**K-MEANS CLUSTERING FOR INDENTIFICATION OF TRENDING SKILL SETS ON JOB ADVERTISEMENT SITES**

**WONG TUCK YEW**

**A project report submitted in partial fulfilment of the**

**requirements for the award of Bachelor of Science**

**(Hons.) Software Engineering**

**Lee Kong Chian Faculty of Engineering and Science**

**Universiti Tunku Abdul Rahman**

**APRIL 2019**

DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at UTAR or other institutions.

|  |  |  |
| --- | --- | --- |
| Signature | : |  |
| Name | : |  |
| ID No. | : |  |
| Date | : |  |

APPROVAL FOR SUBMISSION

I certify that this project report entitled **“K-MEANS CLUSTERING FOR IDENTIFICATION OF TRENDING SKILL SETS ON JOB ADVERTISEMENT WEBSITES”** was prepared by **WONG TUCK YEW** has met the required standard for submission in partial fulfilment of the requirements for the award of Bachelor of Science (Hons.) in Software Engineering at Universiti Tunku Abdul Rahman.

Approved by,

|  |  |  |
| --- | --- | --- |
| Signature | : |  |
| Supervisor | : |  |
| Date | : |  |

The copyright of this report belongs to the author under the terms of the copyright Act 1987 as qualified by Intellectual Property Policy of Universiti Tunku Abdul Rahman. Due acknowledgement shall always be made of the use of any material contained in, or derived from, this report.

© 2019, Wong Tuck Yew. All right reserved.

ACKNOWLEDGEMENTS

I would like to thank everyone who had contributed to the successful completion of this project. I would like to express my gratitude to my research supervisor, Dr. Simon Lau Boung Yew and Dr. Victor Tan Hock Kim for their invaluable advice, guidance and their enormous patience throughout the development of the research.

Lastly, I would also like to express my gratitude to my loving parents and friends who had helped and given me encouragement in this project.

ABSTRACT

Nowadays, online job advertisement websites such as indeed.com.my, monster.com and jobstreet.com.my have become the effective pipeline for employers to hire the employees that having the desired skill sets in worldwide social media. However, the job seekers are not able to track the trending skill sets that are required in the market by using job advertisement websites. This is an underlying issue that job seekers also do not know what skill sets that are needed to be acquired in the specific location. The project mainly aims to develop an K-means text clustering model to identify the trending skill sets on job advertisement websites. Unimportant texts filtering and TF-IDF technique is used to measure the weight score of each important terms before perform K-Means text clustering algorithm. The keywords are clustered into 8 clusters and around 5000 job titles are distributed into these 8 clusters. By visualizing the results generated by the K-Means clustering model, employers or job seekers can perform analysis to identify the most desired skill sets in area of Singapore and Malaysia in year 2019. In this project, the trending skill sets in area of Singapore and Malaysia in 2019 are identified which includes 2485 jobs are required business project management skills, 719 jobs are required technical support skills, 708 jobs are required web development skills, 696 jobs are required software testing skills, and 90 jobs are required cyber security consultation skills. As a conclusion, this research project are able to help the job seekers to understand the trending skill sets in today’s market, which they will have the higher opportunity and chance to get hired if they acquire the welcomed skill sets.

TABLE OF CONTENTS

[DECLARATION 2](#_Toc6531943)

[APPROVAL FOR SUBMISSION 3](#_Toc6531944)

[ACKNOWLEDGEMENTS 5](#_Toc6531945)

[ABSTRACT 6](#_Toc6531946)

[TABLE OF CONTENTS vii](#_Toc6531947)

[LIST OF TABLES x](#_Toc6531948)

[LIST OF FIGURES xi](#_Toc6531949)

[LIST OF SYMBOLS / ABBREVIATIONS xii](#_Toc6531950)

**CHAPTER**

[1 INTRODUCTION 13](#_Toc6618931)

[1.1 Background 13](#_Toc6618932)

[1.2 Problem Statement 14](#_Toc6618933)

[1.2.1 Universities or colleges provide outdated courses in education plan 14](#_Toc6618934)

[1.2.2 Market research is expensive 14](#_Toc6618935)

[1.2.3 Job seeker do not or do not know which expert skill set should be acquired 14](#_Toc6618936)

[1.3 Project Objectives 15](#_Toc6618937)

[1.4 Project Solution 15](#_Toc6618938)

[1.4.1 Explore, gather and filter job advertisement websites 15](#_Toc6618939)

[1.4.2 Pull all the text data and remove irrelevant text data 15](#_Toc6618940)

[1.4.3 Extract and store data into database 16](#_Toc6618941)

[1.4.4 Perform k-means text clustering algorithm 16](#_Toc6618942)

[1.4.5 Visualize the patterns of data 16](#_Toc6618943)

[1.4.6 Analyze group patterns to obtain trending skill sets 16](#_Toc6618944)

[1.5 Project Approach 17](#_Toc6618945)

[1.5.1 Understanding the job advertisement websites 17](#_Toc6618946)

[1.6 Scope of the Project 17](#_Toc6618947)

[2 LITERATURE REVIEW 18](#_Toc6618948)

[2.1 Data Pre-processing Techniques 18](#_Toc6618949)

[1) Convert all letters to lower/ upper case 18](#_Toc6618950)

[2) Remove numbers 18](#_Toc6618951)

[3) Remove punctuations 18](#_Toc6618952)

[4) Remove stop words 18](#_Toc6618953)

[5) Stemming and Lemmatization 18](#_Toc6618954)

[2.2 Feature Extraction 19](#_Toc6618955)

[2.3 Machine Learning Algorithms 21](#_Toc6618956)

[2.4 Clustering Technique 21](#_Toc6618957)

[2.4.1 Unsupervised Learning Algorithms 21](#_Toc6618958)

[2.4.2 Text data mining 26](#_Toc6618959)

[2.5 Trending Skill Sets in Job Advert Websites 26](#_Toc6618960)

[2.6 Visualizing techniques 27](#_Toc6618961)

[3 METHODOLOGY AND WORK PLAN 29](#_Toc6618962)

[3.1 Introduction 29](#_Toc6618963)

[3.1.1 Implementing web scraping module 29](#_Toc6618964)

[3.1.2 Implementing clustering module 29](#_Toc6618965)

[3.1.3 Implementing visualizing module 30](#_Toc6618966)

[3.1.4 Combining the modules 30](#_Toc6618967)

[3.2 Tools and techniques 30](#_Toc6618968)

[4 PROJECT INITIAL SPECIFICATION 33](#_Toc6618969)

[4.1 Introduction 33](#_Toc6618970)

[4.2 Functional requirements 33](#_Toc6618971)

[4.3 Non-functional requirements 34](#_Toc6618972)

[4.4 Data flow diagram (DFD) 34](#_Toc6618973)

[4.5 Project Initiation 34](#_Toc6618974)

[4.5.1 Investigation for several job advertisement sites 34](#_Toc6618975)

[4.5.2 Prototyping 35](#_Toc6618976)

[5 PROJECT IMPLEMENTATION 38](#_Toc6618977)

[5.1 Introduction 38](#_Toc6618978)

[5.2 Code implementation 38](#_Toc6618979)

[5.2.1 Web Scraping Module 38](#_Toc6618980)

[5.2.2 Data Storage 39](#_Toc6618981)

[5.2.3 Data Processing Module 41](#_Toc6618982)

[5.3 Result and discussion 44](#_Toc6618983)

[6 CONCLUSION 48](#_Toc6618984)

[6.1 Project Limitation 49](#_Toc6618985)

[6.1.1 Request Clean and Structured Data 49](#_Toc6618986)

[6.2 Future Development 49](#_Toc6618987)

[6.2.1 Data Complexity 49](#_Toc6618988)

[6.2.2 Job Title Search Application 49](#_Toc6618989)

[6.2.3 Utilize Other Data Visualization Techniques 49](#_Toc6618990)

[REFERENCES 50](#_Toc6618991)

LIST OF TABLES

[Table 1: Comparing Clustering Techniques 23](#_Toc6619167)

[Table 2: Clusters with keyword and job title clustered 44](#_Toc6619168)

LIST OF FIGURES

[Figure 1: Overview of Problem Statements 14](#_Toc6619510)

[Figure 2: General Steps of Development 15](#_Toc6619511)

[Figure 3: HTML Code of Job Advertisement 17](#_Toc6619512)

[Figure 4: TF-IDF Score Calculation 19](#_Toc6619513)

[Figure 5: Frequency of occurrence for top 30 terms 20](#_Toc6619514)

[Figure 6: Machine Learning (“Machine Learning at MATLAB”, n.d.) 21](#_Toc6619515)

[Figure 7: K-means clustering. Retrieved from Firdaouss Doukkali (2017) 22](#_Toc6619516)

[Figure 8: K-Medoid. Retrieved from Jiawei Han (n.d.) 23](#_Toc6619517)

[Figure 9: The elbow method 24](#_Toc6619519)

[Figure 10: Elbow method for 1 million headlines 25](#_Toc6619520)

[Figure 11: Labelled IT Dataset: Administrator (258), Analyst (898), Developer (1308), Engineer (716), Lead (952), Support (524), Tester (99). 27](#_Toc6619521)

[Figure 12: Visualize and interpret result 28](#_Toc6619522)

[Figure 13: Beautiful Soup (Available at: https://www.crummy.com/software/BeautifulSoup/bs4/doc/) 30](#_Toc6619523)

[Figure 14: MongoDB (Available at: https://docs.mongodb.com/manual/) 31](#_Toc6619524)

[Figure 15: Scikit-learn (Available at: https://scikit-learn.org/stable/) 31](#_Toc6619525)

[Figure 16: D3.js (Available at: https://d3js.org/) 32](#_Toc6619526)

[Figure 17: Project data flow diagram 34](#_Toc6619527)

[Figure 18: Web scraping prototype 36](#_Toc6619528)

[Figure 19: Text clustering prototype 37](#_Toc6619529)

[Figure 20: Login to View Salary 39](#_Toc6619530)

[Figure 21: BSON Object for Job Advertisement 40](#_Toc6619531)

[Figure 22: MongoDB Compass 41](#_Toc6619532)

[Figure 23: Elbow method (K=8) 43](#_Toc6619533)

[Figure 24: Clusters keyword and job titles clustered 43](#_Toc6619534)

[Figure 25: Clusters Job Adverts Count 47](#_Toc6619536)

[Figure 26: Clusters against number of job titles clustered 47](#_Toc6619537)

LIST OF SYMBOLS / ABBREVIATIONS

SSE sum of squared error

TF-IDF term frequency – inverse document frequency

## INTRODUCTION

### Background

One of the most common ways today's job seekers uncover employment opportunities is by using online sources. There are hundreds of job boards, both generic and niche, as well as aggregators, social media channels, networking groups and staffing company websites to choose from (Robert Half, 2017). The online job advertisement websites such as monster.com, indeed.com.my and jobstreet.com.my. The mobile application version of indeed.com.my and jobstreet.com.my have up to 50 million of download times and 1 million of download times respectively. This means that many people are applying jobs by using online job advertisement application method. Online job advertisement websites provide a very convenient way to apply for a new job which the job seekers make only a few steps to run through entire process of applying a new job no matter at anywhere.

In addition, online job advertisements create an effective pipeline for employers to hire their desired employees in worldwide social media. On the other hand, job seekers also can seek for jobs that are related to their skill sets. Thus, online job advertisement websites play an essential role for employers to hire new workers and job seekers to get employed. A minimum requirement for applying a job is job seeker must fulfil a desired skill sets that can benefits the company.

According to Izabela A. Wowczko (2016), the project aimed at extracting specific competences from job descriptions, and therefore evidence about skills needs in software engineering. The data was scrapped from online repositories, parsed, and filtered based on a set of predefined keywords. Hierarchical agglomerative clustering and K-means were the mining techniques used to identify groups of skills producing coherent job definitions.

### Problem Statement

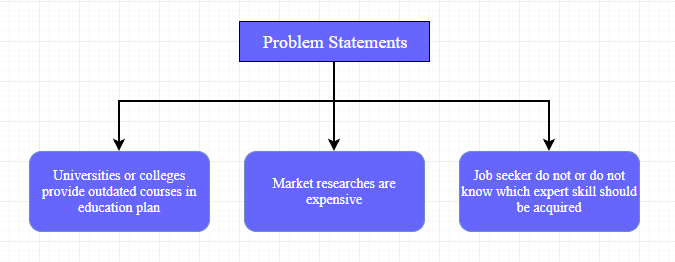


Figure 1: Overview of Problem Statements

#### Universities or colleges provide outdated courses in education plan

Job requirements may include specific skill sets, experience and knowledge in the job field and educational credential. For university student, achieving an specific skill sets is the top first requirement to get hired because most of the employers will ignore those who do not acquire the skills that can benefit their companies, before they are going to justify your personal quality. The company will reduce the pool of applicants based on skill sets before they want to search for a good personality.

#### Market research is expensive

A comprehensive market research may consume high cost and long time. Many business owners do not have the money to spend to analyse the market trend and they try to bypass this crucial step and before producing their products.

Without paying for a market research, business owners are not able to analyse by their own observations. For example, there are millions of job advertisements in job search websites where universities and colleges cannot identify the trending skill sets one by one through the job advertisement.

#### Job seeker do not or do not know which expert skill set should be acquired

Narrowing the pool of job applicants with skill sets which is the first priority for an employer to hire a worker. Undeniably, a job seeker must have a mix skill sets to benefits the company comprehensively. According to Gary Burtless (2014), the manufacturing workers lost their jobs and cannot find a new job because they do not have the specific skill set needed by the company.

When a job seeker with the right skill set is hired, the job seeker can direct benefits the company with his specific expert skill.

### Project Objectives

The key objective of this project is to identify the trending skill sets need to be acquired from job advertisement websites using K-means text clustering algorithm. The project used text pre-processing methods to remove the unimportant texts from the job description. Then measure the weight score of each terms in the job descriptions which scarped from the job advertisement websites. By visualizing the results generated by the K-Means clustering model, an accurate and clear analysis to identify the most desired skill sets from the job advertisement sites.

### Project Solution

The proposed solution consists of 3 modules. The solution is to be developed following steps in Figure 2.

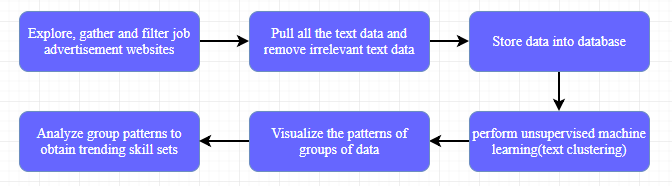


Figure 2: General Steps of Development

#### Explore, gather and filter job advertisement websites

* Explore the job search websites with job advertisements, and select the websites that contains many job advertisements.

#### Pull all the text data and remove irrelevant text data

* Pull all the text data from html, and eliminate “English words” that will affect the result of the analysis

#### Extract and store data into database

Extract and Insert all the text data in string format into NOSQL database.

#### Perform k-means text clustering algorithm

A keyword extraction technique is used to extract the important keywords in the plain text of all job descriptions. After the important keyword features are extracted, the job descriptions with its corresponding titles will then be clustered according to important keyword features. The text data then be taken from NOSQL database as BSON object.

#### Visualize the patterns of data

All the clusters that generated by the machine learning algorithm will be visualized through web front-end in browser.

#### Analyze group patterns to obtain trending skill sets

The patterns that shown in visuals are analyzed to get the accurate results of trending skill sets for every job.

### Project Approach

#### Understanding the job advertisement websites

Figure 3 shows that the html code that contains the important text input data. By accessing into the websites, retrieve all the text data in string format.

The different websites has different html code design, so the job advertisement websites have to be filter in order to reduce the heavy code during implementation.

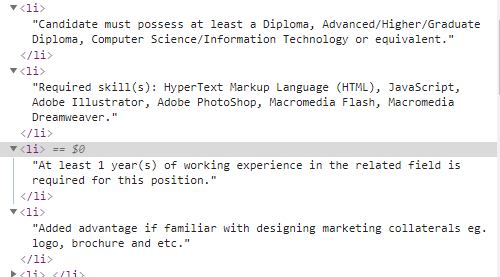


Figure 3: HTML Code of Job Advertisement

### Scope of the Project

The project scope are required to deliver 3 main modules as below:

* Web scarping: Automated scraping of job advert data from selected online job sites to the required depth and stores them in backend database.
* Machine Learning: K-means text clustering algorithm on the scraped data.
* Visualization: Visualizes the associations and relationships of keyword clusters.

## LITERATURE REVIEW

### Data Pre-processing Techniques

After the text data obtained, text normalization is an important step to get the better analytic result from text clustering. (Olga Davydova, 2018) The text normalization steps for the following text “The dog is sitting on the number 38 chair!” include:

#### Convert all letters to lower/ upper case

Converting all the string document into lower or upper case string document.

Output1: “the dog is sitting on the number 38 chair!”

Output2: “THE DOG IS SITTING ON THE NUMBER 38 CHAIR!”

#### Remove numbers

Remove the numbers from the string document.

Output1: “the dog is sitting on the number chair!”

#### Remove punctuations

Remove punctuation marks from the string document.

Output1: “the dog is sitting on the number chair”

#### Remove stop words

Remove the English words from the string documents which will affect the result score of the TF-IDF method.

Output1: “dog sitting number chair”

#### Stemming and Lemmatization

Stemming and lemmatization is a process of reducing words to their word stem, base or root form (for example, books — book, looked — look).

Output1: “dog sit number chair”

### Feature Extraction

TF-IDF stands for Term Frequency – Inverse Document Frequency. (Sammy Ongaya, n.d.) TF-IDF is a statistical method of extract the important words in given documents. Term frequency (TF) represents the number of frequency of a term appears in a document. Inverse document frequency (IDF) is the measurement of weight of word in document. It follows the concept that the terms appears frequently in a document are less important than terms that are rare.

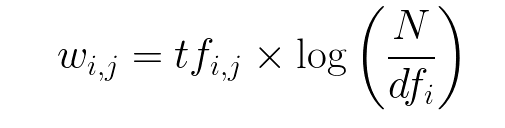


Figure 4: TF-IDF Score Calculation

According to Siddharth Yadav (n.d.), In information retrieval, tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. tf–idf can be successfully used for stop-words filtering in various subject fields, including text summarization and classification. Thus, the important terms in job description was able to be extracted by using TF-IDF vectorizer.

Besides, I[zabela A. Wowczko](https://www.researchgate.net/scientific-contributions/2085206578_Izabela_A_Wowczko) (2016) analyzed the important terms by its frequency of occurrence, it said that “I then visualized the most frequent words within the 7090 examples of the IT dataset. The top 30 words appearing in the job titles are presented in Figure 5 and reference job positions (manager, senior, lead, etc.), IT occupational areas (engineer, developer, analyst, administrator, security, operations, etc.), and technologies (software, application, java, data, net, web, sap, oracle).” The author was able to find out the significance of a term from the job descriptions. The terms found out was interpreted into a bar graph as follow:

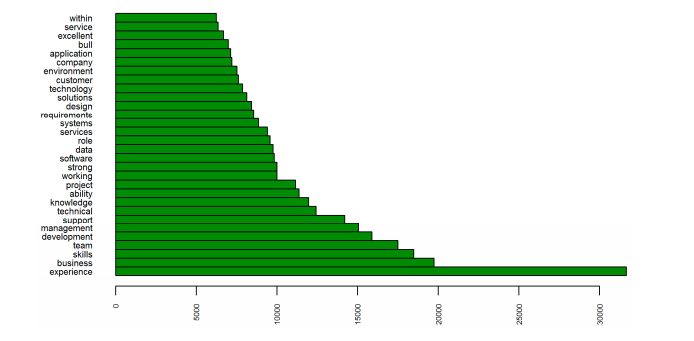


Figure 5: Frequency of occurrence for top 30 terms

In addition, according to Document Clustering with Python (n.d.), if the score of the term is greater than 80% of the documents, it is probably meaningless. The maximum term frequency is 80%, and term exceed 80% is filtered automatically. On the other hand, if the score of the term is lower than 20% of the documents, it is also probably meaningless. The minimum term frequency is 20%, and the term lower than 20% is filtered automatically. Therefore, the noises that affect the results of the project are decreased. For example, the word “salary” is appeared in all job descriptions, with this method, we are able to remove the word “salary” which may affect the result of this project.

### Machine Learning Algorithms

Figure 6 shows that machine learning is divided into 2 main techniques, supervised learning and unsupervised learning. Training a model on labelled input data to predict future output data known as supervised machine learning. Recognizing the hidden patterns on unlabelled input data known as unsupervised machine learning. Machine learning is the process of a computer to learn and build the ability of human. Supervised machine learning has 2 types of techniques, which are classification and regression.

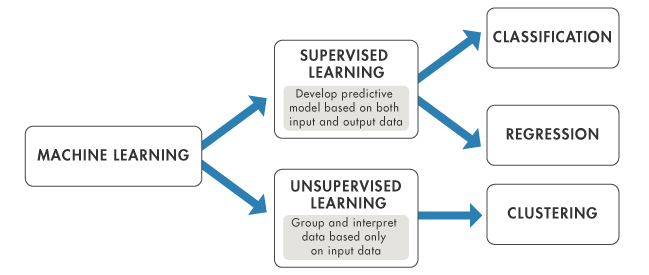


Figure 6: Machine Learning (“Machine Learning at MATLAB”, n.d.)

### Clustering Technique

“Clustering” is the process of grouping similar entities together. The goal of this unsupervised machine learning technique is to find similarities in the data point and group similar data points together. (Anuja Nagpal, 2017) The clusters shows the underlying pattern in different groups which we can identify the insight or hidden information in all these patterns.

#### Unsupervised Learning Algorithms

Two most common clustering algorithms are partitioning algorithm and hierarchical algorithm.

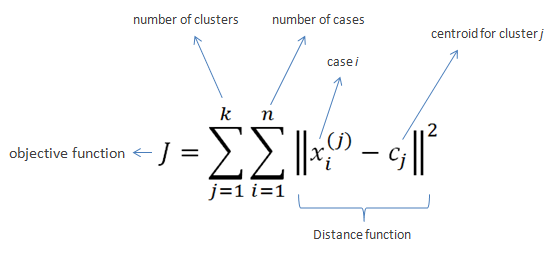
##### Partitioning algorithm

##### K-means clustering

K-means is the cost common partitioning algorithm that used in clustering. A cluster is represented by its centroid, the centroid is the mean of points for entities within the cluster.

Procedure of K-means referred in Swarndeep Saket J and Dr. Sharnil Pandya (2016):

1. The technique requires arbitrary selection of choose k objects from D as the initial centres, where k is the number of clusters and D is the data set containing n objects.
2. Repeat the first step
3. Reassign each object to the cluster to which object is most similar. It is based on the mean value of the objects in the cluster.
4. Calculate the mean value of the objects for each cluster. This had been concluded into the following formulae by Dr. Saed Sayad (n.d.):



1. Until no change

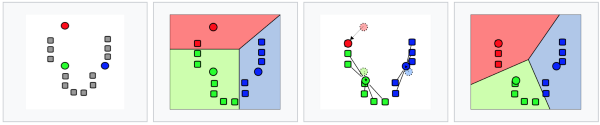


Figure 7: K-means clustering. Retrieved from Firdaouss Doukkali (2017)

##### K-Medoid

K-Medoid is not taking the mean value of the object in cluster as reference point. It use the medoid, the most centrally located point of a cluster.

Procedure of K-Medoid referred in Swarndeep Saket J and Dr. Sharnil Pandya (2016):

1. Arbitrary choose K points as the initial K-medoids, O (e.g. object in cluster).
2. Assign the objects into a set cluster where nearest to the medoid.
3. Randomly choose a non-medoid, Orandom object.
4. Recalculate the SSE, swap the medoid, O with Orandom if clustering quality is improved. Repeat this step until convergence criterion is satisfied.

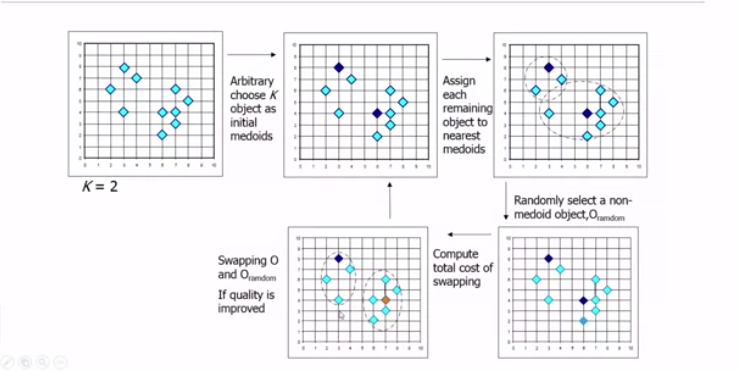


Figure 8: K-Medoid. Retrieved from Jiawei Han (n.d.)

Table 1: Comparing Clustering Techniques

|  |  |  |
| --- | --- | --- |
| Algorithms | Advantages | Disadvantages |
| K-Means | 1. Fast computation 2. It generates tighter clusters | 1. K value is hard to predict 2. Quality of cluster is hard to compare 3. Less work efficiency with global clusters |
| K-medoids | 1. Easy understand and implement 2. Fast computation 3. Least sensitive to outlier among Partition Around Medoid (PAM) | 1. Costly than K-Means because the time complexity 2. Weak scaling for large datasets 3. Total run time depends on first partitions |

##### Finding K-value for Clustering

The initial K-value should be assign to the k-means clustering algorithm, by plotting a line chart of the SSE for each K-value. If the line chart looks like an arm, the “elbow” of the arm will be best the K-value. K-value can be found demonstrated by Asanka Perera (2017) as follow:

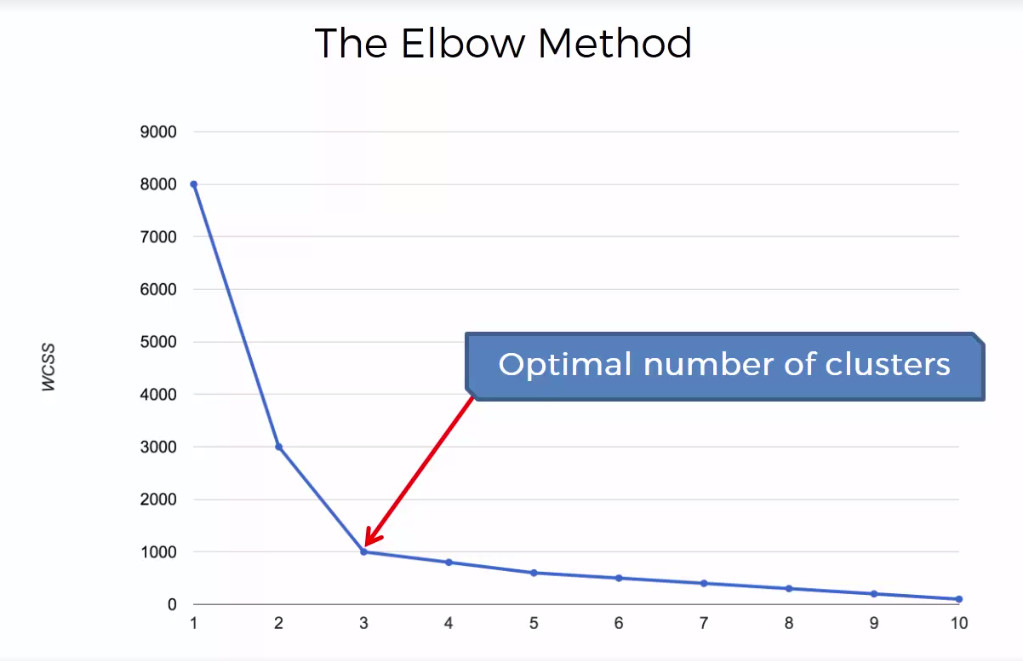


Figure 9: The elbow method

According to Siddharth Yadav, (n.d.), the project is using k-means clustering of 1 million headlines. In order to find out the number of clusters hidden in 1 million headlines, the project used elbow method in Figure 10. In Figure 10, due to the data complexity, the elbow method gave “elbow” to many clusters. So, it showcased the keywords of different amount of clusters to find out the right amount of clusters.

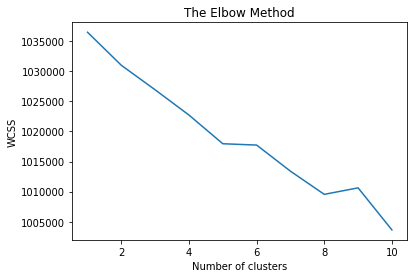


Figure 10: Elbow method for 1 million headlines

Therefore, the high complexity of data can stop the elbow from giving us the exact or correct number of cluster for the job descriptions. Clustering algorithm should be run for few times with different number of cluster to get the meaningful and good results. Term in each cluster is manually analysed to decide whether the term is useful to be the keyword of the clusters.

##### Hierarchical algorithm

There are two top-level methods for finding the hierarchical clusters (Doruk Kilitcioglu, 2016):

##### Agglomerative clustering

Agglomerative clustering follows “bottom-up” approach, it grouping up those similar nodes, then starts to continuously merge those dissimilar nodes together to form into same clusters.

##### Divisive clustering

Divisive clustering follows “top-down” approach, it starts with a macro cluster and split them up into possible micro clusters to build a dendrogram.

#### Text data mining

Text data mining uses the clustering technique, process unstructured text information, to extract the meaning full numeric indices from the text. By summarizing all the words, all words are indexed and counted in a table of documents, for example, matrix of frequencies of each words occurs in each document. Pre-processing should be done to exclude those common “English” word such as “the”, “a”, “with”, etc. In addition, the words should be combined which have only different grammatical formation such as “ring”, “ringing”, “rang” and “rung”. Finally, identify those important words that gives an insight to predict the outcome variable.

### Trending Skill Sets in Job Advert Websites

Text Mining (Big Data, Unstructured Data) (n.d.) states that “Investigating competitors by crawling their web sites.” Another type of potentially very useful application is to automatically process the contents of Web pages in a particular domain.” For instance, Up-to-date skill sets in market demands results in universities and colleges are eliminating those outdated skill sets from the subjects or courses in their education plan, in order to follow and providing the trending skill sets for their customers, students. Universities and colleges keep tracking the relevant trending skill sets that can maximize the profit by adding the up-to-date subject in their education plan and the benefits of the students.

### Visualizing techniques

After the job titles are clustered into clusters, we are able to visualize them in graph form and plot in word cloud. According to Izabela A. Wowczko (2016), in Figure 11 the project identified 7 clusters from job advertisements. The graph is plotted with cluster keyword against count of job clustered. From this, we are able to know developer is most welcomed job position in IT category.

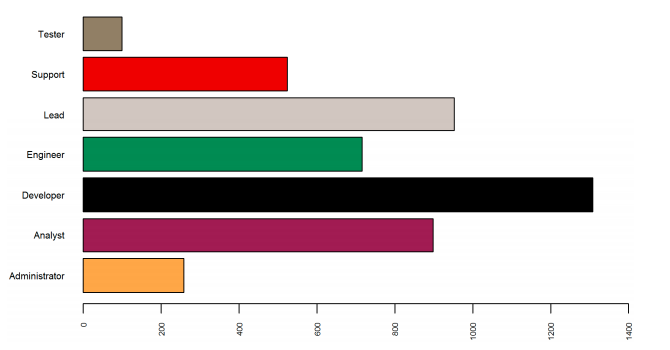


Figure 11: Labelled IT Dataset: Administrator (258), Analyst (898), Developer (1308), Engineer (716), Lead (952), Support (524), Tester (99).

Furthermore, according to Luis Cajachahua Espinoza, Andrea Ruiz Guerrero and Tomás Nieto Agudo (n.d.), the project extracted the keywords for each clusters and plot in word cloud form. In Figure 12, it show one of the clusters identified in the project. The clusters named as “Risk Managers”. Using the keywords appeared in Figure 12, we are able to assume that risk managers are professionals with experience in portfolio and risk management (both credits and investments), preferably analysts and engineers. They are sued for the financial and banking sector. They were also requested domain mainly SQL and SPSS.



Figure 12: Visualize and interpret result

## METHODOLOGY AND WORK PLAN

### Introduction

In this chapter, the methodology of the project, tools and techniques, and project activities along a timeline will be discussed. The methodology consist of the following steps:

#### Implementing web scraping module

These are the job advertisement websites that decided to be the source of text scraping:

* [www.indeed.com.my](http://www.indeed.com.my)
* [www.jobstreet.com.my](http://www.jobstreet.com.my)
* [www.monster.com.my](http://www.monster.com.my)

In the first phase, one of these job advertisement websites are selected to implement the very first prototype of web scraping module, each of the job advert on the same websites has the same format in html code. The aim of the very first module is focus on 1 specific job such as Software engineer, IT, Computer Science and the job related in this field and identify the trending skill sets of this field.

In the second phase, the scope for identifying the trending skill sets will be widen into every job if possible.

The unstructured text data is stored into a backend database in form of string for each text. The “English” words are filtered while inserting the text into backend database.

#### Implementing clustering module

This module is to implement a K-means text clustering algorithm, both clustering algorithm stated in **2.2 Clustering Technique** is implemented to check the clustering quality. The more techniques are tested to improve the clustering quality. The k-value will found by using elbow method stated in **2.2.1.2 Finding k-value for clustering**.

#### Implementing visualizing module

Visualizing module to visualize the patent of group in cluster for evaluation purpose in browser front-end.

The cluster is shown in a graph format with partitioning graph or dendrogram. Besides that, the module will also generates the reports of trending skills sets to understand easily without looking at the graph.

#### Combining the modules

Compile the modules in one shot to perform complete operation.

### Tools and techniques

The modules could be using cloud platform after the first phase implementation succeed.

1. Beautiful soup is a Python library for pulling data out of HTML and XML files. It works with your favourite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work. This library is used to scrap the html text data from job advertisement websites.

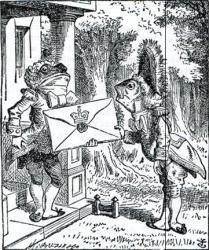


Figure 13: Beautiful Soup (Available at: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>)

1. MongoDB is a document database designed for ease of development and scaling. The Manual introduces key concepts in MongoDB, presents the query language, and provides operational and administrative considerations and procedures as well as a comprehensive reference section. MongoDB is used to store all the text data in document format which scraped from the job advertisement websites.



Figure 14: MongoDB (Available at: <https://docs.mongodb.com/manual/>)

1. Scikit-learn is providing the simple and efficient tools for data mining and data analysis. It is accessible to everybody because it is an open source. It has built on modules such as NumPy, SciPy, Matplotlib and etc. It is used to perform K-means text clustering algorithm, structure data, and plot graph.



Figure 15: Scikit-learn (Available at: <https://scikit-learn.org/stable/>)

1. D3.js is the JavaScript library for manipulating documents based on data. D3 allow us to bring data to life using HTML, SVG, and CSS. It is used to create the data visualizations in front end browser to interpret the results generated by text clustering algorithm.



Figure 16: D3.js (Available at: <https://d3js.org/>)

## PROJECT INITIAL SPECIFICATION

### Introduction

This chapter describes about the functional, non-functional requirements for each deliverable module.

### Functional requirements

* Web scraping module
  + 1. The module is able to scrap the text from selected html of job advertisement sites with automated manner.
    2. The module is able to parse the html codes into texts and filter out the unimportant “English words” and symbols.
    3. The module is able to search and filter the same job advert that appeared on different job advert sites (if any).
    4. The module is able to store the texts into a backend database.
* K-means text clustering algorithm module
  + 1. The module is able to retrieve all the unstructured text data.
    2. The module is able to identify the total number of cluster to compute.
    3. The module is able to compute the text data to perform clustering technique.
* Visualizing module
  + 1. The module is able to show the results generated by K-means text clustering model with dendrogram or partitioning graph.
    2. The module is able to provide the analytical table to understand the percentage of each trending skill sets for each job without looking at the graph.
    3. The module is able to run as an user interface by clicking button to compile the system.

### Non-functional requirements

* The system is able to finish the compilation of all module within 2 hours.
* The modules are able to sync to cloud once all the module runs smoothly.
* The tools and techniques are able to use in cloud platform to increase the flexibility, maintainability, and reusability that not only run on “my computer”.

### Data flow diagram (DFD)

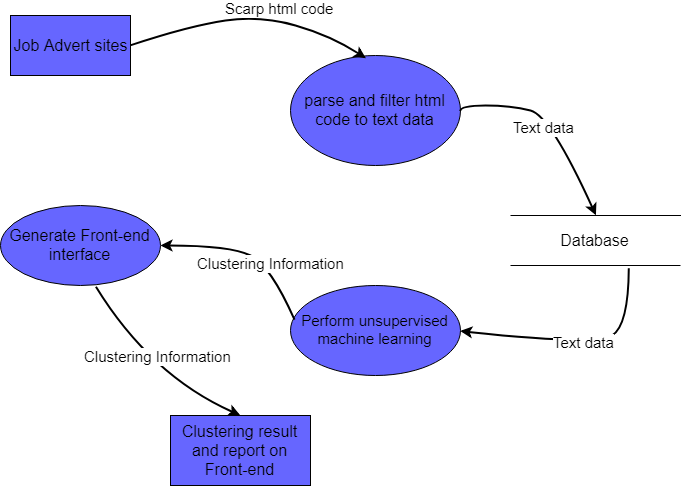


Figure 17: Project data flow diagram

### Project Initiation

#### Investigation for several job advertisement sites

Based on the research Online & Dolnicar (2002), the minimal sample size to include no less than 2k cases (k = number of variables/ clusters), preferably 5\*2k. In jobstreet.com.my has about **14** main job specialization categories (Accounting/ Finance, Admin/ Human Resources, Arts/ Media/ Communications, Building/ Construction, Computer/ IT, Education/ Training, Engineering, Healthcare, Hotel/ Restaurant, Manufacturing, Sales and Marketing, Science, Service and Others), which represents the number of clusters that need to be investigated. With this, we can easily predict the minimum size of documents to perform text clustering. **5\*214 = 16384**, sample size that we need to collect. It has total of **30,975 job adverts** in this website.

In another job advert site, monster.com.my, about **10** main job functions that has been found, which are Human Resources, Construction, Customer Service, Finance& Accounts, Manufacturing/ Engineering, Banking/ Insurance/ Financial Services, Sales/ Business Development, Marketing& Communications, IT, and Others. **5\*210 = 5120** with **1933** job adverts. In Indeed.com.my, there are **14** types of job titles, which need **16384** sample size and the website contains **42,085 job advert.**

As a conclusion, the sample size is predictably enough to perform text clustering with these 3 job advertisement sites. When the project has programming developed, we can use the elbow method stated in **2.2.1.2 Finding K-value for clustering** to find the accurate K-value to calculate sample size. The sample size may be increased by exploring another websites to fit the optimal sample size.

#### Prototyping

A simple prototype has been done by:

##### Scraping the text data from one of the job advert sites

The following python codes shows how it access to the jobstreet.com.my and scrap the text data from the job description for each job title. Then the text data are stored into a text file.

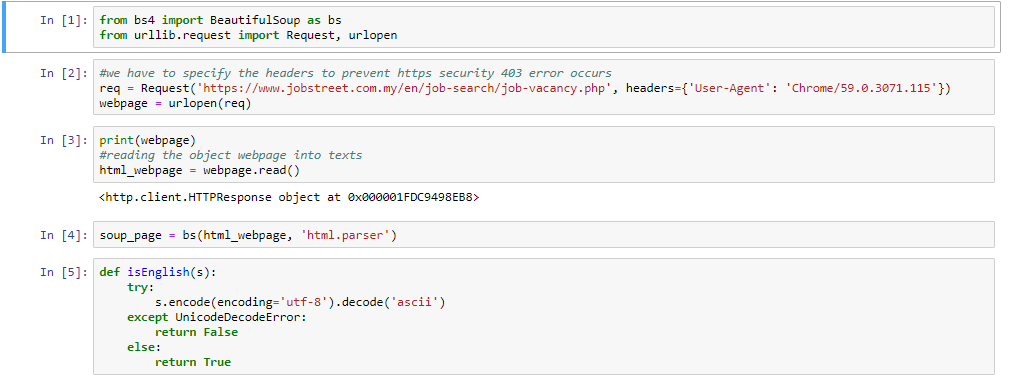


Figure 18: Web scraping prototype

##### Read the text documents to perform basic k-means clustering algorithm

The following python codes shows how it transform text documents into numerical data using TF-IDF vectorizer. TF-IDF vectorizer are able to measures the weight score of an important words in all job advertisement document. Then we will perform K-means clustering algorithm with the clusters defined with top terms per clusters as shown below. The job descriptions are able to be clustered into clusters defined using TF-IDF vectorizer.

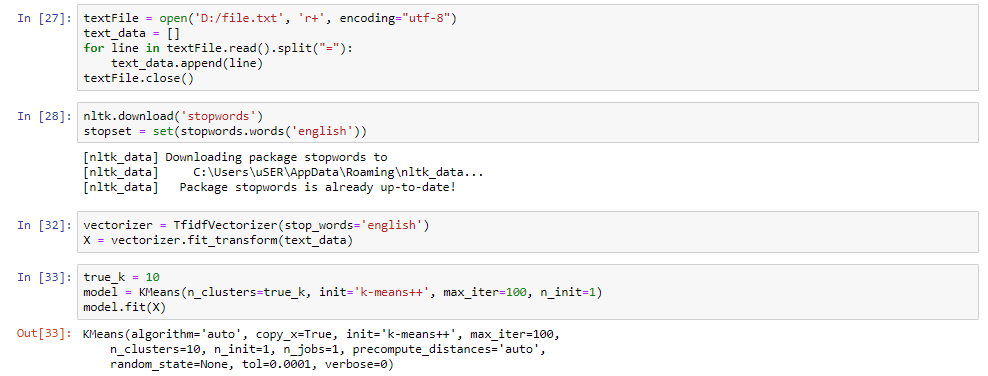


Figure 19: Text clustering prototype

## PROJECT IMPLEMENTATION

### Introduction

This chapter discussed about the processes to implement the whole project. This chapter also discussed about how the results and outcomes being analysed.

### Code implementation

The project has been implemented with 3 modules which are Web Scraping Module, Data Processing Module, and Data Visualization Module. The project is built up mainly using Jupyter Notebook in Python language.

#### Web Scraping Module

##### Html Structure

This module basically is used for scrap all the plain text data from the websites which stated earlier. This module is implemented in a quite manual way because the html structure for each website is different. The module is unable to scrap the plain text data from all the websites using the same code structure. Thus, this module is limited scrap the website that it specified. As the websites proposed earlier, “Jobstreet.com.my”, “Indeed.com”, and “Monster.com”, these websites are in totally different HTML structure.

For example:

**Indeed**

<div class="jobsearch-JobComponent-description icl-u-xs-mt--md">

**Monster**

<div data-v-4fa62c2c="" class="card-panel job-description-content">

**Jobstreet**

<div class="unselectable wrap-text" id="job\_description"> <div>

The job descriptions text are located in these division which are from different job advertisement websites. As we can see, they are using totally different class name and id name to represent the job description. Moreover, in the division there are still some element wrapping job description text. The module need to go deep into each different division to scrap the complete text data. Thus, there are total of 3 web scraping module which performing the same functionality but scrap data in the way that the html structure being.

##### Security

The website security is also a huge problem for me to scrap the sensitive data from the websites. For example, Figure 20 shows that we are unable to view the salary if we are not logged into the website. It is a difficulty to make it automated where the python code can login the website by itself. We are not able to predict what is the next thing will come out from the screen after logged in. The websites keep changing its pattern to prevent hackers.

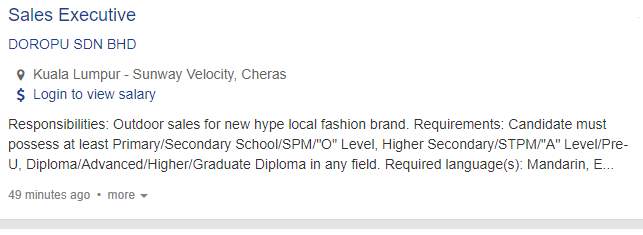


Figure 20: Login to View Salary

#### Data Storage

In the process of scraping the text data for each job advertisement, the Web Scraping Module will also store the data into NoSQL database, MongoDB. Since there is not restriction on the pattern of data storage. As general, we store the data as BSON job advertisement object which contains “title”, “desc”, “salary”(if exists), “location”, “data\_rec”, and “source”.

**title** – the title of the job advertisement

**desc** – the description of the job advertisement

**salary** – the salary offered by the job advertisement

**location** – the location of the job

**data\_rec** – The date of the job advertisement scraped

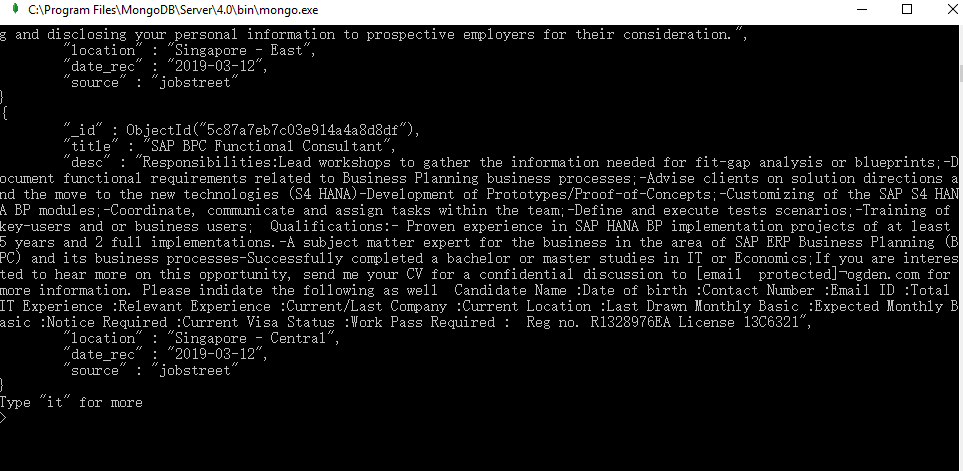


Figure 21: BSON Object for Job Advertisement

For a better visualization for all these data, I was suggested to utilize MongoDB Compass, which is a GUI for MongoDB. With this, I am able to visualize the data scraped easily.

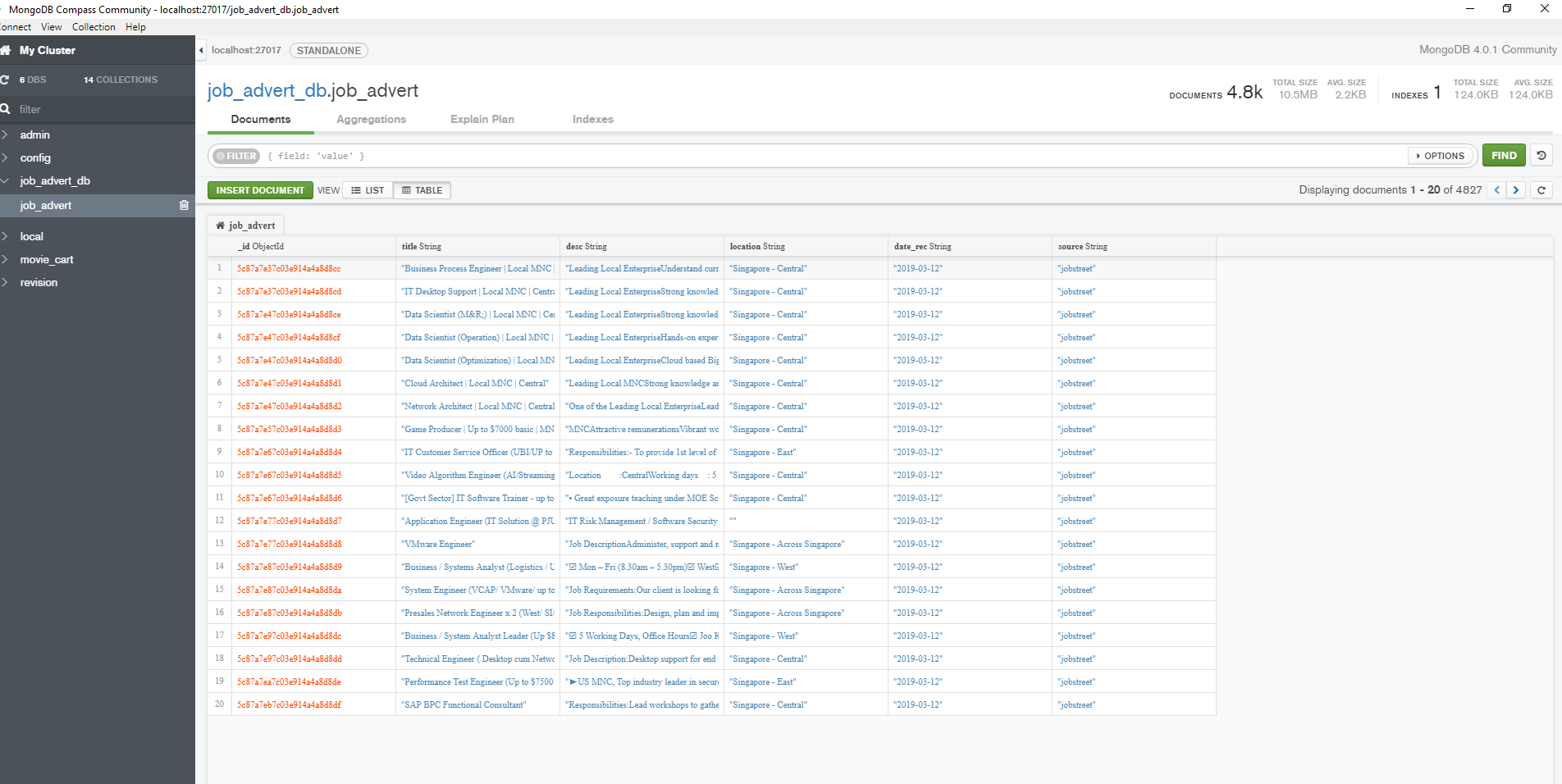


Figure 22: MongoDB Compass

#### Data Processing Module

This module is performing data pre-processing methods, and job advert clustering method. After the data is stored in MongoDB server, implement the data processing module to pre-process the data in order to filter the unnecessary words that will affect the clustering algorithm to get the proper insights from the text data. The techniques that used to pre-process the data are:

##### Data Pre-processing Techniques

###### Noise Removal

When scraped the html text data from job advertisement websites, there should have many noise stick with the text data. Noise removal function are provided by Beautiful Soup framework itself. The noise that need to be removed are text file headers, footers, html, XML, mark-up and metadata, and also extracting the valuable data from JSON format when needed.

###### Tokenization

Tokenization is a process that splits the strings of text into short text pieces, also known as tokens. Long string texts can be split into sentences and sentences can be split into words.

###### Remove Punctuation

Removing punctuations from the sentences and texts, the system are not able to differentiate the string of text and string of punctuation, so we will use remove punctuation method to remove the ‘,’, ‘.’, ‘!’, etc. from the tokenized data. The punctuations are useless for us to perform the clustering algorithm.

###### To Lower Case

Converts all the words into lower case to prevent case sensitive functions that will affect the final result of the project.

###### Filter English Words

The ‘English’ Words such as ‘you’, ‘more’, ‘then’, ‘only’, ‘at’, and so on which are not important in text clustering will be filtered out from the data. We can also add the custom words into the set of stop words which we would not like to see.

###### Stemming and Lemmatization

Stemming are able to stem the words which have the same root meaning, for example, ‘fished’, ‘fishing’ and ‘fishers’ will be stemmed into ‘fish’ keywords. Which reduced the different keywords appears from the text data. This will helps TF-IDF vectorizer to recognize the important keywords features from the text data.

##### TF-IDF Vectorizer

After the text data is pre-processed completely, the module will then proceed to finding keyword features from text data using TF-IDF method. TF-IDF are able to extract the features from the job descriptions. Basically, the pre-processed job description will be the input of TF-IDF vectorizer which transform job description into TF-IDF matrix which contains the scores of each keywords in the matrix.

##### Finding K-Value with Elbow Method

The TF-IDF matrix will then be the input of K-Means algorithm to produce a global optimal number of clusters named The Elbow Method. The elbow method graph is created with y-axis (within cluster sum of squares) against x-axis (number of clusters). By using K-Means, we find sum of squared distances of samples to their closest cluster centre which known as inertia. If the within cluster sum of squares value remains at a number for few next clusters, the elbow is formed.

In Figure 23, although there is no an obvious elbow shown in the graph, we can slightly predict the number of clusters, K-Value is 8. Since there is a slight elbow appears. Finally, we decided to use K-Value as 8 to get clusters result.

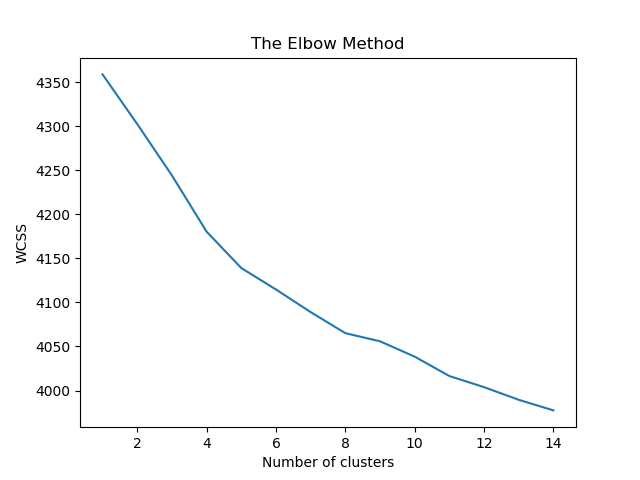


Figure 23: Elbow method (K=8)

After we have got the number of clusters, we are able to see the job advertisement titles are already clustered into the keyword clusters as shown in Figure 24.

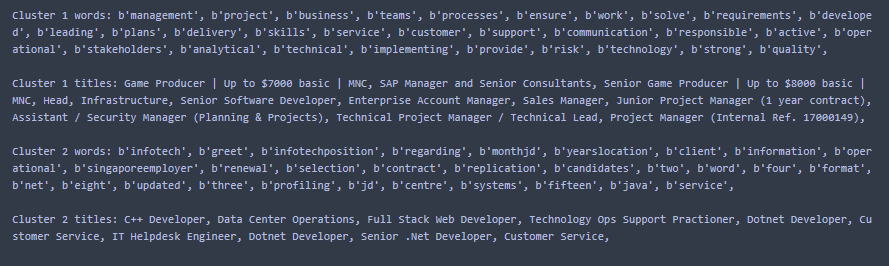


Figure 24: Clusters keyword and job titles clustered

### Result and discussion

The total number of job advertisements that collected from websites is 4827. All the job advertisements are IT computer science related fields. However, the data size is still quite big which we can see the clusters number are not clear when we use Elbow Method. The reason that we used K-Means algorithm is because the example code of K-Means algorithm is much more than the others algorithm, it is very easy to be referred from online source code. After the several runs for K-Means algorithm, we finally decided to use K=8 as our global optimal value for clustering. So, there are total 8 clusters for us to analyse and cluster all the job descriptions:

Table 2: Clusters with keyword and job title clustered

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Keywords | Job title Clustered | Job title counts |
| 0 | networking, support, hardware, system, service, troubleshooting, provide, security, administration, configuration | IT desktop Support, Cloud Architect, IT Customer Service Officer, and etc. | 719 |
| 1 | management, project, business, teams, processes, ensure, work, solve, requirements, developed | Senior Game Producer, Sales Manager, Junior Project Manager, Technical Project Manager and etc. | 771 |
| 2 | infotech, greet, infotechposition, regarding, monthjd, yearslocation, client, information, operational, singaporeemployer | c++ Developer, Dotnet Developer, Senior .Net Developer, Dotnet Developer, Full Stack Web Developer and etc. | 129 |
| 3 | teams, work, productionize, business, growth, markets, developed, customer, management, build | Data Scientist, DevOps engineer, Customer Account Manager, Asset Management IT, Product Designer, System Analyst, IT Project Manager and etc. | 767 |
| 4 | testing, developed, requirements, systems, design, analytical, software, project, documentation, work, solve, business, specifications | Data scientist, Application Engineer, Software engineer- C#, (SA) Software engineer – Java and etc. | 696 |
| 5 | security, cyber, consulting, discriminate, feel, cooperative, raising, client, min, threat | Communication Solution Architect (Presales), Regional Marketing Manager, APAC, Sales Engineering Manager, Senior Sales Manager, Security consultant, Security lead, Cyber security Specialist and etc. | 90 |
| 6 | developed, web, design, Apicentric, java, JavaScript, code, frameworks, testing, work, knowledge, software, programming, skills, net, databases, strong, css, sql | Web/ Front end Developer, PHP Developer, Software Engineer, Web Application Developer (.NET) Central, Senior .Net Developer, Software Engineer – Web (JavaScript/ React) and etc. | 700 |
| 7 | work, skills, requirements, management, developed, good, communication, knowledge, data, project, software, support, business, analytical, design, teams, client, customer | Business Process Engineer, IT software Trainer, Business/ System Analyst, System Analyst Leader, Performance Test Engineer, SAP BPC Functional Consultant and etc. | 947 |

Referring to Table 2 shown above, each clusters have their own clustered keywords which can be used to cluster the job titles by using the job descriptions. By using the keywords for each cluster, the skills are able to be predicted as follow:

* Cluster 0: Technical support skills (719 jobs)
* Cluster 1: Business project management skills (771 jobs)
* Cluster 2: It is not a skill set, but it is categorized as job demand in Singapore (129 jobs)
* Cluster 3: Business project management skills (767 jobs)
* Cluster 4: Software Testing skills (696 jobs)
* Cluster 5: Cyber security consultation skills (90 jobs)
* Cluster 6: Web application development skills (700 jobs)
* Cluster 7: Business project management skills (947 jobs)

Cluster 1, 3, 7 are categorized into business project management skills because they have the similar keywords in the cluster. Figure 25 shows that the number of job adverts located in each clusters.

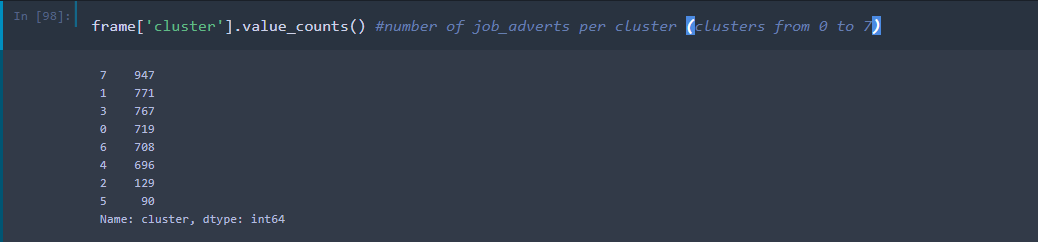


Figure 25: Clusters Job Adverts Count

For a better visualization for the result, a bar graph is created as shown in Figure 26.

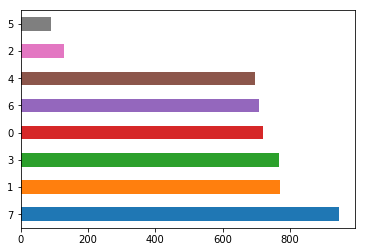


Figure 26: Clusters against number of job titles clustered

In short, the trending skill sets in area of Singapore and Malaysia in 2019 are identified which includes 2485 jobs are required business project management skills, 719 jobs are required technical support skills, 708 jobs are required web development skills, 696 jobs are required software testing skills, and 90 jobs are required cyber security consultation skills. Thus, these skill sets are high demand in the “IT Computer Science” and Singapore and Malaysia area. As we can see, business project management skills has the highest demand in the market which has 2485 out of 4827 jobs required this skill set.

## CONCLUSION

In this report, chapter 1 discussed about the problem statement that motivates me to implement the project. The problem statements are universities or colleges provide the outdate courses in education plan, market research is expensive, and job seekers do not know which expert skill should be acquired. With such, the key objective of this project is to use K-means text clustering algorithm to identify the trending skill sets need to be acquired from job advertisement websites. The project solution discussed about how the objective was achieved. Three modules were created which are web scraping module, data pre-processing and K-means clustering module, and visualizing module. Web scraping module scrap the text data from html structure of job advertisement websites. Data pre-processing and K-means clustering module cleans the data and perform K-means text clustering algorithm to find the relationship between each cluster. Visualizing module shows the relationship between each cluster in graph form. Besides, Chapter 2 discussed about the tools and techniques used in this project and benchmarking other similar project. Moreover, Chapter 3 discussed about the methodology and work plan for this project. All the tools and techniques discussed were used to implement the modules. Beautiful Soup in python and MongoDB were used in web scraping module. Scikit-Learn was used in data pre-processing and K-means clustering module. D3.js was used to visualize the clustered data which allow us to interpret the result easily. In addition, chapter 4 mainly discussed about the specification of the project and the prototype created. It discussed the functional and non-functional requirements, data flow diagram allows us to visualize the data flow in each module, and the simple prototype created at initializing phase. Furthermore, chapter 5 discussed about the processes to implement the whole project. This chapter also discussed about how the results and outcomes being analysed. Lastly, chapter 6 discussed about the conclusion of the report, limitation of the project, and the work that can be done in future.

### Project Limitation

#### Request Clean and Structured Data

Data complexity is the major problem that will affect the accuracy of the result. The html data scarped from job advertisement websites is very complex. Sometimes Beautiful Soup is not able to scrap the specific data attribute from every job advertisement. For example, the job salary is not shown in job description and the salary attribute will be empty. I failed to receive any reply from the admin of the job advertisement websites to get the clean and structured data.

### Future Development

#### Data Complexity

The data complexity scraped from html job advertisement sites is very high. For future development, we can try to request the structured data from the official job advertisement websites. This can remove unnecessary noise from the html plain text data. Besides, the project can be scale bigger to analyse more topics such as the job salary range offered by different location.

#### Job Title Search Application

The data visualization is not enough for us to visualize all the data with total number of 4827. The search application allow people to key in skill set keywords and the application is able to show all the job titles related with the skill sets. Although job advertisement sites are able to search for skill keywords, but they do not rank the important keywords which this project are using TF-IDF vectorizer to measure the importance of a keyword. The job title will be listed accordingly with the frequency of keywords appear in that particular job title.

#### Utilize Other Data Visualization Techniques

The project is mainly using Python language to develop. According to the researches made earlier, mostly data visualization techniques are provided in R language. There is quite less example code and implementation for us to refer using Python language.

## REFERENCES

Robert Half. 2018. 10 Best Job Search Websites. [ONLINE] Available at: <<https://www.roberthalf.com/blog/job-market/10-best-job-search-websites>>. [Accessed 13 July 2018].

Izabela A. Wowczko, 2016. Skills and Vacancy Analysis with Data Mining Techniques. *Institute of Technology Blanchardstown*, [e-journal]. Available through: ResearchGate website <<https://www.researchgate.net/publication/284217689_Skills_and_Vacancy_Analysis_with_Data_Mining_Techniques>> [Accessed 26 July 2018].

Gary Burtless. 2014. Unemployment and the “Skills Mismatch”. Story: Overblown and unpersuasive. [ONLINE] Available at: <<https://www.brookings.edu/opinions/unemployment-and-the-skills-mismatch-story-overblown-and-unpersuasive/>>. [Accessed 14 July 2018].

Machine Learning at MATLAB, n.d.. [Online] Available at: <<https://www.mathworks.com/help/stats/machine-learning-in-matlab.html>>.

Olga Davydova, 2018. Text Preprocessing in Python: Steps, Tools, and Examples. [Online] Available at: <<https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908>>. [Accessed 2 April 2019].

Sammy Ongaya, n.d.. Feature Extraction with TF-IDF. [Online] Available at: <https://galaxydatatech.com/2018/11/19/feature-extraction-with-tf-idf/>.

Siddharth Yadav, n.d.. K-means Clustering of 1 million headlines. [Online] Available at: <<https://www.kaggle.com/thebrownviking20/k-means-clustering-of-1-million-headlines#TF-IDF>>. [Accessed 2 April 2019]

Document Clustering with Python, n.d.. [Online] Available at: <<http://brandonrose.org/clustering#Tf-idf-and-document-similarity>>. [Accessed 2 April 2019]

Luis Cajachahua Espinoza, Andrea Ruiz Guerrero and Tomás Nieto Agudo, n.d.. Specializations for the Peruvian Professional in Statistics: A Text Mining Approach. [e-journal]. Available through: CEUR Workshop Proceedings <<http://ceur-ws.org/Vol-1478/paper2.pdf>>. [Accessed 2 April 2019]

Swarndeep Saket J and Dr. Sharnil Pandya, 2016. An Overview of Partitioning Algorithms in Clustering Techniques. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, [e-journal] 5(6), pp. 1943-1946. Available through: Ijaret website <<http://ijarcet.org/wp-content/uploads/IJARCET-VOL-5-ISSUE-6-1943-1946.pdf>>. [Accessed 26 July 2018].

Firdaouss Doukkali, 2017. Clustering Using K-Means algorithm. [ONLINE] Available at: <<https://towardsdatascience.com/clustering-using-k-means-algorithm-81da00f156f6>>. [Accessed 26 July 2018]

Jiawei Han, n.d.. The K-Medoids Clustering Method. [ONLINE VIDEO] Available at: <<https://www.coursera.org/lecture/cluster-analysis/3-4-the-k-medoids-clustering-method-nJ0Sb>>. [Accessed 26 July 2018]

Dr. Saed Sayad, n.d.. K-means clustering. [ONLINE] Available at: < <http://www.saedsayad.com/clustering_kmeans.htm>>. [Accessed 26 July 2018]

Text Mining (Big Data, Unstructured Data), n.d.. [ONLINE] Available at:< <http://www.statsoft.com/Textbook/Text-Mining>>. [Accessed 27 July 2018]

Rob Petersen, 2016. 20 companies do data mining and make their business better. [ONLINE] Available at: < <http://barnraisersllc.com/2016/11/companies-data-mining-business-better/>>. [Accessed 27 July 2018]

Asanka Perera, 2017. Finding the optimal number of clusters for K-Means through Elbow method using a mathematical approach compared to graphical approach. Available at: <<https://www.linkedin.com/pulse/finding-optimal-number-clusters-k-means-through-elbow-asanka-perera>>. [Accessed 27 July 2018]

Doruk Kilitcioglu, 2016. Hierarchical Clustering and its Applications. Available at: <<https://towardsdatascience.com/hierarchical-clustering-and-its-applications-41c1ad4441a6>>. [Accessed 27 July 2018]

Online, R., & Dolnicar, S. (2002). *A Review of Unquestioned Standards in Using Cluster Analysis for Data-Driven Market Segmentation.* Retrieved from <http://ro.uow.edu.au/commpapers/273>