \quad (2.4)$$

For the squared Euclidean distance in Equation (2.5),  $\phi$  is strictly convex, let  $\phi(x) = |x|^2$  on  $\mathbb{R}^d$ , so  $d_\phi$  is defined as:



$$\begin{aligned}
d_\phi(x, y) &= \|x\|^2 - \|y\|^2 - \langle x - y, \nabla \phi(y) \rangle \\
&= \|x\|^2 - \|y\|^2 - \langle x - y, 2y \rangle \\
&= \|x\|^2 - \|y\|^2 - 2\langle x, y \rangle + 2\|y\|^2 \\
&= \|x - y\|^2
\end{aligned} \tag{2.5}$$

For the relative entropy in Equation (2.6), let random variables  $P, Q$  to be  $E_P(\log \frac{p(x)}{q(x)})$ , where  $p(x), q(x)$  are the probability density functions,  $p = (p_1, \dots, p_d)$  be a discrete probability distribution:  $\sum_{j=1}^d p_j$ , define the negative entropy by  $\phi(p) = \sum_{j=1}^d p_j \log p_j$ , So  $d_\phi$  is defined as:

$$\begin{aligned}
d_\phi(\mathbf{p}, \mathbf{q}) &= \sum_{j=1}^d p_j \log p_j - \sum_{j=1}^d q_j \log q_j - \langle \mathbf{p} - \mathbf{q}, \nabla \phi(\mathbf{q}) \rangle \\
&= \sum_{j=1}^d p_j \log p_j - \sum_{j=1}^d q_j \log q_j - \sum_{j=1}^d (p_j - q_j)(\log q_j + 1) \\
&= \sum_{j=1}^d p_j \log \left( \frac{p_j}{q_j} \right) - \sum_{j=1}^d (p_j - q_j) \\
&= \sum_{j=1}^d p_j \log \left( \frac{p_j}{q_j} \right)
\end{aligned} \tag{2.6}$$

## 2.6 Challenges and Opportunities

Text pre-processing is a challenging task in job analysis. During the traditional text pre-processing steps, experts select valid data based on their subjective judgment and the manual work of explicit programming. Besides, since many datasets involve different aspects of knowledge and technical terms, researchers often need self-defined dictionaries that contain relevant terms for use (Debao et.al, 2021). These methods consume a lot of time and lead to inevitable bias from humans. Therefore, to improve the efficiency of the work and reduce the bias, AI technology can contribute to pre-processing the textual data of job postings. In May 2020, a language model Generative Pre-Trained Transformer-3 (GPT-3) is released by OpenAI, which is a milestone for the Transformer-based language models within the NLP field. GPT-3 is an unsupervised ML model being able to produce human-like written language replies.

Through a large number of parameters, the AI model has the trend of expanding language size. The dataset owned by OpenAI for training the model has 300 billion tokens (Chan, 2022). Therefore, without the need for fine-tuning the text GPT-3 can complete language tasks such as text translation, text summarization or answering questions. Its level of performance has not been seen before in any NLP model, and the enormous scale of the model significantly the breadth, accuracy and quality of generated content. The GPT-3.5 model is an advanced version of the GPT-3, one of the largest language models and is more effective in answering universal questions and responding to various commands. GPT-3.5 is user-friendly and can perform question answering, text completion and language translation by delivering a deep understanding of human communication based on a transformer neural network trained on a super large dataset. Thus, this thesis will combine the GPT-3.5 model in text mining pre-processing to improve the usability of the textual data.

High dimensional space produced by NLP is another challenging task in job analysis. The high dimensional space damages the effectiveness of the text clustering process. To achieve an accurate and efficient clustering algorithm, the main concepts of the text need to be extracted by reducing the high dimensionality of the data, which transforms existing features into a low dimensional feature space. Mohamed (2020) compared three dimension-reduction algorithms for text clustering, which are Principal Component Analysis (PCA), Nonnegative Matrix Factorization (NMF) and Singular Value Decomposition (SVD). Mohamed conducted a series of experiments for Arabic and English, and showed that PCA provided the most interpretable results for both Arabic and English documents. Besides, the clustering quality was improved and the processing time was reduced with the PCA implementation. Thus, this thesis will use PCA in clustering to improve the effectiveness of the model.

## **CHAPTER 3**

### **RESEARCH METHODOLOGY**

#### **3.1 Research Operational Framework**

This chapter will illustrate the operational framework for job posting analysis by the K-Means algorithm with the data pre-processing using an AI model of GPT-3.5. In the chapter, the process including problem formulation, data collection, data pre-processing, modeling and outcomes discussion are addressed.

Figure 3.1 shows the flow of the operational framework and the aim of the thesis. Each phase will be discussed with the corresponding objectives in the chapter.

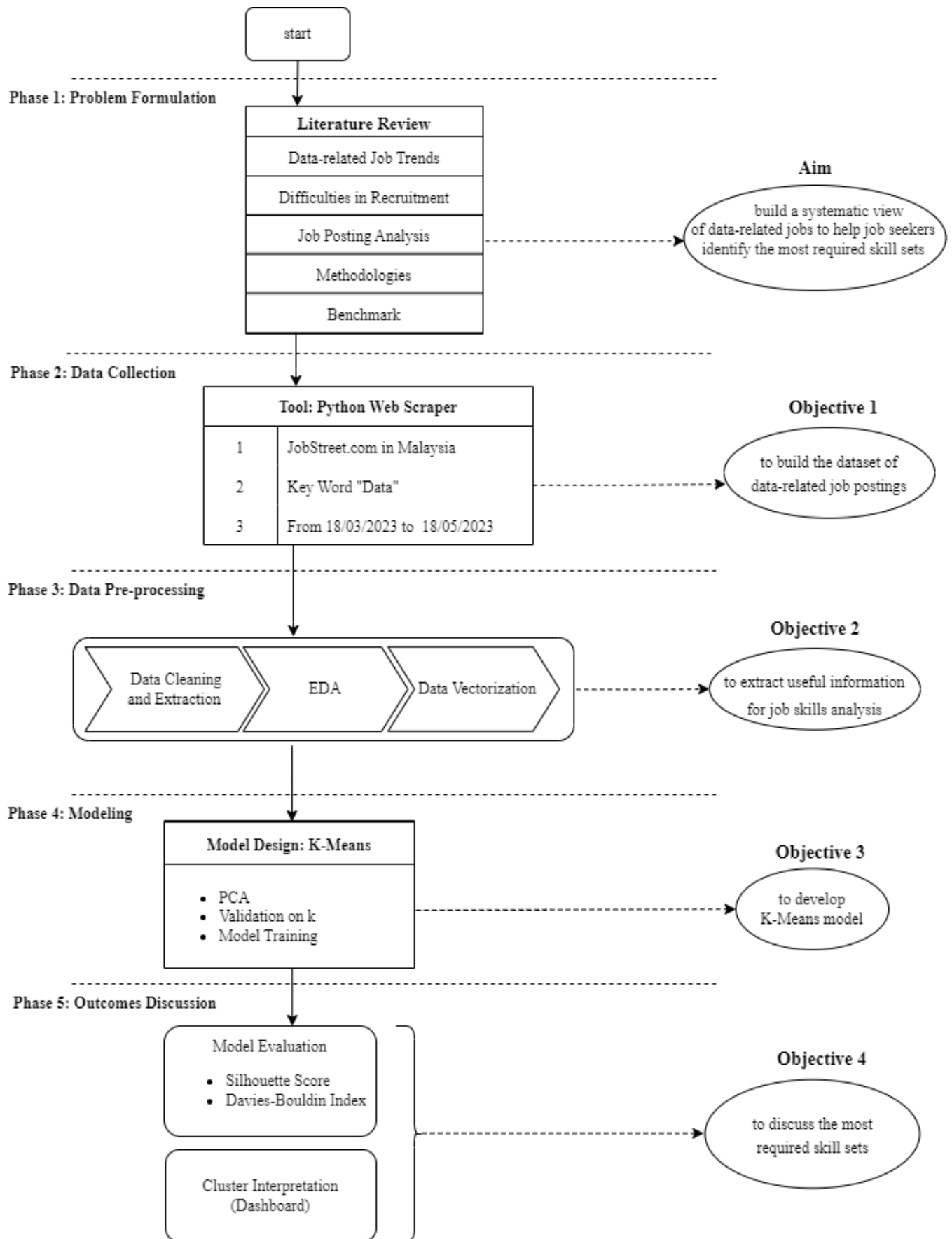


Figure 3.1 Operational Framework of Job Posting Analysis

### **3.1.1 Phase 1: Problem Formulation**

The modern labor market has a great demand for data talent to deal with the surging quantity of data and discover the latent business value underneath the pattern of data. Organizations and job seekers regard online job advertisements as an effective channel to connect employers and employees, however, due to the issues of subjective definition from corporations and unclear scope in the data-related job description, there is still a gap between job seekers' cognition and companies' requirements towards the data-related jobs. This thesis aims to help job seekers in Malaysia identify the most required skill sets and build a systematic view of job requirements in data-related industries.

The literature reviews cover the studies based on the demand for data talent in the current labour market, difficulties in recruitment, the changing trend in job posting channels and the corresponding analytical methodologies. Traditional content analysis as a statistical approach has a size limitation on the dataset and requires a large manual work, by contrast, text mining on job description analysis combines the ML techniques that can analyse textual data of job descriptions with effectiveness and simplicity. By applying the AI techniques at the data pre-processing stage, text mining not only can discover more latent information within the textual data than traditional content analysis without the large manual work, but also can get less biased outcomes in a short time. Besides, the unsupervised ML technique is suitable for the unlabelled dataset. More specifically the K-means algorithm as a partitioning technique corresponds to the thesis aim of grouping  $n$  observations into  $k$  clusters.

### **3.1.2 Phase 2: Data Collection**

The dataset of data-related job postings is scraped from Malaysian JobStreet.com by searching the keyword “data”, which has a period of three months from 18/02/2023 to 18/05/2023, with 15504 online postings. Python as the scraping tool is used. Privacy for collected is preserved, and processing will base on an appropriate legal public interest. Figure 3.2 illustrates the flow of the data collection.

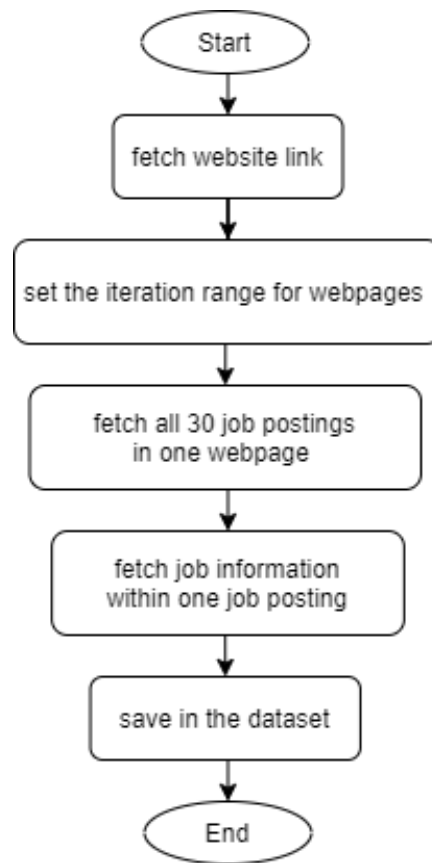


Figure 3.2 Data Collection

### 3.1.3 Phase 3: Data Pre-processing

The collected raw data will undergo the pre-processing flow from data cleaning, data extraction, till data vectorization as stated in Figure 3.3.

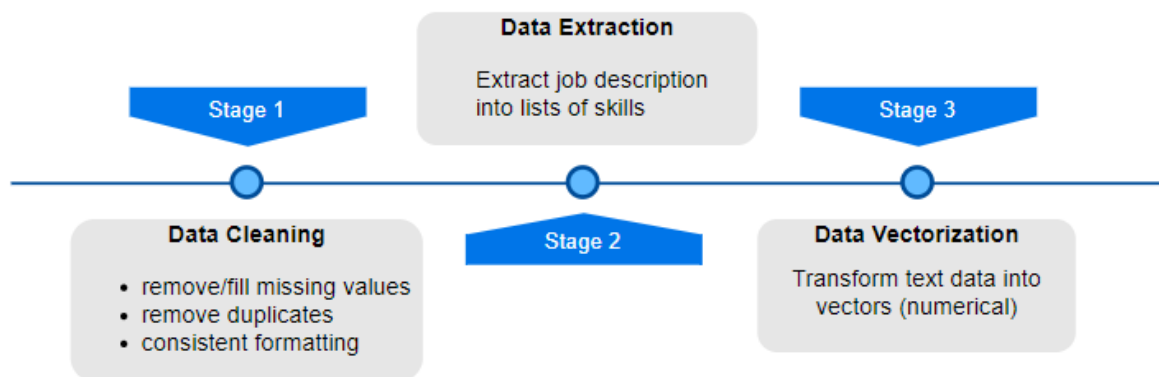


Figure 3.3 Data Pre-processing Stages

At the data cleaning stage, the missing values in the dataset are identified and processed accordingly, either be fixed or removed. Also, data that in inconsistent format will be processed to keep consistency of the dataset. Besides, duplicated values will be removed to avoid data redundancy. At the data extraction stage, long sentences in the job description are extracted into a list of short words that only containing information about skills. At the data vectorization stage, text is converted into numerical data so that ML can use it as the input to undergo the model analysis.

#### 3.1.4 Phase 4: Modeling

This phase will explain how the K-means model is used to cluster the skills that extracted from the job descriptions. K-means implements partitional clustering which divides data into groups without overlapping as shown in Figure 3.4. In this way, skills can be divided into groups based on their similarity, and each group will have their own characteristics. Through analysis on similarity within the group and distinctions between groups, insight can be derived for the data-related job requirements in Malaysia.

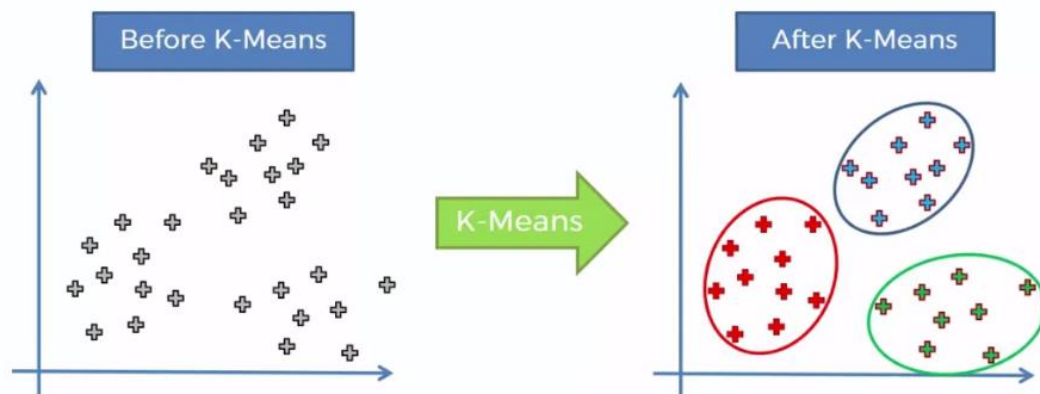


Figure 3.4 K-means Clustering

### 3.1.5 Phase 5: Outcomes Discussion

Since clustering based on the similarity of the data point, most of the clustering metrics are designed to maximize intra-cluster similarity and reduce inter-cluster similarity (Wijaya et al., 2021). This thesis uses the silhouette coefficient and the DBI to assess clustering performance by measuring the similarity of a data point with other data points inside the cluster and outside the cluster. Besides, a dashboard will be constructed to present the result visualization, and insight discovered from the model outcomes will be interpreted.

## 3.2 Performance Measure

The silhouette coefficient calculates how well the point fits in its cluster compared to other clusters. Equation (3.1) shows the function of the silhouette coefficient for a data point where  $S(i)$  is the silhouette coefficient for data point  $i$ ,  $a(i)$  is the average distance between  $i$  and all other data points in the cluster to which  $i$  belongs,  $b(i)$  is the average distance from  $i$  to all clusters that  $i$  does not belong to. By the equation, the average silhouette value will be calculated for every  $k$ . The silhouette coefficient is between -1 and 1, where a higher silhouette coefficient indicates that clusters are clearly distinguished from each other and a lower silhouette coefficient indicates clusters are assigned in the wrong way.

$$S(i) = \frac{b(i) - a(i)}{\max \{a(i), b(i)\}} \quad (3.1)$$

The Davies Bouldin index (DBI) in Equation (3.2) assesses clustering performance by obtaining the intra-cluster similarity and inter-cluster similarity (Wijaya et al., 2021), where  $S$  is the average distance between the feature vectors in cluster and the centroid of the cluster,  $M_{i,j}$  is a measure of separation between cluster  $i$  and cluster  $j$ . It calculates the average value of the maximum ratio of the intra-cluster distance and the inter-cluster distance for each cluster. A lower DBI value indicates a better number of clusters.



$$DB = \frac{1}{N} \sum_{i=1}^N \max \frac{s_i + s_j}{M_{i,j}} \quad (3.2)$$

### 3.3 Summary

This chapter states the overall flow of the operational framework, including 5 phases and corresponding objectives. The processing steps and all related works in each phase are explained thoroughly.

## **CHAPTER 4**

### **RESEARCH DESIGN AND IMPLEMENTATION**

#### **4.1 Data Preparation**

This section will explain the flow of data collection and data preprocessing, including methods and tools used for cleaning, extraction and vectorizing data. The overall flow of data preparation is shown in Figure 4.1. After the suitable data preparation, the dataset can be fed into the model to further explore meaningful insight.

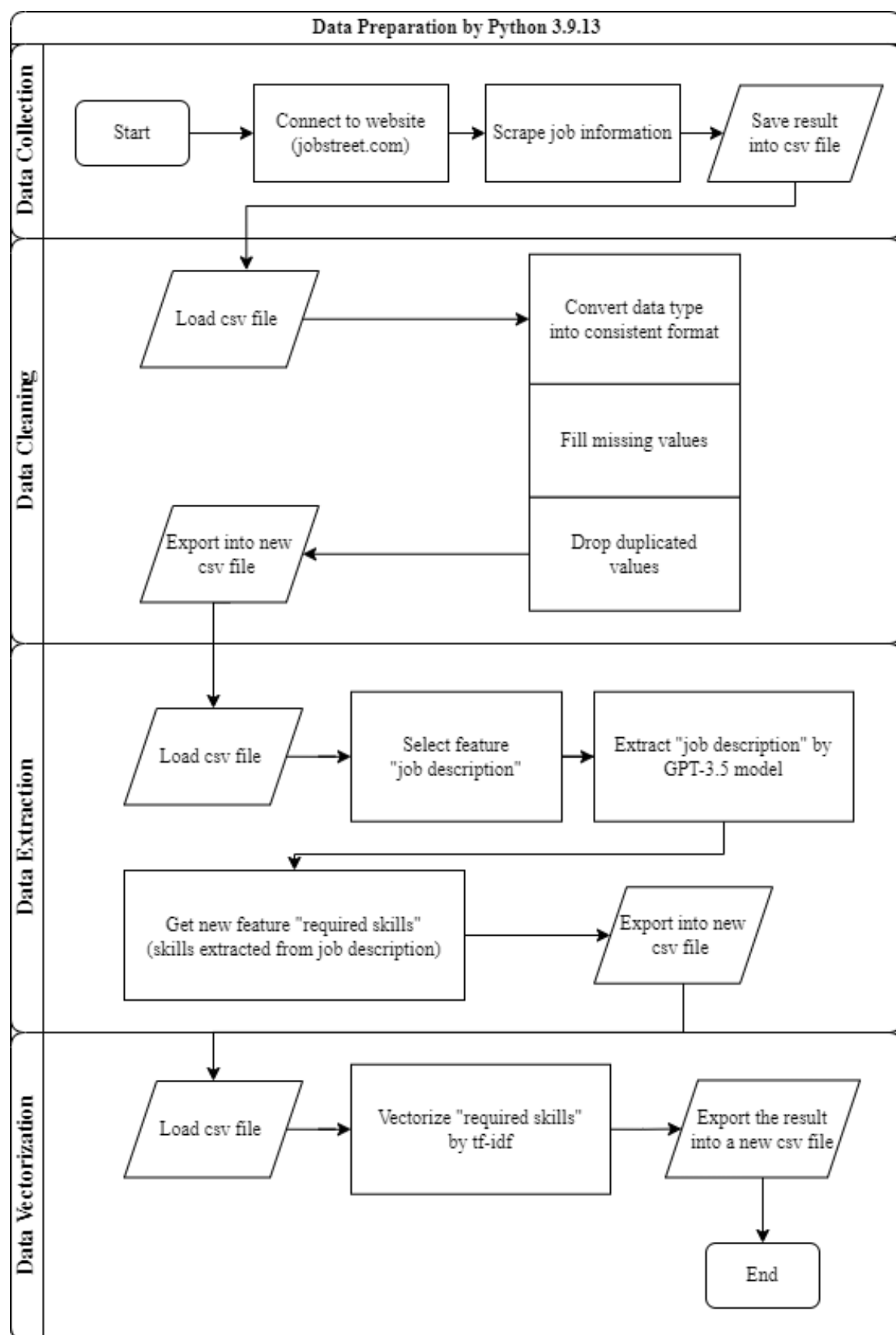


Figure 4.1 Data Preparation Flow

### 4.1.1 Data Collection

As aforementioned, the dataset of data-related job postings is scraped from Malaysian JobStreet.com. This website is one of the most common Malaysian online recruitment websites, which contains over 30 million profiles as mentioned in Table 4.1. There are many other popular online recruitment websites in Malaysia, however, through a bunch of experimentations it is found that most websites have restrictions on conducting web scraping on their web information, whether denied the access request or interrupted connection after a period of attempting. Only the JobStreet website accepts long-term web scraping. Besides, the website contains over 10k job postings under the field of data, which is acceptable in quantity for performing data analysis.

Table 4.1 Description of Data Source

Data Source	JobStreet.com
Characteristics	<ul style="list-style-type: none"><li>• Southeast Asia's largest online employment company.</li><li>• Contains over 30 million profiles.</li><li>• Over 49500 job postings in Malaysia.</li><li>• Only display postings within the last 30 days</li></ul>
Strength	<ul style="list-style-type: none"><li>• Allows to web scraping.</li><li>• Over 10k job postings related to data industry.</li></ul>

Figure 4.2 shows the web page of JobStreet. It has 3 main filters on the top which are job keyword, job location and job specialization. As the objective of the thesis is to extract insights from data-related job postings in Malaysia, the first step is to set the job keyword as "data" and the location as "Malaysia". Secondly, due to the website policy, the website only keeps postings published within the last 30 days, by setting the different time period, the desired job information will be found as the result. In this thesis, the earliest job postings can be found are published on 18/02/2023, and at the end of the data collection stage, the latest postings are published on 18/05/2023.

JobStreet by SEEK Job search MyJobStreet Company profiles Career advice New Login For employers

1.key words

data Malaysia

Salary Job type Last 30 days Clear Filters Sort By Relevance

1-30 of 13,246 jobs Apply Now View in new tab 3.result Close

**Job 1:**

- Company:** PEERHEALTH MALAYSIA SDN BHD (PISIL GROUP)
- Position:** Data Visualization Executive - (Market Research)
- Location:** Petaling Jaya
- Work Type:** Hybrid Work
- Posted:** 1d ago

**Job 2:**

- Company:** Rahi
- Position:** Asset Management Specialist (DATA CENTER OPERATIONS, Johor MY - KULAL , NUSAJAYA)
- Location:** Johor

**Job 3:**

- Company:** PEERHEALTH MALAYSIA SDN BHD (PISIL GROUP)
- Position:** Data Visualization Executive - (Market Research)
- Location:** Petaling Jaya
- Posted:** 18-May-23

Figure 4.2 JobStreet.com Website

<https://www.jobstreet.com.my/data-jobs/in-Malaysia?createdAt=7d>

Figure 4.3 shows the search result from the website. There are two parts to the result webpage. The first part on the left side is a vertical list that contains the qualified jobs. The second part on the right side is the detail of the selected job, which thoroughly states the job information about the title, company, location, job type, and job description, etc.

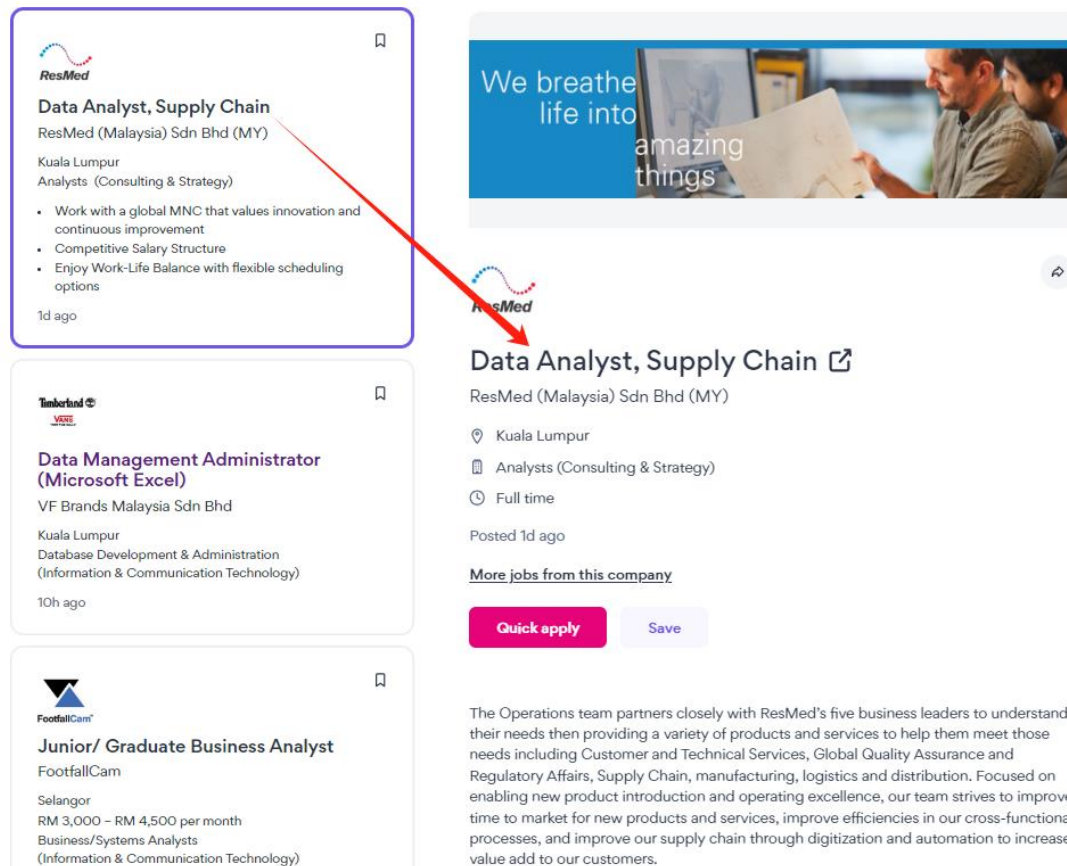


Figure 4.3 Search Results

From the search results, there is a lot of information contained in each job posting. However, it is unnecessary to fetch all job information. In this thesis, the key consideration on how to select information as features comes from 3 aspects: technologically achievable, structured text and relevant to the analysis. The first aspect is to consider if it is technically achievable. Python is used as the scraping tool by the package “Beautiful Soup”, which is specifically for parsing and extracting data from HyperText Markup Language (HTML), while the web page is in the format HTML. As “Beautiful Soup” works by searching and iterating web pages in Python loops, the HTML format of the target data in each web page must be consistent to ensure that

automatically fetched data is the target data. Secondly, if the information on the web page is semi-structured or unstructured, it will be hard to conduct analysis and interpret model outcomes. Thus, only information in a structured format is accepted. Thirdly, if the information is less relevant to the analysis, then no need to scrape those data. For instance, Mirjana et al. (2020) used Location, Employment type, Seniority level and Job description as features to implement job analysis. Figure 4.4 shows the flow to decide whether the job information can be the feature.

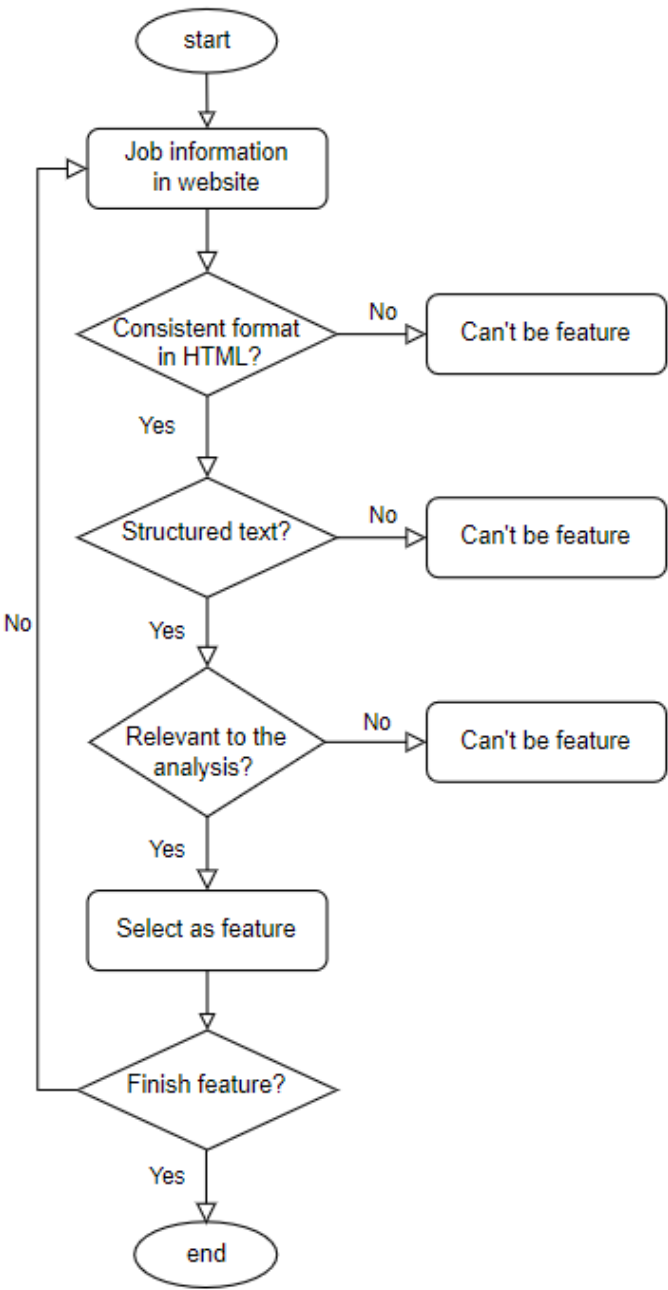


Figure 4.4 Selection of Job Features

Through the above flow, there are 6 types of job information selected as the feature, which are “job\_title”, “location”, “career\_level”, “experience”, “job\_type” and “job\_description”, the detail of each feature is illustrated in Table 4.2.

Table 4.2 Feature Description

Feature	Description
job_title	Job title listed in the job posting.
location	First job location listed in the posting.
career_level	Career level mentioned in the posting.
experience	Required experience mentioned in the posting.
job_type	Job type mentioned in the posting, weather it is full-time or part-time.
job_description	Job description includes job requirements and thorough explanation.

The job information is scraped step-by-step. The first step is to set the iteration range of webpages. Secondly, since there are 30 postings on each page, conduct the recursive processing 30 times to fetch all job postings on the page. Lastly, for each posting fetching the features selected at the former step. The detailed Python code is shown in Appendix B. As a result, 15504 data-related job postings are scraped and saved into the comma-separated values (CSV) file.

Figure 4.5 shows the first 5 samples of the raw dataset which has 6 columns. As shown below the scraped data is easy to read and understand. However, the data format is not fully consistent, some in upper case and others in lower case. Moreover, some column contains information which is not matching with the column title.



	job_title	location	career_level	experience	job_type	job_description
0	Data Analyst	Johor Bahru	Non-Executive	Full-Time	Computer/Information Technology, IT-Software	ZEMPOT MALAYSIA SDN BHD is a newly branched co...
1	Director- Data Analytics	Petaling Jaya	Senior Manager	12 years	Full-Time	Job overviewWe are looking for a highly motiva...
2	Solutions Architect - Data Lake Specialist (Ba...	Kuala Lumpur	Senior Executive	8 years	Full-Time	We are looking for a Solutions Architect who s...
3	Automotive Analyst – Vehicle Valuation	Kuala Lumpur	Senior Executive	3 years	Full-Time	DescriptionThe Automotive Analyst will be acco...
4	IT Business Analyst	Kuala Lumpur	Senior Executive	4 years	Full-Time	Position Objective:- Responsible to be the IT B...

Figure 4.5 First 5 Samples of Raw Dataset

Through the elementary analysis conducted by Python, Figure 4.6 shows the general pattern of the raw dataset. There are a lot of missing values and duplicated values.

```

The number of rows: 15504

The number of columns: 6

Missing value:
  job_title      1941
  location        0
  career_level    0
  experience      0
  job_type      1231
  job_description  0
  dtype: int64

Duplicated value: 2431

```

Figure 4.6 Elementary Analysis Result

### 4.1.2 Data Cleaning

To improve the availability of the collected data, the raw data undertakes cleaning as the first preprocessing step. Figure 4.7 shows the flow of the data cleaning. At first, all strings in the dataset are converted into lowercase. Then, if “job\_title” is missing then there is no indicator to explain the job posting. Therefore, drop the data if “job\_title” is empty. For missing values in other columns, filled missing values by “not specified”, to keep as much as information and be meaningful for result interpretation. Afterward, to keep the consistent of the data format, if there is any unmatching information among columns, adjust the column to ensure it only contains the most relevant information. Lastly, if there are any duplicates then drop the duplicates.

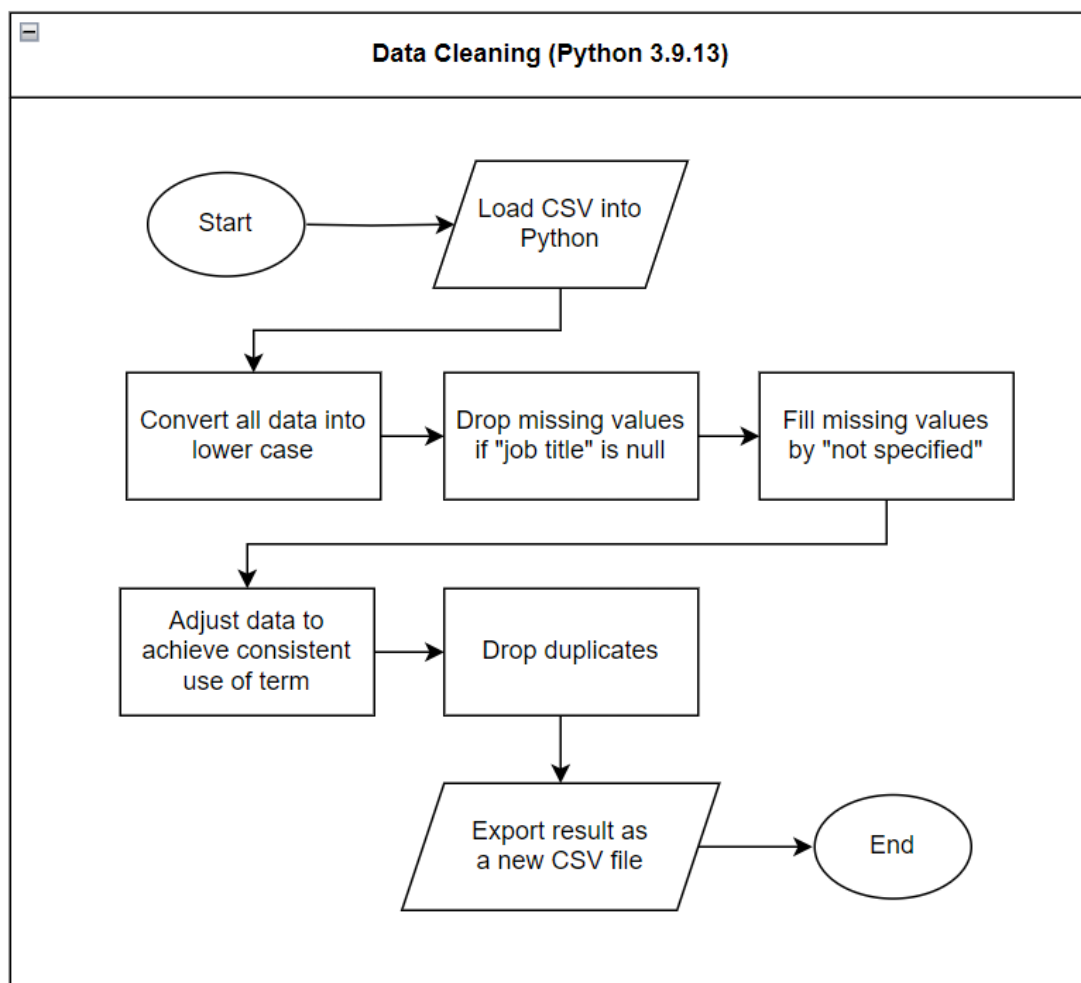


Figure 4.7 Data Cleaning Flow

As the result shown in Figure 4.8, the dataset after cleaning left 11184 rows without missing values, and all duplicated rows are removed. This result is saved into a new CSV file. Details of the dataset after each processing are shown in Appendix B.

```
shape: (11184, 6)

Duplicated value: 0

Missing value:
  job_title      0
  location       0
  career_level   0
  experience     0
  job_type       0
  job_description 0
dtype: int64
```

Figure 4.8 Data Cleaning Result

#### 4.1.3 Data Extraction

There is one unique feature in the dataset, which is “job\_description”. As aforementioned, “job\_description” includes job requirements and a thorough explanation of the job. However, there are a lot of limitations if using traditional text pre-processing to extract data from “job\_description” into desired information. Firstly, must self-define the relevant terms based on subjective judgment and construct the dictionaries of those terms for use. Secondly, need a lot of manual programming work. Therefore, to improve the usability of the textual data in “job\_description”, and to increase the efficiency of data extraction, the AI model GPT-3.5 is used in this step.

To implement the GPT-3.5 model, permission from the company OpenAI is required. OpenAI provides an application programming interface (API) to allow users to access their products. Then, the prompt question needs to be defined to formulate the way to extract the input text. By iterating the question on each sentence of the “job\_description”, it can be extracted into lists of skills mentioned in the text. Figure 4.9 illustrates the flow of the data extraction.

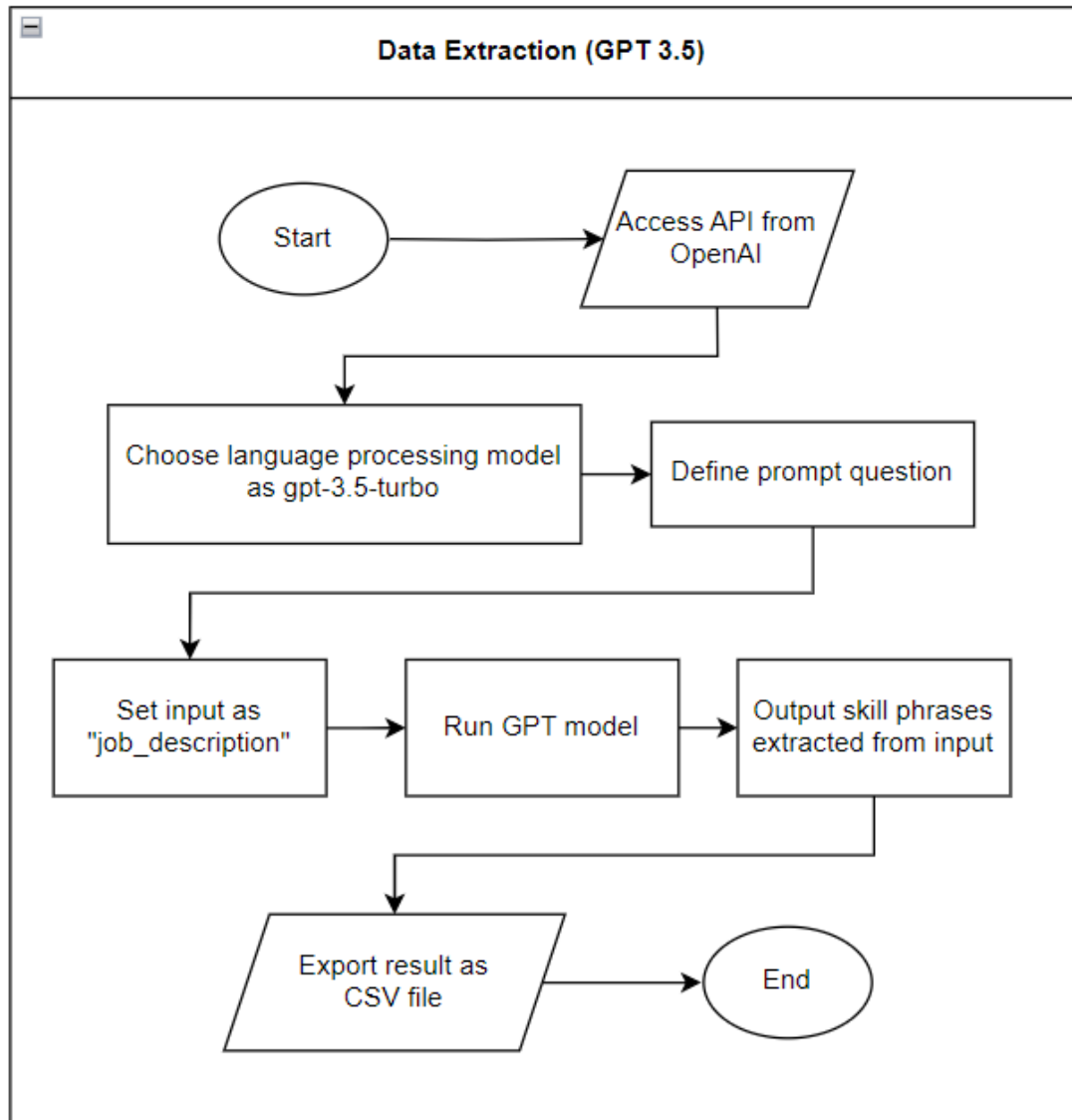


Figure 4.9 Data Extraction Flow

The prompt question is defined as “Output a JSON object that list out the hard skills, tools, programming skills”. Firstly, the reason for output in JSON is that GPT-3.5 can’t guarantee consistency in output format due to its unstable cognition, if using a word like "dictionary", it may be ambiguous on the result. While JSON as the technical term is more robust and understandable. Secondly, setting the keywords as hard skills, tools and programming skills is trying to give a specific scope towards the word “skill” as much as possible. Thus, it is more possible that the GPT-3.5 model will understand the prompt question and fetch the desired information. By a bunch of experiments, the GPT-3.5 model works best on this command, it is able to fetch all

kinds of mentioned skills and required knowledge in the "job\_description". The detail of the code is put in Appendix B.

GPT-3.5 is not totally free but it provides a free trial to users. Figure 4.10 shows the usage consumed for the data extraction. There is a free trial of 18 dollars, and the 11184 records of job descriptions consumed around 14 dollars. Meanwhile, OpenAI has enforced rate limits on the requests of the gpt-3.5-turbo model, which only allows 3 times access per minute. It leads to an extremely long time to process the whole dataset, which takes around one week to complete the extraction.

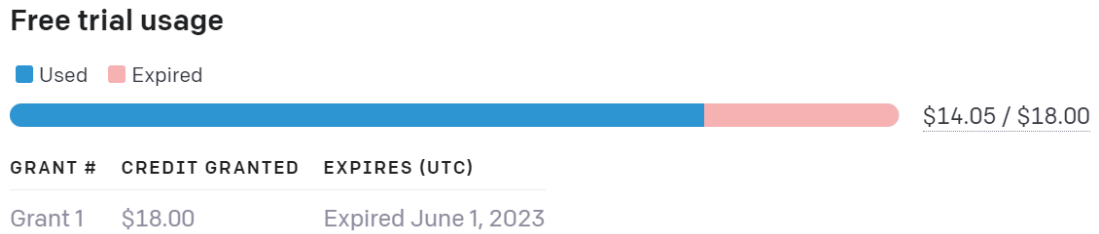


Figure 4.10 API Usage

Table 4.3 shows an example of data extraction result. For each job description, the extracted information is stored as a string but the appearance of list-like. If there is no related information detected in the job description, nothing will be returned. As aforementioned, the useful information is identified and summarized automatically by the GPT-3.5 model, which is much more effective and subjective than traditional text pre-processing.

Table 4.3 Data Extraction Example

Raw data in Job description	Extracted Skills
<p>description1. support the head of department in <u>all phases of the project</u> from feasibility study up start-up and performance test.2. development and preparation of project definition manual and related documentation.3. review of existing <u>system process, data analysis</u> and <u>troubleshooting</u> to enhanced new <u>software development</u> or <u>modification existing software</u>. 4. developing <u>pre-commissioning</u> and <u>commissioning</u> and start-up procedures, <u>performance test</u> procedures and <u>operational manual</u> company transwater brings together latest technology and engineering from global manufacturers to provide innovative solutions and services to our customers with a local touch and professionalism. transwater is staffed with a group of dedicated employees who are experienced and suitably qualified in the field of work they work in. we provide business solutions to our customers through the correct application of products, services and solutions that contribute to our customerâ€™s enterprise in the areas of control, safety, efficiency, and reliability.</p>	<p>[ project management, system analysis, troubleshooting, pre-commissioning, commissioning, performance testing, operational manual development, data analysis tools, software development tools]</p>

#### 4.1.4 Exploratory Data Analysis

To better understand the distribution of each feature, RapidMiner and PowerBI are used to demonstrate the overview of data. For the feature “job\_title”, the distribution is displayed in Figure 4.11. As the graph shows, data-related jobs in Malaysia have a wide range so there is a high-diversity of titles.

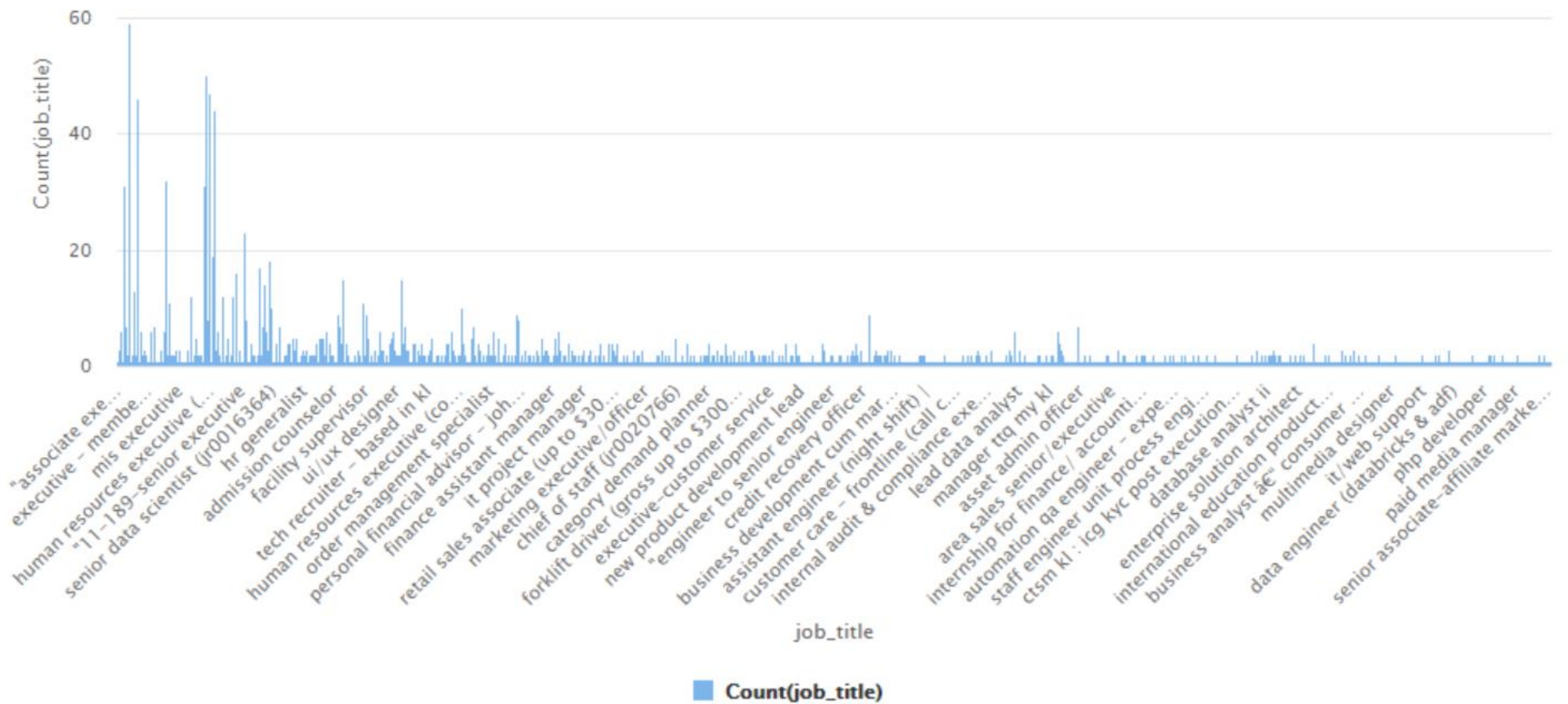


Figure 4.11 Job Titles Distribution

For the feature “location”, Figure 4.12 shows the frequent locations, the top three locations are Kuala Lumpur, Selangor and Petaling Jaya. Kuala Lumpur as the capital city of Malaysia has the most job opportunities, which occupies 26%.

Index	Nominal value	Absolute count	Fraction
1	kuala lumpur	1411	0.260
2	selangor	409	0.075
3	petaling jaya	374	0.069
4	shah alam	225	0.041
5	penang	214	0.039
6	johor bahru	171	0.031
7	bayan lepas	163	0.030
8	multiple work locations	120	0.022
9	subang jaya	106	0.020
10	puchong	97	0.018

Figure 4.12 Top Frequent Job Locations

For the feature “career\_level”, Figure 4.13 shows that 32.6% is not specified, junior executives and senior executives are the most in need with 23.9% and 17.9%. The need for senior managers is the lowest, only 1.9%.



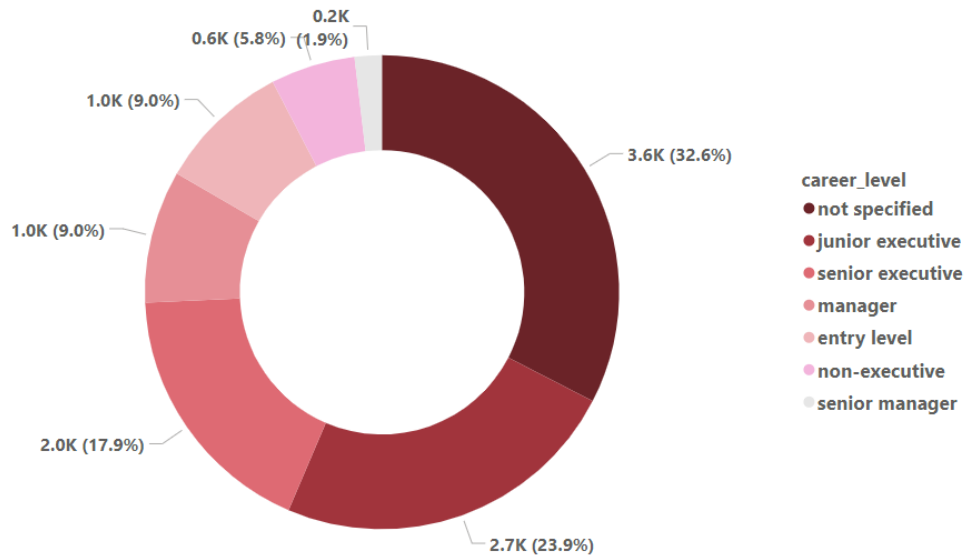


Figure 4.13 Career Level Distribution

For the feature “experience” Figure 4.14 shows that most companies require employees to have experience from one year to five years.

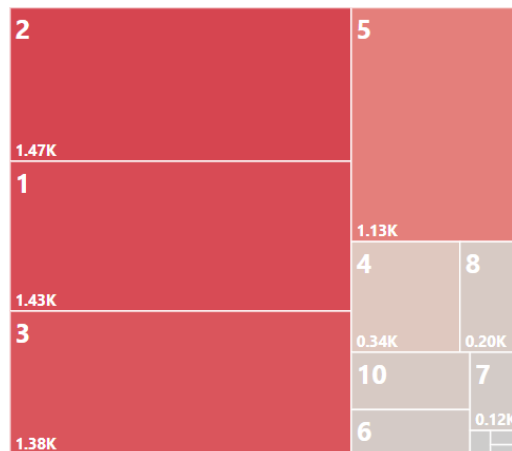


Figure 4.14 Required Experience Distribution

For the feature “job\_type”, Figure 4.15 shows that more than 80% of the jobs is full-time. While job for temporary and part-time is very limited, which is less than 3%.

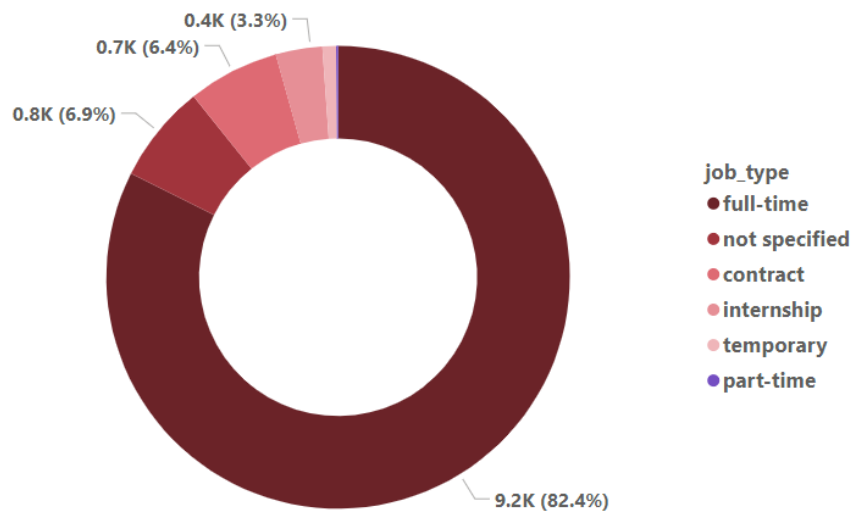


Figure 4.15 Job Type Distribution

For the feature “required\_skills” extracted from “job\_description”, Figure 4.16 shows the overview of the first 10 rows. Since the objective of the thesis is to cluster skills, this feature will be used as input, fed into the K-Means model. Then, the clustering results will be combined with other features, to discover the insight from the Malaysian data-related job postings.

	required_skills
0	['data analytics', 'knowledge in online table ...
1	['data science', 'business intelligence', 'mac...
2	['data lake architecture', 'data modelling', '...
3	['data analysis', 'market research', 'database...
4	
5	['programming languages like mysql', 'database...
6	['machine learning', 'deep learning', 'data qu...
7	['BI ETL', 'DWH projects', 'Informatica', 'Azu...
8	
9	

Figure 4.16 Required Skills Example

### 4.1.5 Data Vectorization

Before the vectorization, only the skill phrase that consists of a maximum of 5 words is selected to conduct the following processing ((Mirjana et al., 2020). This approach resulted in 24234 phrases of the required skills and knowledge that occurred in the job description as shown in Figure 4.17. The most required skills include data analysis, data entry, project management and so forth. The code of vectorization is put into Appendix B.

```
number of total skills: 24234

Element Count:
data analysis          483
data entry             266
project management     184
financial analysis     175
inventory management   152
...
drill                  1
inventory forecasting  1
parts management       1
meticulousnesssupply chain management software  1
aml/kyc certificationsms office products        1
```

Figure 4.17 Skill Phrases Overview

In this section, the tf-idf approach is applied to convert extracted skills into numerical data. The reason for selecting the tf-idf method is that it considers the frequency of the documents which is more intelligent than one-hot encoding. More importantly, extracted skills involve single words and phrases, such as project management, data science, etc., while only tf-idf can process phrases. The tf-idf method measures a phrase's occurrence in a document and the total number of documents the phrase is in. Even though Word2Vec has the strength that considering the semantic context of the text, it will split text into a single word which may lead to incomplete information. Therefore, this thesis uses the tf-idf as the vectorization method. As shown in Figure 4.18, 24234 samples are converted into vectors and 6283 features are generated.

```
n_samples: 24234, n_features: 6283

Document index, Specific word-vector index, TFIDF score:
  (0, 344)      0.8627008873865464
  (0, 1425)     0.5057145231278862

[0. 0. 0. ... 0. 0. 0.]
```

Figure 4.18      Vector Example

## 4.2      K-means Model Development

In this section, the scikit-learn machine learning library for Python is used to conduct the K-means model. As shown in Figure 4.19, each step and method involved in model training will be explained explicitly, including PCA to simplify the high-dimensional data, elbow method to identify the suitable number of clusters, etc.

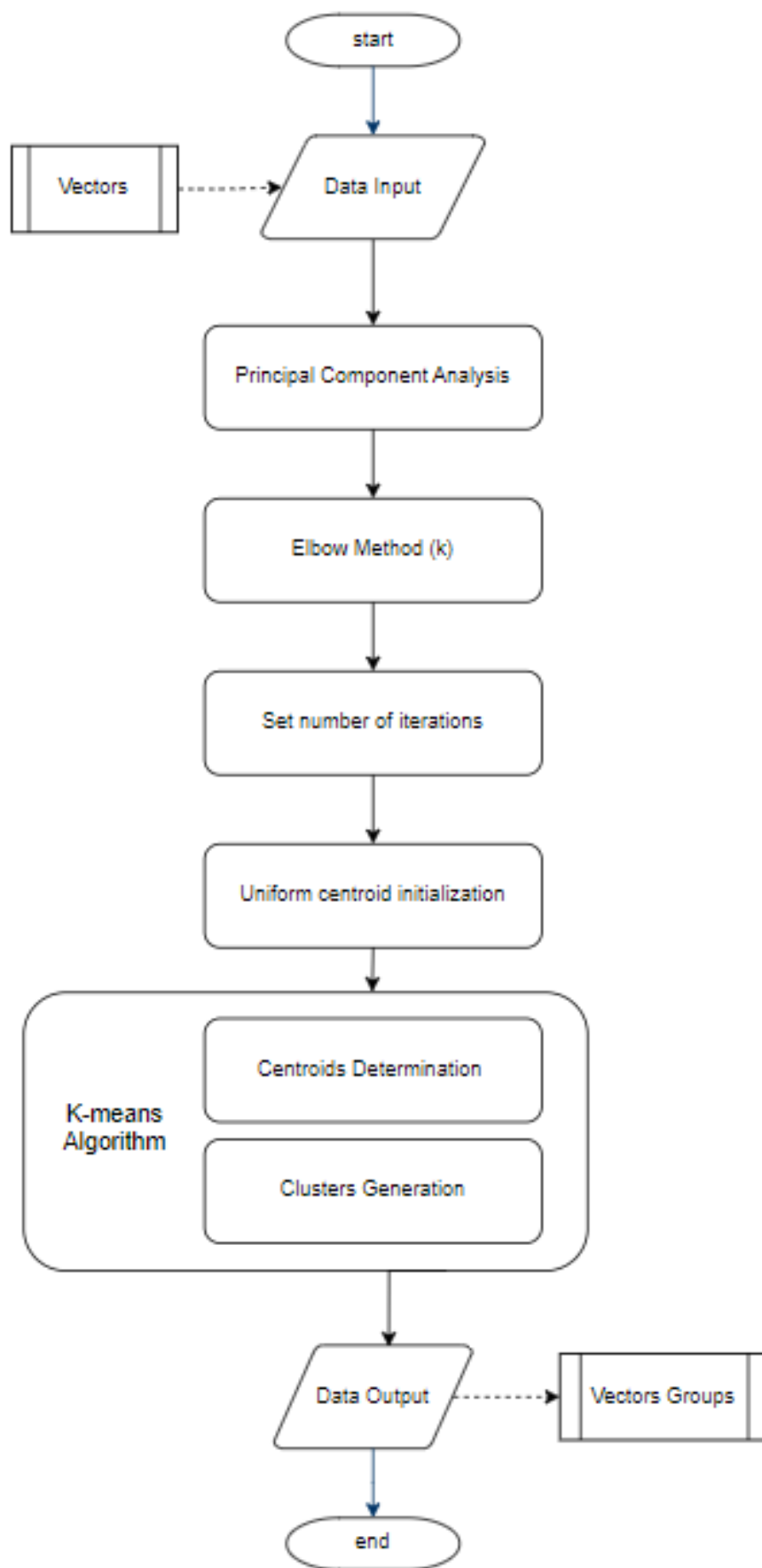


Figure 4.19 K-means Model Development

### 4.2.1 Principal Component Analysis

PCA is a popular dimensionality reduction technique and has been widely used in many fields (Francesco et al., 2023). Since PCA uses the most effective principal components of the data, most of the patterns and trends of data can be retained during the reduction of the dimensionality. In this section, due to the high dimension of 6384 generated by the tf-idf vectorizer, it is hard for the K-means model to effectively differentiate groups. Specifically, K-means clustering determines centroids and clusters based on intra-cluster distance and inter-cluster distance, without the dimension reduction, the distance among data points is prone to be very small and the model can't distinguish similarity among data points effectively.

By setting the dimension range from 50 to 2, the K-means model performance is measured by silhouette coefficient and DBI. Silhouette coefficient can evaluate how similar a data point is within-cluster compared to other clusters, a score of 1 denotes the best which means the data point is compact within the cluster it belongs to and far away from other clusters. Oppositely, the score -1 denotes the worst case. Similarly, DBI calculates the ratio of the within-cluster distance and the between-cluster distance, the lower the value denotes the better separation of the clusters and compact inside the clusters. The result is shown in Figure 4.20. It can be seen that with the decrease of the dimension, the performance is better. When the dimension is 2, the silhouette coefficient is the closest to 1 and DBI is the lowest.

However, to retain the patterns and trends of the data as much as possible, select 4 as the final reduced dimension, where the silhouette coefficient is relatively high (above 0.7) and DBI remains low. The Python code is put in Appendix B.

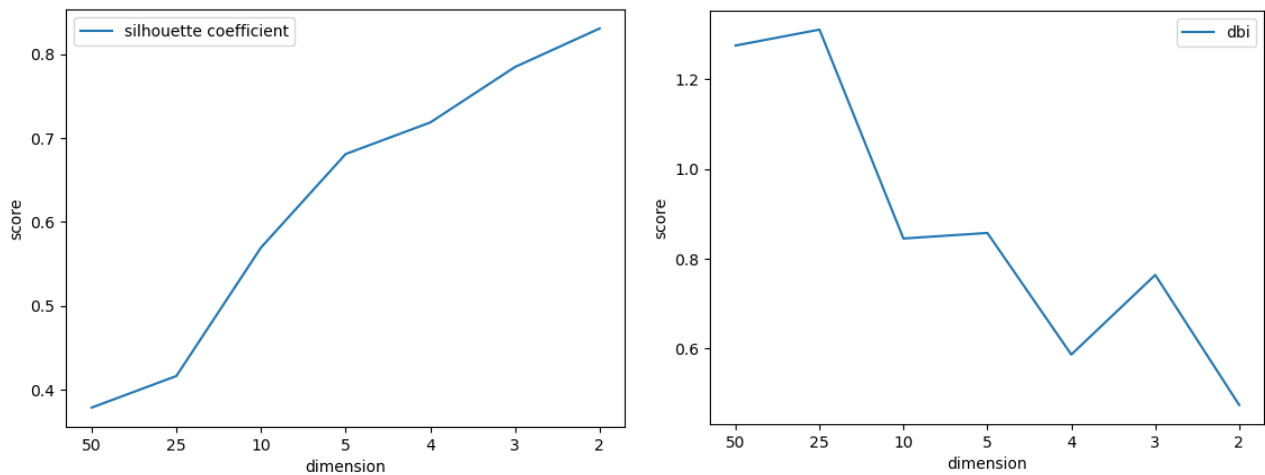


Figure 4.20 PCA Performance (k=3)

### 4.2.2 Elbow Method

The elbow method is a technique to determine the optimal number of clusters. By plotting the inertia over a range of clusters, the significant change of the point in the rate of decline can be identified. Then, select the point as the optimal number of clusters. The result is shown in Figure 4.21, “elbow” point is identified as 6. The Python code is put in Appendix B.

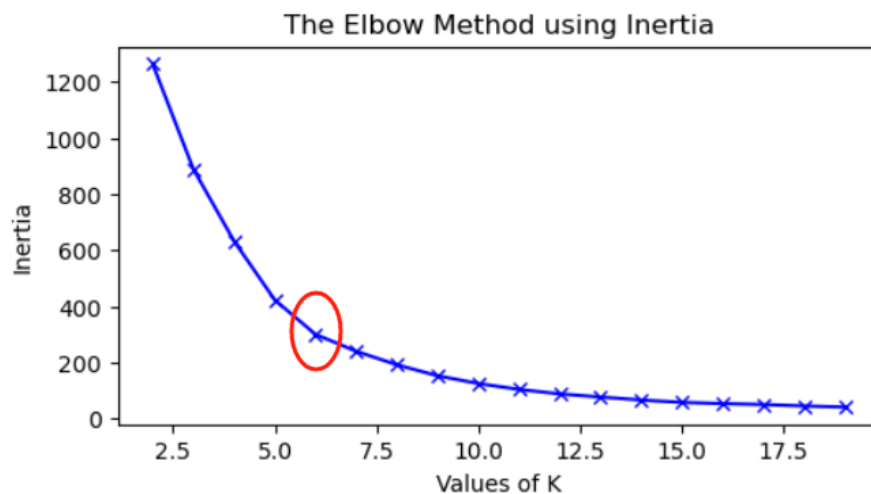


Figure 4.21 Elbow Method

### **4.2.3 Generate K-means based on selected k**

As aforementioned, the input data is the vector with 4 dimensions and the optimal number of clusters is 6. Besides, setting the iteration of times that K-means would run with different sets of starting points as the default number 10, due to there is no difference in result with the increase of the iteration times through experimentation. Also, setting the random state as 0 so the random number generation for centroid initialization is deterministic, which leads to the same clusters every time. Then, by the K-means algorithm, the input data has been trained and grouped into clusters as the final result. The silhouette coefficient and DBI will be used to evaluate the model performance. The Python code is put in Appendix B.

### **4.2.4 Model Evaluation**

The silhouette coefficient in Equation 3.1 measures the similarity of a data point with other data points inside the cluster and outside the cluster. Also, the DBI in Equation 3.2 calculates the average value of the maximum ratio of the intra-cluster distance and the inter-cluster.

## **4.3 Summary**

In this chapter, the overall flow of data preparation and K-means model development are explained thoroughly. In the next chapter, model outcomes will be analyzed.



## CHAPTER 5

### OUTCOMES DISCUSSION

#### 5.1 Introduction

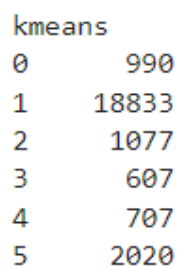
In this chapter, the performance of the model will be evaluated, clustering result will be analysed in detail and interpreted by the dashboard.

#### 5.2 Performance Evaluation

The silhouette coefficient is 0.808 and the DBI is 0.604 for the clustering, the clustering performance is good since the silhouette coefficient is close to 1. The data point is compact within the cluster it belongs to and far away from other clusters. Meanwhile, DBI is not high so it has a good separation of the clusters and is compact inside the clusters.

#### 5.3 Result Analysis

There are 6 clusters generated by the K-means model. The distribution of data points is shown in Figure 5.1. The cluster 2 has the most data points of 18833, cluster 4 and cluster 5 have the least data points of 607 and 707.



kmeans	
0	990
1	18833
2	1077
3	607
4	707
5	2020

Figure 5.1 Data Points Distribution

The overview of the skill requirements can be obtained by combining the clustering result with other features, which can identify whether the job description contains the skill grouped in the cluster (contain as 1, not contain as 0). Then, the objective of the thesis that analysing the most demanded skillsets in data-related jobs can be achieved. The shortcut of the combination result is shown in Figure 5.2.

	job_title	location	career_level	experience	job_type	job_description	skill_group1	skill_group2	skill_group3	skill_group4	skill_group5	skill_group6
0	data analyst	johor bahru	non-executive	not specified	full-time	zempot malaysia sdn bhd is a newly branched co...	1	1	1	1	1	0
1	director- data analytics	petaling jaya	senior manager	12	full-time	job overviewwe are looking for a highly motiva...	0	1	1	0	0	0
2	solutions architect - data lake specialist (ba...	kuala lumpur	senior executive	8	full-time	we are looking for a solutions architect who s...	0	1	1	0	1	1
3	automotive analyst – vehicle valuation	kuala lumpur	senior executive	3	full-time	descriptionthe automotive analyst will be acco...	1	1	1	0	1	0
4	it business analyst	kuala lumpur	senior executive	4	full-time	position objective:-responsible to be the it b...	1	1	1	0	0	0
5	system analyst	kuala lumpur	junior executive	2	full-time	job purpose as part of our expansion plan in m...	1	1	1	0	0	1
6	data scientist / artificial intelligence assoc...	kuching	senior executive	2	full-time	job summarywe are looking for a data scientist...	1	1	1	0	0	0
7	data engineer (business intelligence)	kuala lumpur	junior executive	2	full-time	why us?in aia digital+, we serve as aia's grou...	1	1	1	0	0	0
8	analyst, it business	selangor	junior executive	not specified	full-time	job responsibilities:identify the needs of the...	0	1	1	0	0	1
9	associate executive - data management & system...	sarawak	junior executive	3	full-time	job description:handling pmsys (performance ma...	0	1	1	0	1	1
10	data analytics developer / data analyst	kuching	junior executive	2	full-time	job summary:your focus will be on designing, c...	0	1	1	0	0	0
11	asset management specialist (data center opera...	johor	junior executive	3	full-time	the asset management specialist will work in t...	0	1	1	0	1	1
12	database administrator	kuala lumpur	senior executive	3	full-time	benefits of this role:13th months salary (upo...	1	1	1	0	1	1
13	scm analyst	kuala lumpur	senior executive	3	contract	1. supporting management in communication with...	1	1	1	1	1	1
14	senior finance analyst (fp&a) new req	perai	senior executive	8	full-time	key areas of responsibility: • the financial p...	1	1	1	0	1	1

Figure 5.2 First 15 Samples of Clustering Distribution

### 5.3.1 Result of Initial Clusters

The pie chart is plotted to observe the distribution of each cluster among all 11184 jobs. As shown in Figure 5.3, all jobs mentioned skills grouped in cluster 2 and cluster 3. 42.8% of jobs mentioned skills in cluster 1, 10.8% of jobs mentioned skills in cluster 4, 50.4% mentioned skills in cluster 5 and 61% mentioned skills in cluster 6.



Figure 5.3 Initial Observation of Skillset Distribution

### 5.3.2 Validation of Words in Each Cluster

To explore the textual-data validity in clusters, the word’s length of skills is checked for each cluster. In Table 5.2, the words whose length are less and equal to 8 characters are displayed. It is shown that for cluster 2, the words “c” and “r” only have one character, and there are many words consisting of a few characters that are hard to understand or explain. Besides, for cluster 3, the word “data” has 4 characters, which may lead to the 100% mention in all job descriptions. The table shows that when the word’s length achieves 8 characters, the skills are more understandable and meaningful in each cluster. Therefore, if the word has less than 8 characters, it is removed from the skillsets to improve the availability and informativeness of skills in each cluster.

Table 5.1 Skill Validation

<b>Length of Word (No. of character)</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>
Cluster 1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	{ 'analysis' }
Cluster 2	{ 'c', 'r' }	{ 'js', 'gl', 'fx', '3g', 'sw', '5g', 'gb', 'mm', 'ai', 'fw', 'ge', 'it' ... }	{ 'sap', 'gx2', 'gmp', 'wcf', 'ehs', 'bas', 'ftp' ... }	{ 'tool', 'awwa', 'xero', 'ssrs', 'sara', ... }	{ 'revit', 'vxlan', 'moves', 'ansys', 'regex', 'react', 'apics', 'alarm', ... }	{ 'capcut', 'setter', 'vbasas', 'xstore', 'vmware', 'filing', 'safety' ... }	{ 'fastapi', 'hlookup', 'tik tok', 'as2java', 'flutter', 'filmora', 'editing' ... }	{ 'problem-', 'balances', 'labeling', 'orcaflex', 'std work', 'sftp/ftp', 'teaching', 'clusters' ... }
Cluster 3	N/A	N/A	N/A	{ 'data' }	N/A	N/A	N/A	{ 'dsd data', 'data api', 'data etl', 'big data' }
Cluster 4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cluster 5	N/A	N/A	N/A	N/A	{ 'excel' }	N/A	N/A	{ 'ms excel' }
Cluster 6	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

### 5.3.3 Result of Final Clusters

After removing the unavailable skill words, the distribution of the clustering result is shown in Figure 5.4. The skills grouped in cluster 2 are the most demanded skillsets in data-related jobs, which are almost 100% mentioned among 11184 job descriptions. Then, skills in cluster 6 are highly required which are mentioned by 61% of jobs. Skills in cluster 1 and cluster 3 are medium-level required, which is mentioned by 42.8% and 36% of jobs. Lastly, skills in cluster 4 and cluster 5 are less required with the percentage of 10.8% and 22.3% mentioned.



Figure 5.4 Skillset Distribution of Each Cluster (validated)

To explore the correlation and characteristics among skills in each cluster, the most frequent 30 skills are displayed in Table 5.3. As shown in table, the most demanded skills in Malaysia are in cluster 2, with the central theme of programming skills and relevant fields of customers, accounting, sales, troubleshooting, communication and so forth. Next, skills in cluster 6 are highly required with the central theme of management, which indicates that in Malaysia, jobs for data-related industries value management ability. The most required ability includes project management, inventory management, database management, risk and budget management, etc. Next highly required skills are in cluster 1 with the central theme of financial and analysis, which indicates jobs for data-related industries are often related to the financial field and require analysis from various aspects, including risk, market, cost, competitors, and so forth. The skills in cluster 3 are commonly required, which are about how to manipulate data, from data entry, data collection, data management, data modelling to visualization, etc. Then, skills in Microsoft and Excel in cluster 5 are required. Lastly, skills that mainly indicate the way of data analysis are the least required in Malaysia, which involves tools, software, and reporting of the data analysis.

Table 5.2 Most common skills in cluster

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
financial analysis', risk analysis', financial reporting', financial reports', root cause analysis', market analysis', statistical analysis', cost analysis', variance analysis', system analysis', sales analysis', financial modeling', financial reporting software', competitive analysis', competitor analysis', financial software', financial accounting', financial analysis software', market research and analysis', chemical analysis', analysis', financial statements', financial planning', research and analysis', problem analysis', trend analysis', business process analysis', statistical analysis tools', cost-benefit analysis', business analysis')	sql', customer service', accounting', python', sales', budgeting', javascript', troubleshooting', java', reporting', powerpoint', market research', problem-solving', process improvement', r', quality control', communication', accounting software', analytical skills', record keeping', auditing', crm software', documentation', filing', word', forecasting', mysql', negotiation', tax preparation', marketing')	data entry', data management', data collection', data analytics', data visualization', data processing', data modeling', data transformation', data mining', data warehousing', data migration', data extraction', data visualization tools', data integration', data compilation', data recording', data cleansing', data validation', data modelling', data governance', data accuracy', data architecture', google data studio', data structures', data quality management', data verification', master data maintenance', data entry software', data analysismicrosoft excel data maintenance')	data analysis', data analysis tools', data analysis software', data analysis and reporting', data collection and analysis', financial data analysis', sales data analysis', statistical data analysis', data trend analysis', technical data analysis', business data analysis', research and data analysis', data analysis and visualization', data entry and analysis', data processing and analysis', project data analysis', performance data analysis', basic data analysis', data and problem analysis', tableau data analysis', spreadsheet data analysis', data verification and analysis', big data analysis', data relationship analysis', data analysis tools e.g. excel', process data analysis', document data analysis', data/systems analysis and development data analysis and management')	microsoft excel', excel', microsoft word', microsoft powerpoint', microsoft office', microsoft office suite', ms excel', microsoft azure', microsoft office proficiency', microsoft outlook', excel proficiency', microsoft', microsoft sql server', project managementmicrosoft excel', microsoft excel proficiency', microsoft project', microsoft office word', microsoft office applications', microsoft office products', ms office excel', microsoft sql', microsoft server', microsoft iis', excel formula', advanced excel', microsoft excel 2013 onwards', excel vlookup', inventory managementmicrosoft excel', excel sheets', filingmicrosoft excel')	project management', inventory management', database management', time management', project management software', risk management', budget management', customer relationship management', inventory management software', supply chain management', vendor management', social media management', cash flow management', sales management', payroll management', compliance management', team management', project management tools', stakeholder management', stock management', warehouse management', logistics management', quality management', contract management', database management software', financial management', document management', cost management', cash management', performance management')



## **5.4 Insight by Dashboard**

The dashboard for cluster 2 is shown in Figure 5.5. It can be seen that for programming skills, SQL is the most required, followed by Python, JavaScript and Java. Apart from programming skills, skills in customer service are the most required. Then, skills in accounting, sales and budgeting are required, which are all related to the financial field. Besides, skills in troubleshooting and reporting are commonly required as well. As for the job locations, most jobs are located in Kuala Lumpur, Petaling Jaya, Selangor, Penang and Shah Alam. The requirement for job experience is mostly from 1 year to 2 years. The 88% of jobs require full-time workers. For the career level, the most demanded are junior executives and senior executives, followed by managers and entry-level workers. More dashboards of the other 5 clusters have been put into Appendix C.

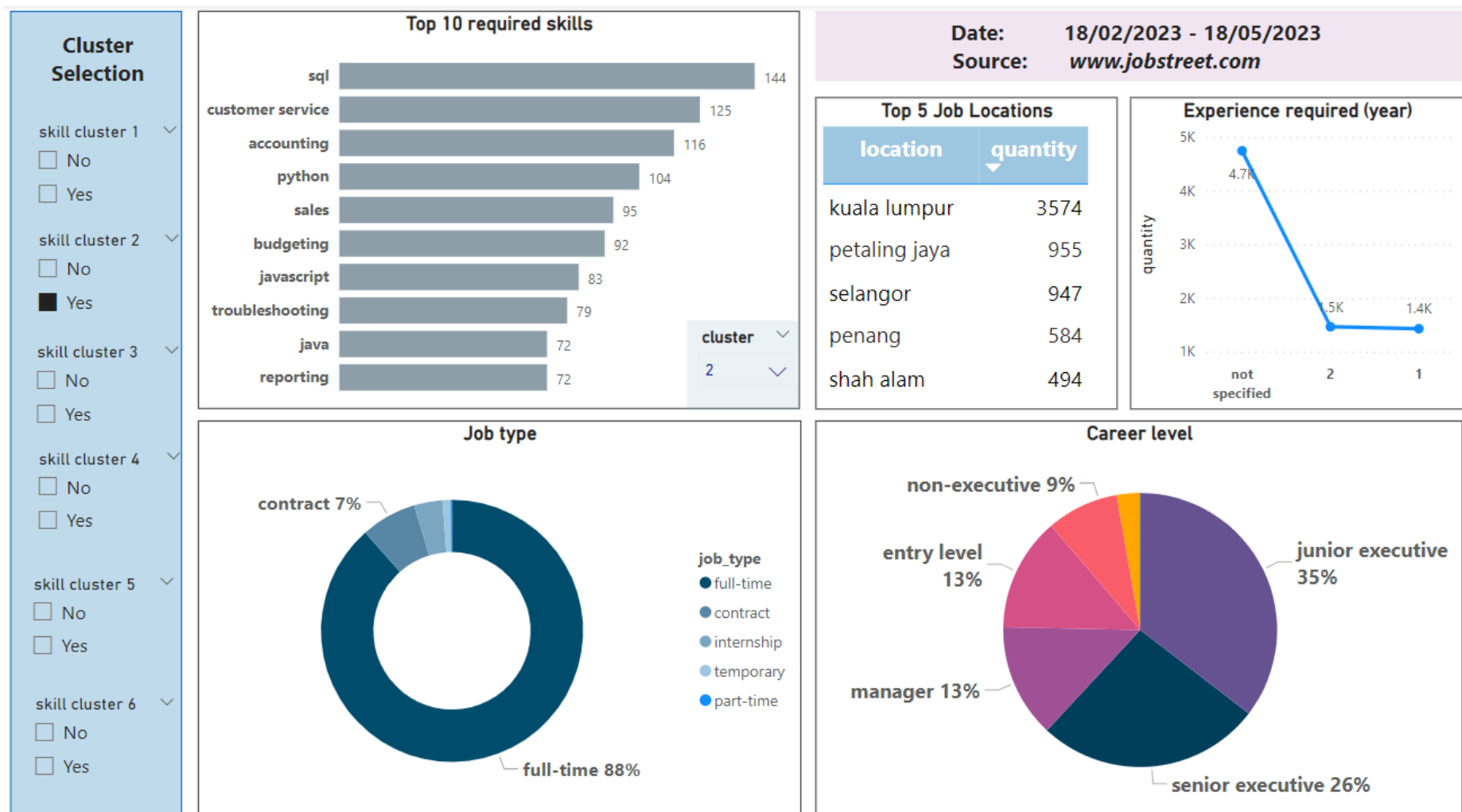


Figure 5.5 Insight of Job Information (Cluster 2)

## **5.5 Summary**

In this chapter, the clustering result from a large number of job descriptions is analyzed and interpreted by the dashboard. The relationship between job skills and related job information is clearly illustrated, in other words, the trends and manifesting technologies in the Malaysian data-related labor market are identified: most data-related jobs are located in Kuala Lumpur; companies prefer employees with 1-3 years of experience; companies in Malaysia prefer full-time employees; companies have the most job vacancy for the executive level, less vacancy for non-executive level jobs. With the help of result analysis, job seekers are able to gain useful information and match themselves with jobs in Malaysian data-related industries.

## **CHAPTER 6**

### **CONCLUSION**

#### **6.1 Introduction**

The result of the clustering has been analysed and the interpretation has been illustrated by the dashboard. In this chapter, the accomplishment of the thesis objectives will be stated, besides, the limitation and future works of the project will be discussed.

#### **6.2 Achievements of Project Objectives**

All objectives of the thesis are achieved. Objective 1 to build a dataset of data-related job postings has been completed by scraping data from the Jobstreet website. Objective 2 to extract useful features for the job skills analysis has been completed by the GPT-3.5 model. Objective 3 to develop a K-means model has been completed with the parameter tuning by PCA and elbow method. Objective 4 to analyse the clustering outcomes and display findings through a dashboard has been completed. The most demanded data-related job skillsets were identified, the trends and manifesting technologies in the Malaysian data-related labor market were displayed by the dashboard. Overall, the progress of the thesis keeps in line with the aim of identifying the skillset groups and the most required skills in data-related industries.

#### **6.3 Project Limitation**

The dataset used in the thesis only collects the job postings published from 18/02/2023 to 18/05/2023, and the data source is single. Firstly, since the trend in labour market keeps changing, if the data is not up-to-date and only covers a short period, the outcomes analysis may not fit the latest situation. Secondly, due to website restrictions only one website is used as the data source, the outcomes may not represent

the overall situation of the Malaysian labour market. Thirdly, the GPT-3.5 model as the advanced AI techniques, is freshly released and still undergoing development. Its stability and maturity in processing text data can't be 100% guaranteed, so the outcomes generated by the GPT-3.5 model have flaws in some way. Also, there is a rate limitation of using the GPT-3.5, so the text can't be processed at a fast speed. Lastly, there is no standard method to select relevant job features among job postings, so the model result can't promise superb performance.

## **6.4 Future Work**

This thesis has several potential areas for improvement and future works. Firstly, if there is more preparation time for data collection, the dataset can have the larger size and more job information can be extracted. Secondly, this thesis accepted the outcomes from GPT model as the final result of text processing. It can be improved by combining the human work and AI model to increase the quality and availability of the processed data.

For future work, the thesis analysis can be generalized as shown in Figure 6.1. By setting the filter on the left side, the requirement distribution towards skills in each cluster is displayed on the right side, and relevant job titles are listed as well. For instance, if the job seeker wants to search for a data-related job in Johor, with the career level of manager and the type of full-time, the requirement level for skills in each cluster is displayed. From the result, skills in cluster 2 and cluster 6 are the most mentioned. Job titles for the job seeker include account manager, assistant IT manager, etc. If there is more job information available, the application of the analysis generalization can be more accurate and complete.

Filter

location

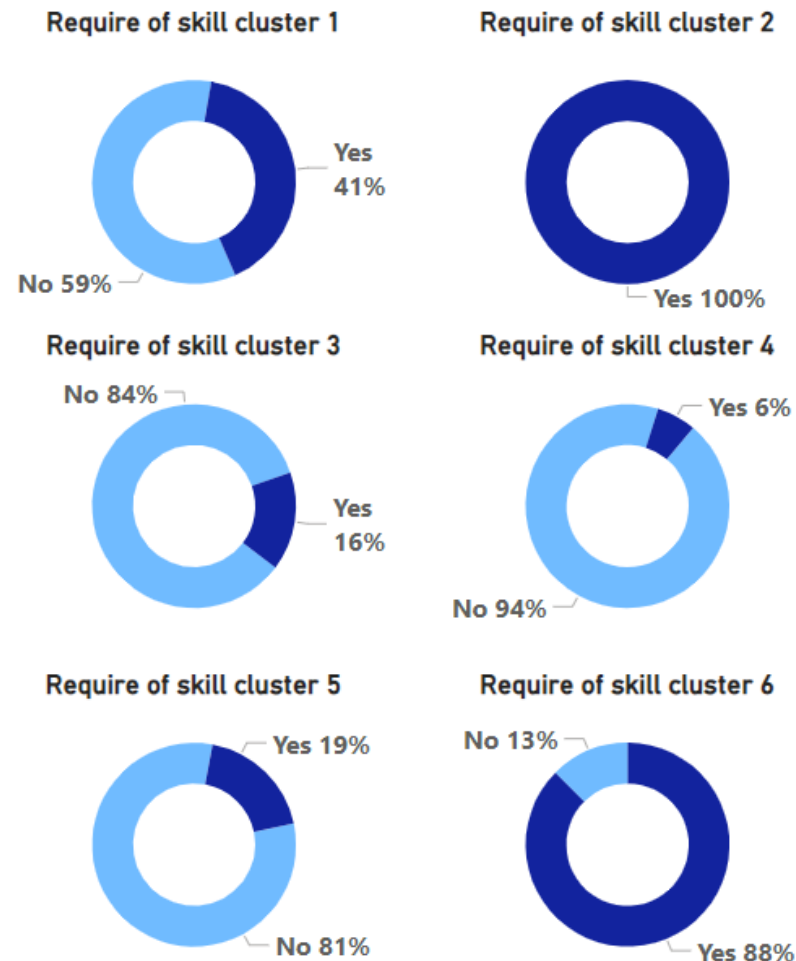
johor

career\_level

☐ entry level
☐ junior executive
☒ manager
☐ non-executive
☐ not specified
☐ senior executive
☐ senior manager

job\_type

☒ full-time



job_title
account manager
administrative manager
anodizing process manager
assistant it manager
assistant manager (process engineering) - bio-tech/chemical manufactures
commercial manager
digital marketing manager
electrical engineer (kulai, johor)
elv engineer (kulai, johor)
it manager
maintenance manager (kulai, johor)
maintenance manager (up to 18k / kulai / data center)
manager - production packing
manager, critical operations (kulai, johor)
manager, documentation (kulai, johor)
manager, human resources
planning manager
plant manager (us mnc / rubber manufacturing / gelang patah)
production manager
production manager (id: 569042)
purchasing manager
qa manager
quality process specialist
retail store manager

Figure 6.1 Generalized Analysis Example

## REFERENCES

- Almaleh, A., Aslam, M.A., Saeedi, K. and Aljohani, N.R. (2019). "Align My Curriculum: A Framework to Bridge the Gap between Acquired University Curriculum and Required Market Skills." *Sustainability*, 11(9), p.2607.
- Almgerbi, M., De Mauro, A., Kahlawi, A. and Poggioni, V. (2021). "A Systematic Review of Data Analytics Job Requirements and Online-Courses." *Journal of Computer Information Systems*, 62(2), pp.422–434.
- Bach, M.P., Krstić, Ž., Seljan, S. and Turulja, L. (2019). "Text Mining for Big Data Analysis in Financial Sector: A Literature Review." *Sustainability*, 11(5), p.1277.
- Banerjee, A., Dhillon, I., Ghosh, J. and Lafferty, J. (2005). "Clustering with Bregman Divergences." *Journal of Machine Learning Research*, 6, pp.1705–1749.  
Available at:  
<https://www.jmlr.org/papers/volume6/banerjee05b/banerjee05b.pdf>
- Chan, A. (2022). "GPT-3 and InstructGPT: technological dystopianism, utopianism, and ‘Contextual’ perspectives in AI ethics and industry." *AI and Ethics* 3, pp.53–64.
- Chen, A. (2017). *Projected new jobs by major occupational group, 2016–26: Career Outlook: U.S. Bureau of Labor Statistics*. [www.bls.gov](http://www.bls.gov). Available at:  
[https://www.bls.gov/careeroutlook/2017/data-on-display/projections-occupational-group.htm?view\\_full](https://www.bls.gov/careeroutlook/2017/data-on-display/projections-occupational-group.htm?view_full).

- De Mauro, A., Greco, M., Grimaldi, M. and Ritala, P. (2018). "Human resources for Big Data professions: A systematic classification of job roles and required skill sets." *Information Processing & Management*, 54(5), pp.807–817.
- Debao, D., Yinxia, M. and Min, Z. (2021). "Analysis of big data job requirements based on K-means text clustering in China." *PLOS ONE*, 16(8), p.e0255419.
- Francesco, T., Qinghua, T., Panagiotis, P. and Johan, S. (2023) "Deep Kernel Principal Component Analysis for multi-level feature learning." *Neural Networks*, 2023.11.045
- Grüger, J. and Schneider, G. (2019). "Automated Analysis of Job Requirements for Computer Scientists in Online Job Advertisements." *Proceedings of the 15th International Conference on Web Information Systems and Technologies*.
- Gurcan, F. and Cagiltay, N.E. (2019). "Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets Using LDA-Based Topic Modeling." *IEEE Access*, 7, pp.82541–82552.
- Hayati, H., Khalidi , I.M. and Bennani, S. (2020). "Automatic Classification for Cognitive Engagement in Online Discussion Forums: Text Mining and Machine Learning Approach." *Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science*, 12164, pp.114–118.
- Jeong, B., Yoon, J. and Lee, J.M. (2019). "Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis." *International Journal of Information Management*, 48, pp.280–290.



- Jothi, A., Bhargavi, B. and Rani, D. (2023). "Prediction of dyslexia severity levels from fixation and saccadic eye movement using machine learning," *Biomedical Signal Processing and Control*, 79(1), 104094.
- Kim, J., Warga, E. and Moen, W. (2013). "Competencies Required for Digital Curation: An Analysis of Job Advertisements." *International Journal of Digital Curation*, 8(1), pp.66–83.
- Linkedin.com. (2017). *LinkedIn's 2017 U.S. Emerging Jobs Report*. Available at: <https://economicgraph.linkedin.com/research/LinkedIns-2017-US-Emerging-Jobs-Report>.
- Lyu, W. and Liu, J. (2021). "Soft skills, hard skills: What matters most? Evidence from job postings." *Applied Energy*, 300, p.117307.
- Markow, W., Braganza, S., Taska, B., Miller, S.M. and Hughes, D. (2017). *THE QUANT CRUNCH HOW THE DEMAND FOR DATA SCIENCE SKILLS IS DISRUPTING THE JOB MARKET*. Available at: [http://www.burning-glass.com/wp-content/uploads/The\\_Quant\\_Crunch.pdf](http://www.burning-glass.com/wp-content/uploads/The_Quant_Crunch.pdf).
- Mirjana, P.B., Bertoncel, T., Meško, M. and Krstić, Ž. (2020). "Text mining of industry 4.0 job advertisements." *International Journal of Information Management*, 50, pp.416–431.
- Mohamed, A.A. (2020). "An effective dimension reduction algorithm for clustering Arabic text." *Egyptian Informatics Journal*, 21(1), pp.1–5.

- Oh, H., Park, S., Lee, G.M., Heo, H. and Choi, J.K. (2019). "Personal Data Trading Scheme for Data Brokers in IoT Data Marketplaces." *IEEE Access*, 7, pp.40120–40132.
- Ramzan, M.J., Khan, S.U.R., Inayat, U.R., Khan, T.A., Akhunzada, A. and Naseeb, C. (2021). "A Conceptual Model to Support the Transmuters in Acquiring the Desired Knowledge of a Data Scientist." *IEEE Access*, 9, pp.115335–115347.
- Sabri, T., Beggar, O.E. and Kissi, M. (2022). "Comparative study of Arabic text classification using feature vectorization methods." *Procedia Computer Science*, 198, pp.269–275.
- Saka, A., Taiwo, R., Saka, N., Salami, B. A., Ajayi, S., Akande, K., and Kazemi, H. (2024). "GPT models in construction industry: Opportunities, limitations, and a use case validation." *Developments in the Built Environment*, 17, p.100300.
- Smaldone, F., Ippolito, A., Lagger, J. and Pellicano, M. (2022). "Employability skills: Profiling data scientists in the digital labour market." *European Management Journal*. j.emj.2022.05.005.
- Sridevi, G. and Suganthi, S.K. (2022). "AI based suitability measurement and prediction between job description and job seeker profiles." *International Journal of Information Management Data Insights*, 2(2), p.100109.
- Thielen, J. and Neeser, A. (2020). "Making Job Postings More Equitable: Evidence Based Recommendations from an Analysis of Data Professionals Job Postings Between 2013-2018." *Evidence Based Library and Information Practice*, 15(3), pp.103–156.

- Usabiaga, C., Núñez, F., Arendt, L., Gałęcka-Burdziak, E. and Pater, R. (2022). "Skill requirements and labour polarisation: An association analysis based on Polish online job offers." *Economic Modelling*, 115, p.105963.
- Uymaz, H.A. and Metin, S.K. (2022). "Vector based sentiment and emotion analysis from text: A survey." *Engineering Applications of Artificial Intelligence*, 113, p.104922.
- Wijaya, Y.A., Kurniady, D.A., Setyanto, E., Tarihoran, W.S., Rusmana, D. and Rahim, R. (2021). "Davies Bouldin Index Algorithm for Optimizing Clustering Case Studies Mapping School Facilities." *TEM Journal*, 10(3), pp.1099–1103.
- Xiao, Q., Xin, Z. and Chenghua, Z. (2019). "Application Research of KNN Algorithm Based on Clustering in Big Data Talent Demand Information Classification." *International Journal of Pattern Recognition and Artificial Intelligence*, 34(06), p.2050015.
- Xu, X., Wang, X., Li, Y. and Haghighi, M. (2017). "Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors." *International Journal of Information Management*, 37(6), pp.673–683.

## Appendix A Gantt Chart

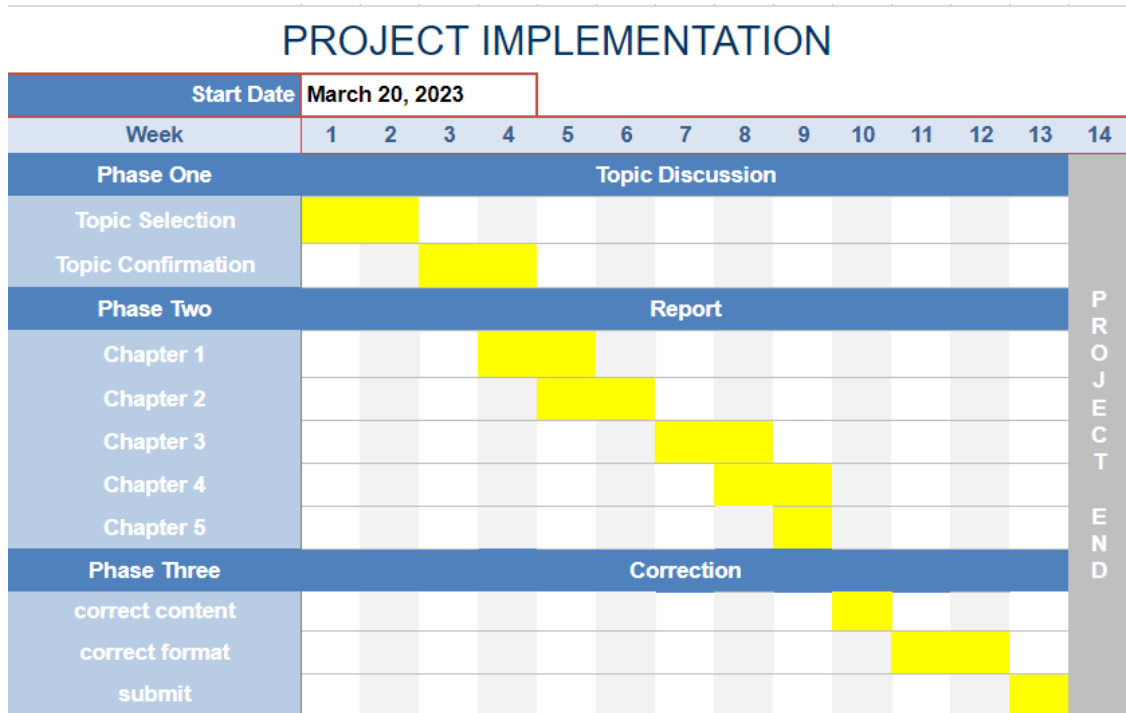


Figure A. 1 Gantt chart for Project 1

### Project 2

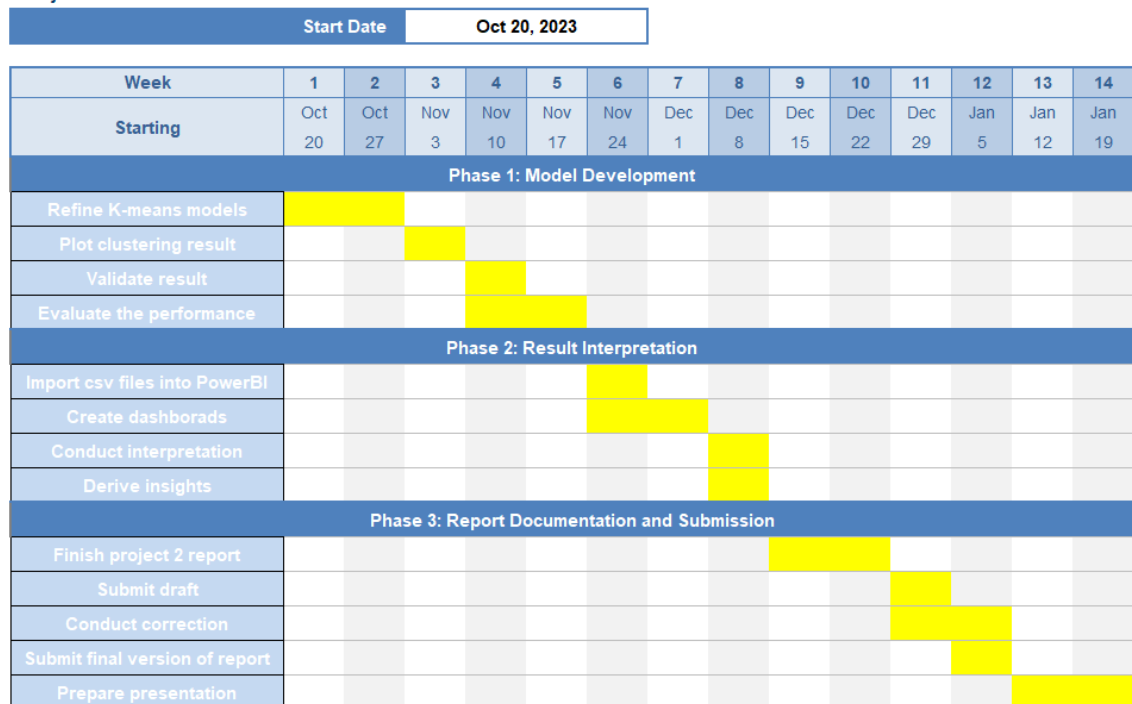


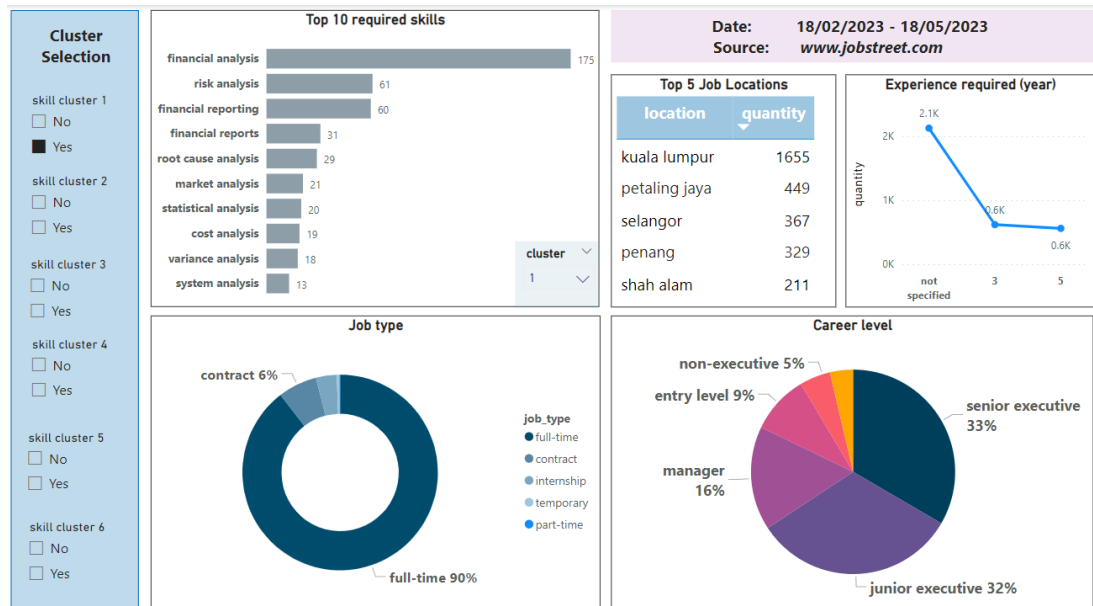
Figure A. 2 Gantt chart for Project 2

## **Appendix B   Code**

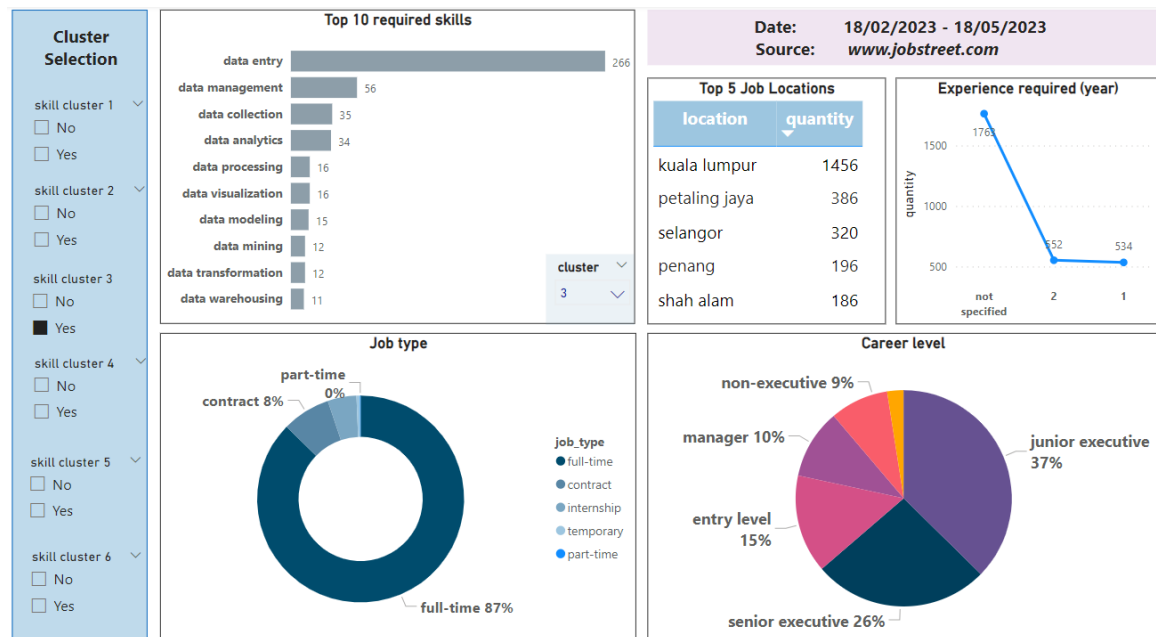
The implementation of all phases is using Python programming language. The source code can be accessed from the GitHub platform with the provided link <https://github.com/qiting3158/FYP-code>

## Appendix C Dashboard of Clusters

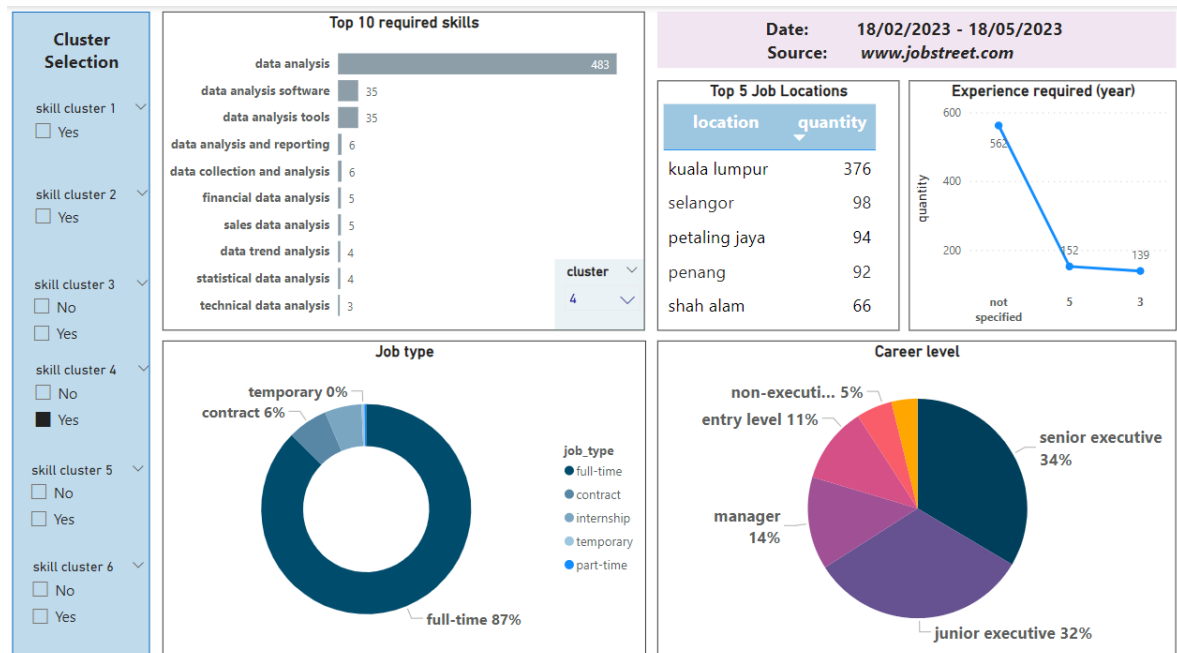
### Cluster 1 Output:



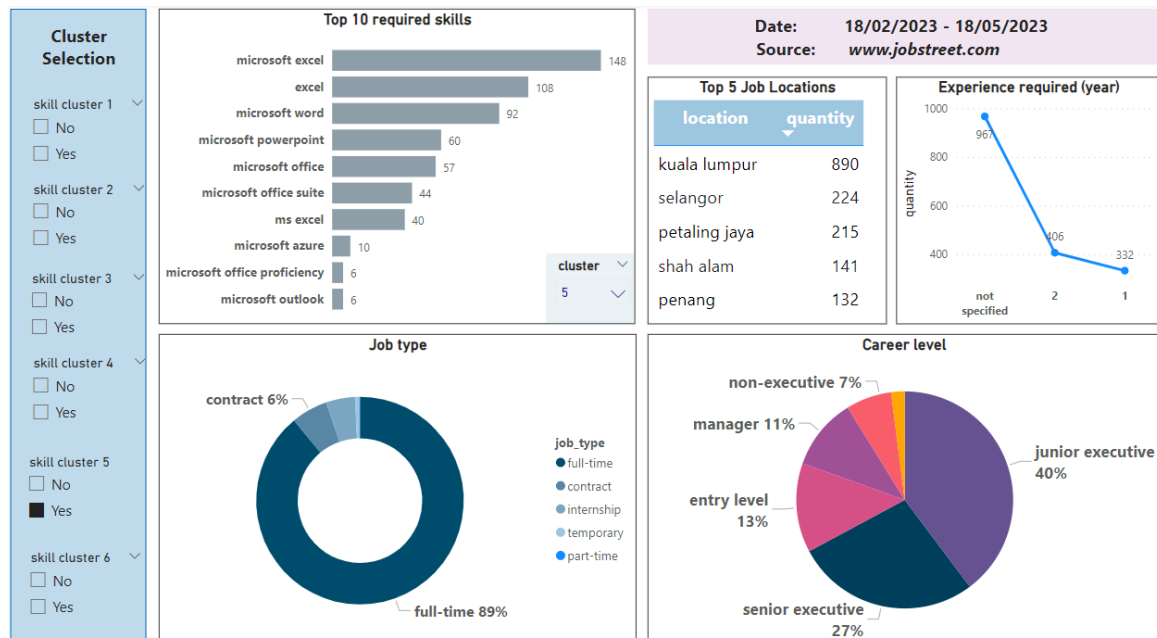
### Cluster 3 output:



## Cluster 4 output:



## Cluster 5 output:



Cluster 6 output:

