CONTEXT AWARE APP RECOMMENDATION USING EMAIL SEMANTIC
ANALYSIS
BY
TANG ZI CONG

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION SYSTEM (HONS)

INFORMATION SYSTEM ENGINEERING

Faculty of Information and Communication Technology

(Kampar Campus)

JAN 2019

REPORT STATUS DECLARATION FORM

Title: CONTEXT AWARE APP RECOMMENDATION USING EMAIL
      SEMANTIC ANALYSIS

Academic Session: JAN 2019

I   TANG ZI CONG
        (CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in Universiti Tunku Abdul

Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

                  Verified by,

_____   _____

(Author's signature)   (Supervisor's signature)

Address:
25, Laluan Chemor Maju 10,
Taman Chemor Idaman,   DR. AUN YICHIET
31200, Chemor, Perak.   Supervisor's name

 Date: 08/04/2019   **Date**: 08/04/2019

CONTEXT AWARE APP RECOMMENDATION USING EMAIL

SEMANTIC ANALYSIS

BY

TANG ZI CONG

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION SYSTEM (HONS)

INFORMATION SYSTEM ENGINEERING

Faculty of Information and Communication Technology

(Kampar Campus)

JAN 2019

DECLARATION OF ORIGINALITY

I declare that this report entitled "**CONTEXT AWARE APP RECOMMENDATION USING SEMANTIC ANALYSIS**" is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature   :   _____

Name   :   TANG ZI CONG

Date   :   08/04/2019

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

ACKNOWLEDGEMENTS

With great pleasure, I would like to send my utmost gratitude to my Final Year Project's Supervisor, Dr. Aun Yichiet, who has selflessly channelled his valuable knowledge, experience and guidance to me throughout this research. This dissertation would had made impossible without his invaluable guidance.

Last but not least, I would like to express appreciation to my family and friends who had been there supporting me all the way.

ABSTRACT

Email is a kind of electronic tool that people used to exchange message. There are many functionalities such as spam filtering, categories section, malicious email identification and etc are available in the email application that had been launched in the existing market. In order to increase the functionalities for the email application and further enhance the user experience, it is proposed to develop a recommendation system to work with the email application that provides a recommendation based on the email content. This means that people can easily receive some recommendation for their next action from the email application based on the email they received. The recommendation system is planned to develop by using data mining and machine learning. The email content as the raw data and sent to the test mining to perform text classification and category the data into different categories. Then the data will go through the machine learning algorithm to perform predict and provide a recommendation that suitable for the content.

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

TABLE OF CONTENTS

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

LIST OF FIGURES

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

LIST OF TABLES

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

LIST OF ABBREVIATIONS

MIT   Massachusetts Institute of Technology

TF   Term Frequency

IDF   Inverse Document Frequency

HMM   Hidden Markov Model

UML   Unified Modeling Language

IDE   Integrated Development Environment

RAM   Random Access Memory

JDK   Java Development Kit

API   Application Programming Interface

NLP   Natural Language Processing

ARR   Average Rate of Return

UTAR   Universiti Tunku Abdul Rahman

Chapter 1: Introduction

1.1 Problem Statement and Motivation

The recommendation system is an information filtering system that looks to forecast the "rating" or "preferences" a user would provide for a thing. In simple, the recommendation system is a system that provides a list of information on items that can help the user to decide something or to make a conclusion on it based on user preferences. In the real world, there are many applications that integrate with the recommendation system to provide a recommendation in order to enhance user experience in using their application. Those apps are Netflix, Amazon, Lazada, YouTube and etc. That app normally using the video watching history or the search history to provide a recommendation that similarity to the user's history.

In an email, the recommendation system had been used to suggest text auto compose in Gmail. But this feature only can be used on the web, not in application level. In order to use the Smart Compose feature, the user had to complete their email via the Gmail website. This project is using AI to understand the semantic and context to suggest the right app to use based on the email. By provide a recommendation to the user, it can help them to solve their action that required for the task in the email.

1.2 Project Scope

This project is aimed to enhance the user experience of an email system by developing a module on a recommendation based on the email context resolving. This project involves text mining and machine learning. The overcome of this project is to provide a recommendation based on the user email context. It will be using text mining to perform text classification and also deep learning to perform accurate recommendation for the user. There are few limitations founded in this project. Because of the limited data and time constraint to complete this project, the result accuracy of the recommendation system that is created may compromise. Below are the limitations of the proposed natural speech reconstruction system:

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

1.2.1 Noise

The expression "noise" has negative undertones in scientific contexts. Noise is mutilation in information, that is undesirable by the perceiver of information. Noise is anything that is misleading and incidental to the first information, that isn't expected to be available in any case, however, was acquainted due with defective catching procedure. if train too much noise data into the machine learning algorithm it might affect the final result.

1.2.2 Writing Style

Each user has his/her own style in writing the message, some of them might use their own style / own create a short form of words. It will be bringing some trouble for the text mining technique. It is because the text mining technique cannot recognize all type of short form created by the millions of users.

1.3 Project Objective

The ultimate goal to conduct this research is to propose a module to enhance the user's experience of using the email system.

The objectives of the research are as followed:

1. To provide useful recommendation information for the user as in suggesting the right Apps to use.
2. To develop a text mining technique to understand the email semantic.
3. To design an app dictionary to profile applied functionalities-pervasively.
4. To design a recommendation system to suggest apps based on email semantic using categories and verb frequency analysis.
5. To optimize the model accuracy using keyword manual weight and hyperparameter tuning.

1.4 Impact, Significance, and Contribution

Despite the blooming of mobile phone communication applications, email system still serves as one of the major communication tools at both workplace and personal social life. People spend couples of time on reading, analyzing and interpreting email text message, followed by making a decision based on the content. Moreover, when a user needs to find an external solution based on the email message, he/she need to switch between different applications. More and more time is needed when users have no idea of getting the solution, and this will affect the user experience. In this case, a module which can provide recommendations based on the email content can certainly solve this issue via a technique such as text mining and machine language.

1.5 Background Information

The first email sample was found on the computer located in MIT in 1965. But, the first email message which can send between computer and computer was stated in 1969. With technology advancement nowadays, email has become an imperative part of daily business activities in nearly all aspects of commerce. the huge breadth of populations exploitation email often may be attributed, in massive half, to its accessibility and general quality.

Email system not only uses for the daily business activities but contribute to personal usage. There are more and more email application providers in the market, such as Gmail, Outlook, Yahoo, iCloud and etc. Using the email system as the communication tool, it can decrease the waiting time to receiving the mail letter. Over the years, the developer has been working on increasing the features which are allocated in the email system. The email system nowadays in the existing market which consists of many functionalities compare with the first version of the email system which only able to send and receive email.
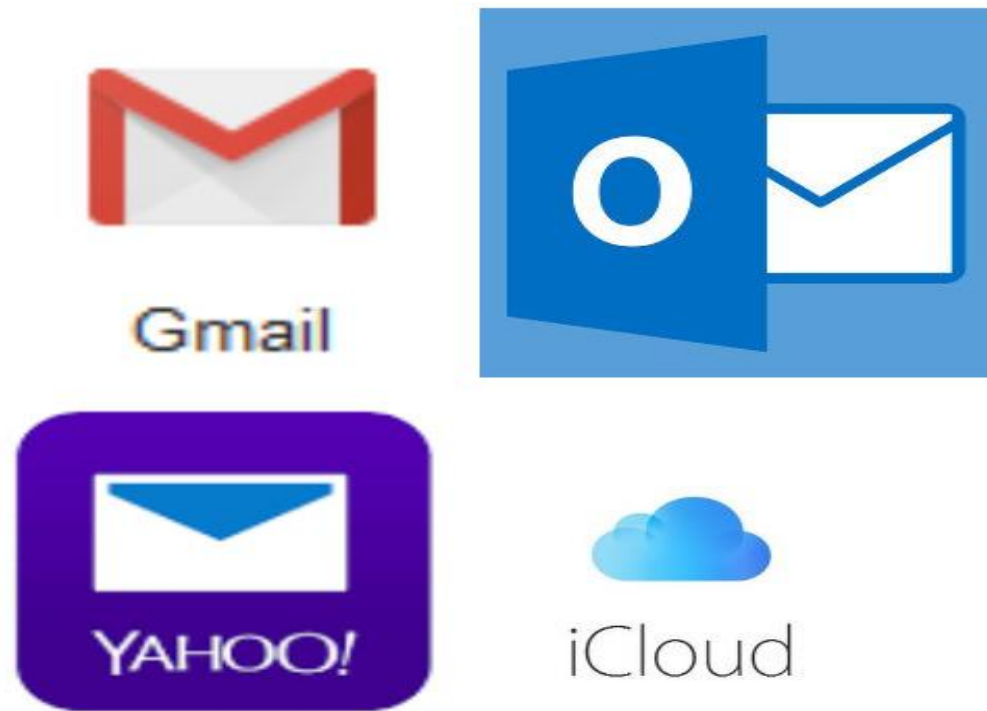
Figure 1.0 Logo of Email Application Providers

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

Chapter 2: Literature Review

2.1 Introduction

This chapter presents a review of the concept and the related work of text classification. as well as the concept of text mining and machine learning and the similar application that using text mining techniques and machine learning algorithm.

2.2 Email System Usage and Statistics

With the technology advancement nowadays, people tend to go for simpler, more convenient options for communication, leading to addiction towards social media applications such as WhatsApp, Telegram, Messenger, etc (Andersson, 2018). Still, the email system remains as one of the major communication tools over the years. Based on the report from Statista released in 2018, the number of email-users worldwide for the year 2017 is about 3718 million users, which is more than 100 times of Malaysia's current population that only recorded at 31.63 million (The office of Chief Statistician Malaysia, 2018). As depicted in the figure below, the number of email-users increases from year to year.
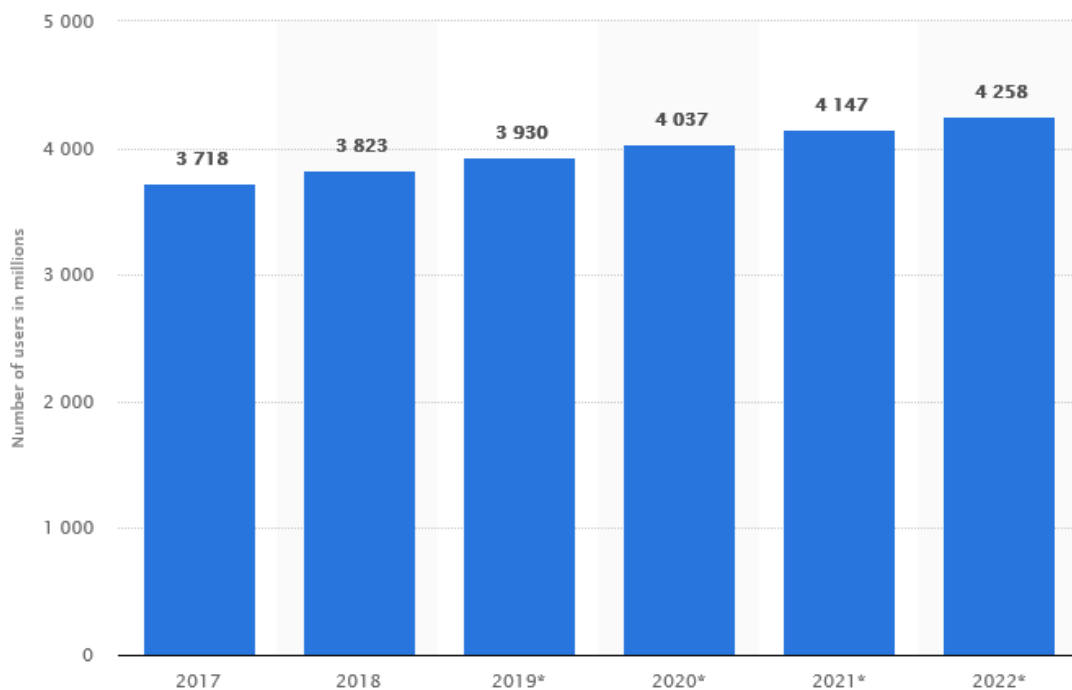
Figure 2.0 Number of e-mail users worldwide from 2017-2022 (in millions) (Statista,2018)

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

Over the years, developers have been working on increasing new features and improving existing features in the email system, from the original email system that only provides send and receive information feature to the advanced current email system with malicious email identification, spam email identification, auto-reply features, and etc.

There are plenty of email service providers including Gmail, Outlook, Yahoo, iCloud, and etc. Besides the basic features, every email system provider has its own features. For example, Gmail had its own features like personalized information filter. Besides, Gmail will automatically filter the email that receives by the user into five categories which are primary, social, promotions, updates, forums to aid users in organizing the incoming emails (Gilad, 2013). One of the popular email providers, which is Outlook (previously known as Hotmail), Microsoft provides that the outlook can work with the calendar changes. For example, Microsoft will add the reminder for the bill, travel reservation, travel itinerary into the calendar that has been appearing in the outlook.com. Besides, Microsoft also bind the advanced camera features into the Outlook.com. The advanced camera is known as Office Lens, it can capture the information that located at the whiteboard or take photos of documents and it will automatically crop the photo into a perfect size like scan it from a printer or scanner (Warren, 2018).

2.3 Contextual Email Content Resolving

Contextual email content resolving is deemed as an act of extracting, organizing and analyzing related word-based information in the email by using certain techniques to enhance the user's experience (Skulimowski & Kacprzyk, 2016). In simpler words, the system needs to perform text classification and understand the email's contents to aid the user's further actions. Text classification refers to separate the contents/words into different categories (Omar & Tjahyanto, 2018). For instance, some features that allocated in the email system like spam filtering, categories the email are using the text classification method.

Text classification which is different from the document class based on the category (Zhang & Wan, 2017). To perform text classification, there are several techniques and methods available, such as Support Vector Machine (SVM), WordNet, decision tree and etc.

Next, (Lu & Ding , 2010) has stated that by utilizing Vector Space Models (VSM) and WordNet to develop the semantic vector of emails with a specific end goal to screen and arrange emails into suitable classifications. Those two tools can easy to arrange the documents vectors into a matrix format for other processing. Unfortunately, VSM has a limitation on understanding the human language.

Moreover, Microsoft Outlook using text classification with Support Vector Machine (SVM) method to perform identify junk email. Microsoft Outlook moving from manual construction pattern to support vector machine to identify junk email in order to save the time-consuming. By using the words recognized by the system to build a dictionary to have more effect on auto reorganization system for the identify junk email system (Woitaszek & Muhammad Shaaban, 2003).

Furthermore, monitoring the email transaction logs can keep track of the unordinary occasion that occurrence in the email conversation. (Esichaikul, Guha, & Juntapoln, 2010) stated that go through the email content one by one it is very time-consuming. Therefore, reduce the number of email content or filter out the keywords before mining content is necessary. By monitoring the email transaction logs, it can easily to find out the important contents. It would easily for the analyze work. Besides, it can also increase the correctness of the system and reducing the processing time at the same time.

2.4 Text Mining

Text mining is an intelligence technique that performs conversion on unstructured text into meaningful data. There are several research fields related to the text mining which is text summarization, text classification, text clustering and text association analysis (Zhang & Wan, 2017). In the functional stage, text mining algorithm several into four big categorize which are pre-processing tasks, core mining operations, presentation layer components, and browsing, functionality and, refinement techniques (Feldman & Sanger, 2007). Pre-processing tasks which will group the information from the original source and convert into a canonical format before extraction method. Core mining operation which is the heart of the text mining technique. It consists of a pattern and trend analysis and also the knowledge discovery algorithms. As shown in the figure below. There are some application/ features are using text mining technique which are spam filtering, automatic text translation, fraud detection and etc. Spam filtering uses text mining in order to differentiate between the

needed email and unwanted email. Automatic text translation uses text mining technique to increase the accuracy on converting the input text from one language to another language (Miner, et al., 2012).
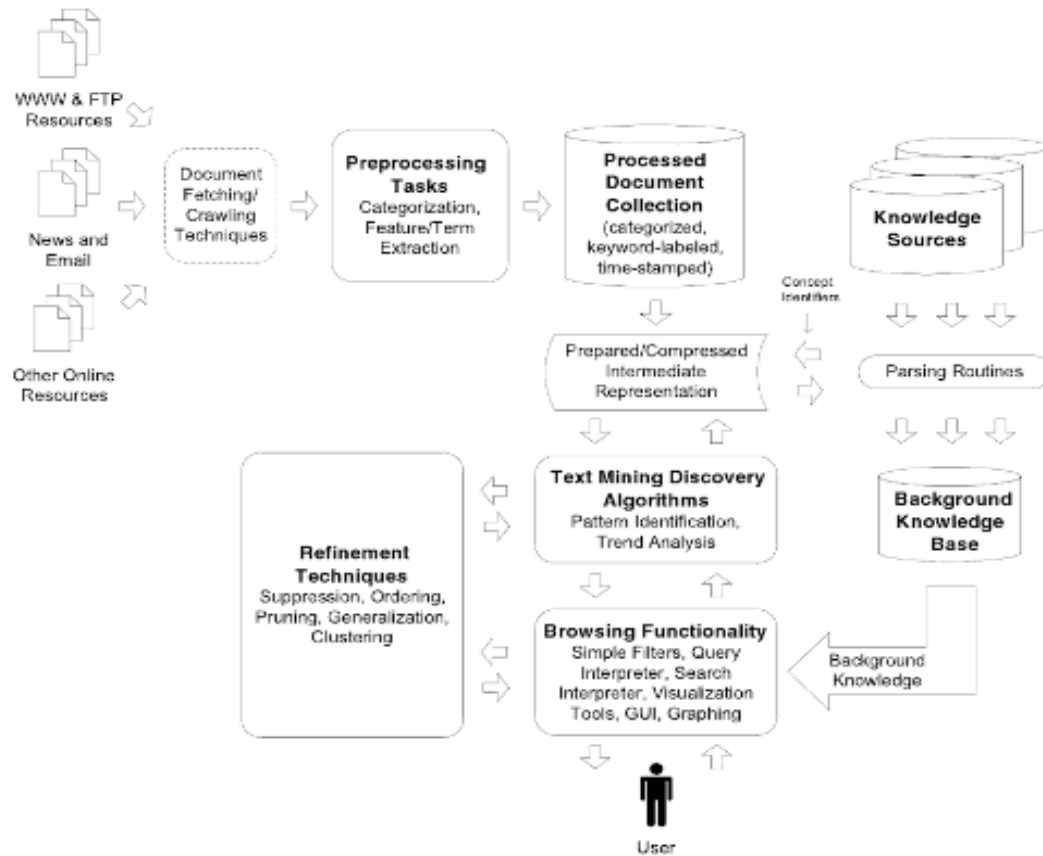
Figure 2.1 Process of Text Mining

2.5 Machine Learning

Learning forms incorporate the obtaining of new revelatory information, the advancement of the engine and psychological abilities through guideline or practice, the association of new learning into general, viable portrayal and the disclosure of new realities and hypotheses however perceptions and experimentation (Carbonell, Michalski, & Mitchell, 2014). Machine learning is a kind of study which is a model to learn it automatically from the experience based on the statistical models. Over time, train with more data and the prediction will more accurate (Dangeti, 2017). By using the users' existing purchase data, can provide recommendation product for the users through machine learning (Benchettara, kanawati, & Rouveirol, 2010).

2.6 Recommendation

The recommendation is a list of information of items that can help the user to decide something or to make a conclusion on it based on the user preferences (Paul, Neophytos , Mitesh, Peter , & John , 1994) (Will, Larry , Mark, & George , 1995). There have 4 types of recommendation algorithm which are a content-based recommendation, collaborative filtering recommendation, knowledge-based recommendation and hybrid recommendation (Fujiang , Yu, & Weiping, 2013). Data of behavior user and the features of the item is used to generate a recommendation in Content-based recommendation (Fujiang , Yu, & Weiping, 2013). Collaborative filtering recommendation normally is depending on the mass of the item to provide a recommendation, but it had a serious limitation which is low accuracy on the calculation for the similarity content (Ying Wei , Xin, & Yong Ge, 2012). A recommendation cannot fit all the user, it might be satisfied by an individual but dissatisfy by another member of group, so hybrid recommendation it used to provide a recommendation that satisfies by all the members in a group (Yuankun, Shujuan, Yongquan, Jianli, & Yongfeng, 2015).

2.7 Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency is unity and used to recognize the substance of content. In Term based strategy each term in the archive is related to important known as weight. This is utilized to quantify the significance of the term (Brindha, Prabha, & Sukumaran, 2016). Term Frequency is calculating the frequency of the word in each document. It represents the ratio of a number of words that appear in a document compared to the total number of words in each document. The formula for Term Frequency as below:

$$\text{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

n = the number of words

Inverse Data Frequency(idf) is using to calculate the weight of the word that appears very less time within the text. The word that normally appears almost every sentence or appears many within the text, it will be given a high IDF score. In order to calculate the idf, we use formula, the formula as follow:

$$\text{idf(w)} = log\frac{N}{df_t}$$

By combining two scores which is term frequency score and inverse data frequency score, it will show which word among all the word are more significance. It using the formula which is the product of tf and idf.

$$W_{i,j} = tf_{i,j} * log\frac{N}{df_t}$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

2.8 Remarks

From the previous research works, there is no similar application on providing a recommendation based on the email contents. Thus, little literature or information available for this research topic. However, review on text mining, machine learning, recommendation method, and term frequency showed that their principles and mechanisms are workable to complete this proposed idea. Therefore, it is confirmed to proceed to the evaluation of the techniques through the building of the module, which will be discussed in the next chapter.

Chapter 3: System Design

3.1 Design Specification

The current email application didn't serve the user with the recommendation service. We are only able to use this kind of service on some website like the video-sharing website, shopping website, social website and etc. Most of the recommendation service in the existing market, that provide a recommendation based on the history that created by the users.

A recommendation module that works together with the email application, can help the user to completed tasks without having a solution. By using the recommendation, it can help the user to choose the right apps to help them to solve the action required in the email. It might enhance the user experience and decrease the time consuming on thinking the idea at the same time.

3.2 System Design
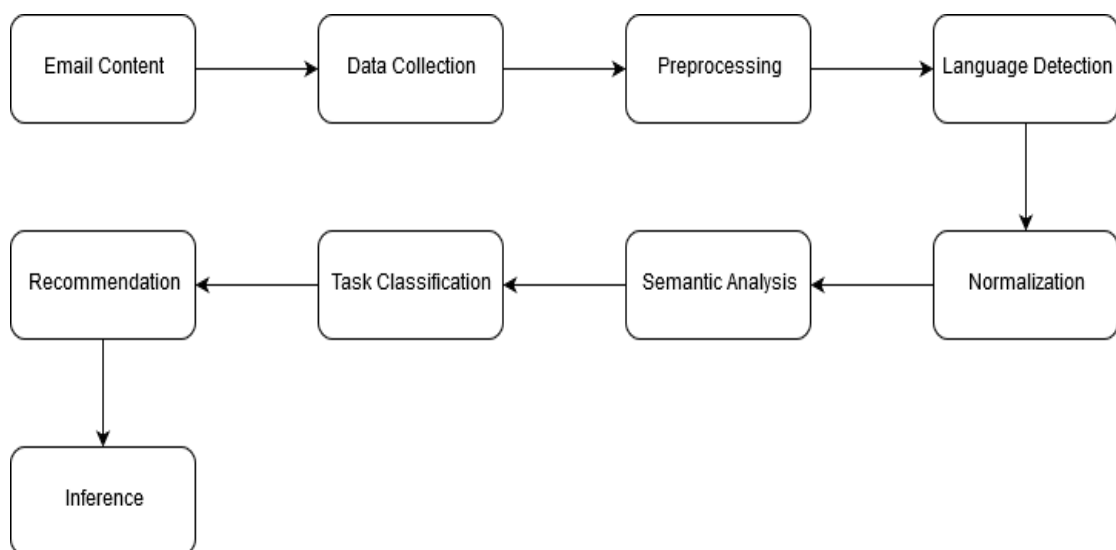
Figure 3 System Design

Email content as the input data will become the raw data for this recommendation module. The raw data will pass to data collection, it will be gathering all the necessary data from the input data, then pass to the preprocessing stage for processing. It will be cleaning all the not necessary data during this stage.

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

Besides, all the data will be pass to language detection to make sure that all the data is in English. After verifying the data language, it will pass to the normalization stage for structuring the data. Thus, it will also remove all the redundancy data during this stage. In order to improve data integrity. Then, it will pass to the semantic analysis stage for verb analysis by using the Google NPL service. Then, by getting the verb analysis result it will pass to the task classification stage. During the task classification, it will be matching the verb with the prepared data set to generate the recommendation. Then, the final recommendation will pass to the inference stage. Based on the recommendation result, the module can inference that the recommendation result can help the user to solve their task that required in the email.

3.3 Block Diagram
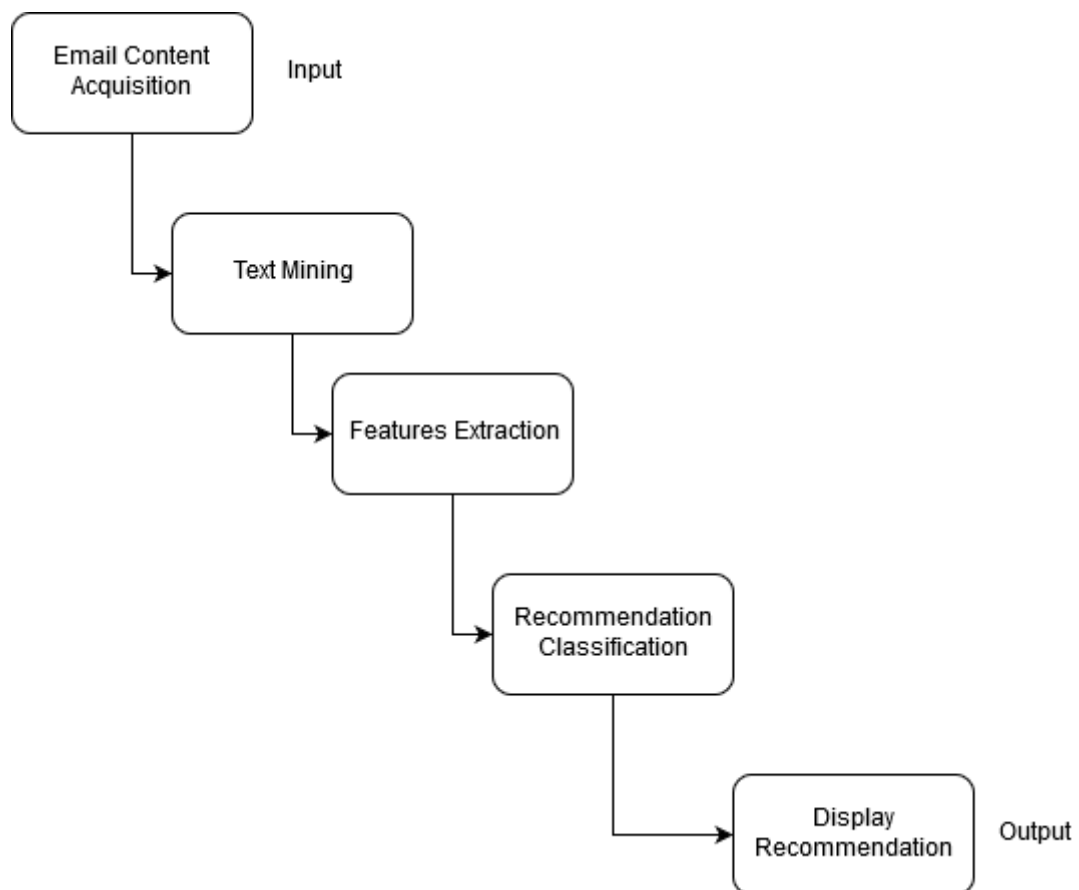
Figure 3.1 Block Diagram

Figure 3.1 shows the overall system design of the email content recommendation module. The system can separate into five stages which are email content acquisition, text mining, features extraction, recommendation classification, display recommendation. Each stage must be completed before the pass to the next stage. The email content acquisition is the stage that the module captures the message as the input and prepare it for the next stage. The next three stages are for the core processing stage of this module. The last stage which is output and displays the recommendation to the end user.

3.4 Use Case Diagram
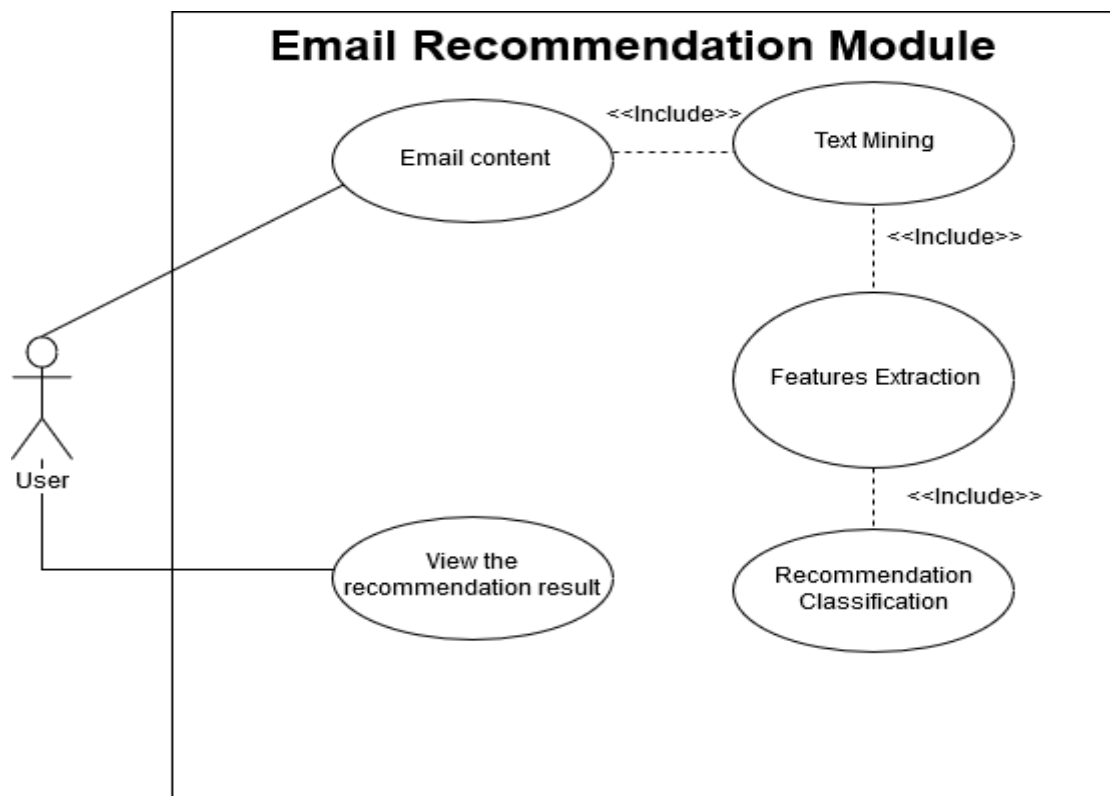
Figure 3.2 Use case Diagram

Figure 3.2 shows that the use diagram has one actor which is the user only. The user is the one that can provide email content and receive the recommendation result from the module. Besides, this use case diagram consists of multiple use case which is email content acquisition, text mining, features extraction, and recommendation classification. The use cases of this module as below show the action perform by the actors and the expected outcome.

Chapter 3: System Design

Use Case 1: Email Content Acquisition

Actor: User

Goal: To capture the user's email content as the system input.

Overview: Read the email content and store in an array list for the next stage.

Use Case 2: Text Mining

Actor: User

Goal: To process the email content and extract the information out.

Overview: To perform analysis of the email content, to convert the content into meaningful data.

Use Case 3: Features Extraction

Actor: User

Goal: To obtain a set of rule features information used for analysis.

Overview: It is the process to transform the data into a set of grammar data set such as entity, verb, noun, adverb and etc.

Use Case 4: Recommendation Classification

Actor: User

Goal: To match the important data with the dataset in order to provide a recommendation.

Overview: It is the process to match the keywords that get from the email content to match with the dataset to generate the recommendation.

3.5 Activity Diagram

Activity 1: Input the email content to the module.
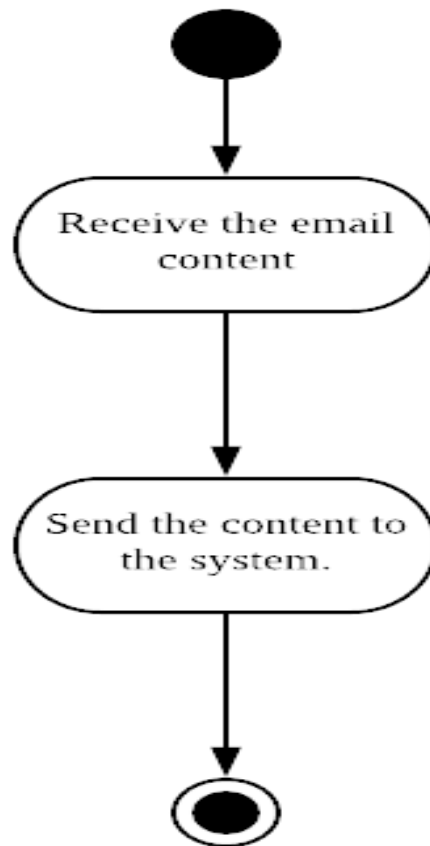
Figure 3.3 Activity Diagram 1

This activity diagram shows the start state of this module. The user needs to input their email content to the text file. Then, the module will read the text file and send to the system.

Activity 2: store the content with a score larger than 0.5 into the array list.
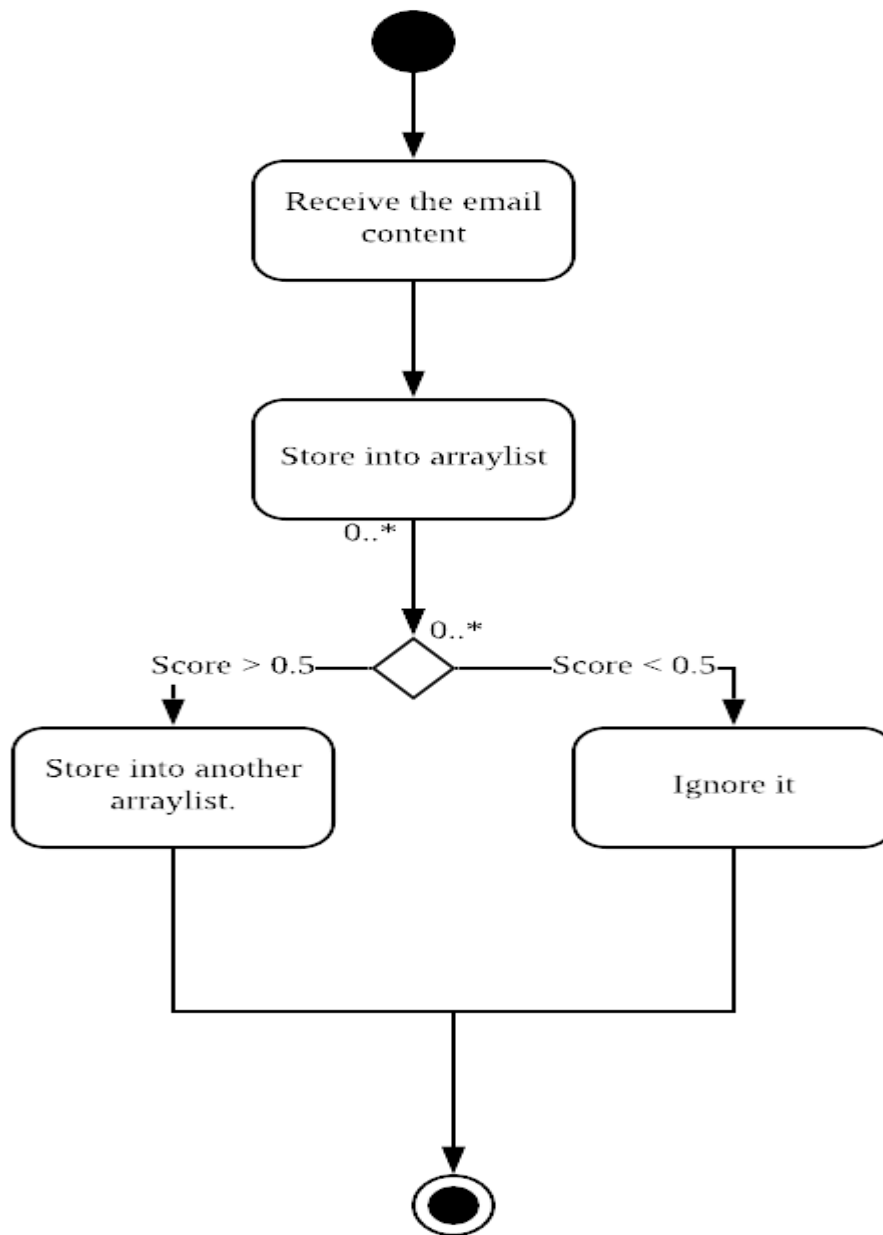
Figure 3.4 Activity Diagram 2

This activity diagram shows the process of selecting the important sentence among the text. The module will receive the email content from the system which is input by the user. The module will perform the calculation on each sentence and tell the module among all the score of the sentence. Then, the module will store the score that more than 0.5 into another list, and it will ignore all the sentence that less than 0.5. The score range is between   -1 and 1.

Activity 3: Analysis of the email content
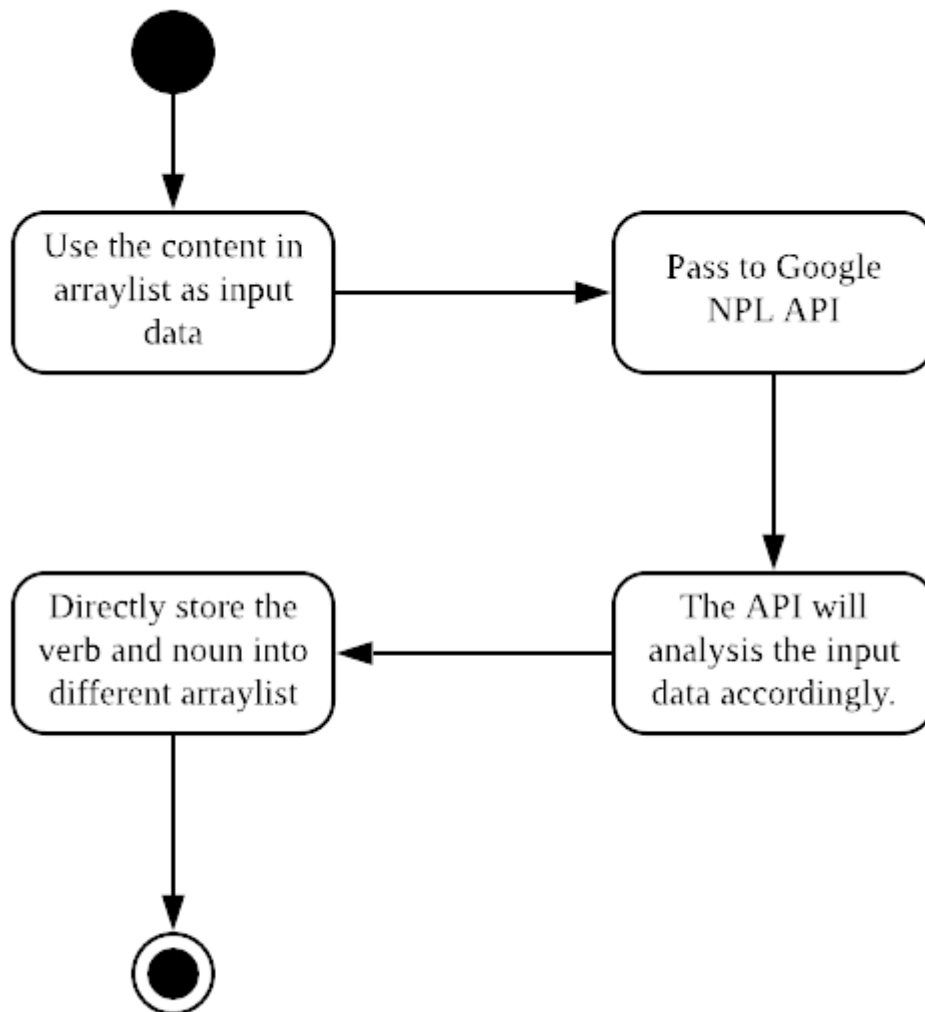
Figure 3.5 Activity Diagram 3

This activity shows how module applies Google Natural Language Processing API to perform the job. The module will pass the sentence that higher or equal to 0.5 scores to this API to perform analysis on it. The API will perform itself when sending the data to it. The result of the analysis will generate a larger number of different grammar type, such as verb, noun, adjective, adverb, etc. But this module only needs a verb and noun to perform the next action. Thus, the module will only capture and store the noun and verb that output by the API.

Activity 4: Perform Matching on multiple results
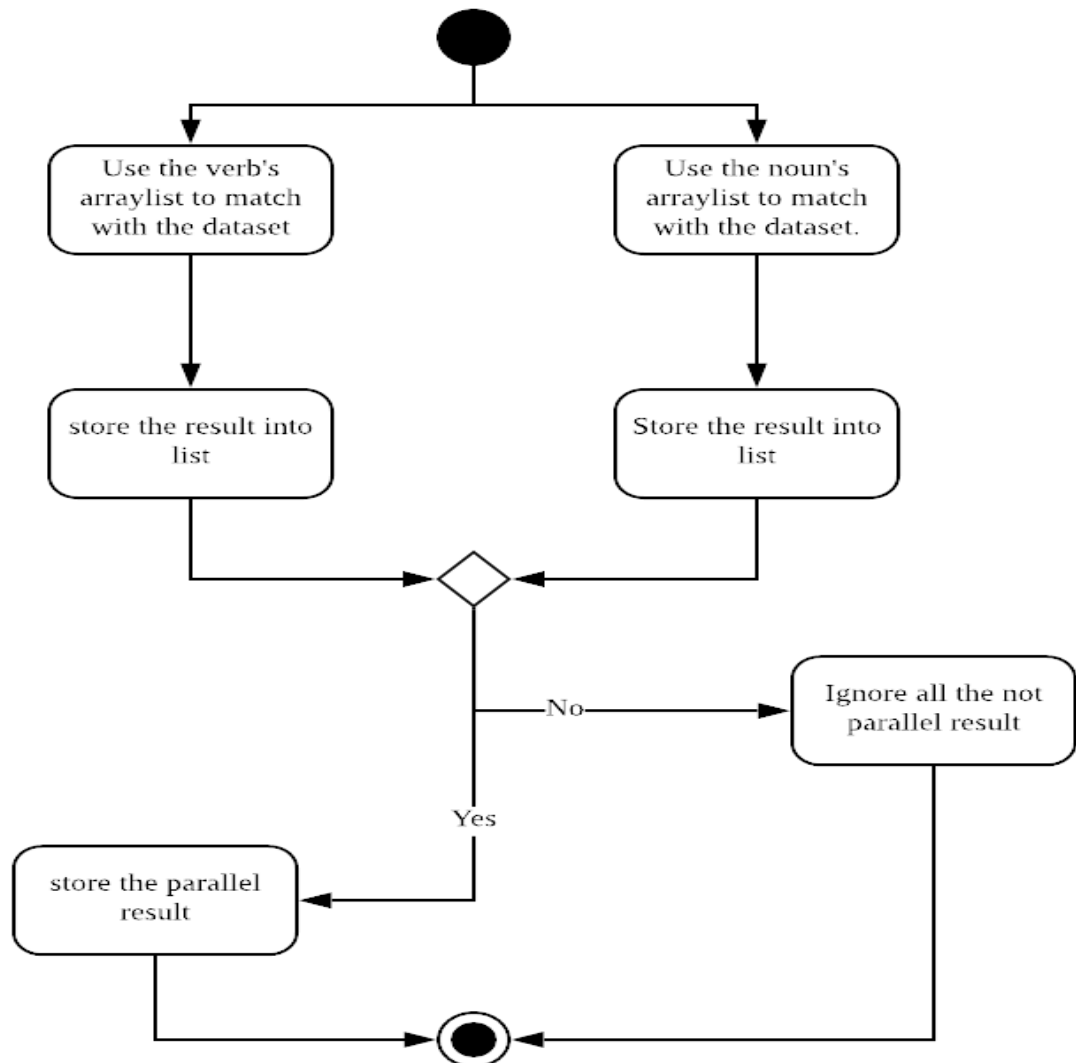
Figure 3.6 Activity Diagram 4

This activity shows the matching process on the result of the recommendation for the user. This module will perform a parallel process on this stage, it will run the matching on verb and noun according to the different dataset that prepared in the early stage. The matching result of verb and noun will match again in order to increase the accuracy on providing the recommendation.

Activity 5: Save the result and display to the user

Figure 3.7 Activity Diagram 5

This activity shows the flow on how the module shows the recommendation result to the user. The module will capture the result that output by the system. Then, the module will store the final answer into a result list before the show to the end user.

3.6 Term Frequency – Inverse Document Frequency

This module also using a text optimization call term frequency. It is a numerical statistic to reflect on how important a word to a particular document or text. It is calculating the frequency of each word appear within a text, and calculate the weight of each word of the text. Term Frequency using the formula is the number(n) of times a term(t) appears in a document(d) to the total number(N) of terms in the documents. [tf(t,d) = n/N]. Besides, in order to remove out the most common word used in the documents, it will apply Inverse Document Frequency to calculate the ratio of the number of documents present in the data set to the number of documents that contain the term t.

3.7 Hidden Markov Model
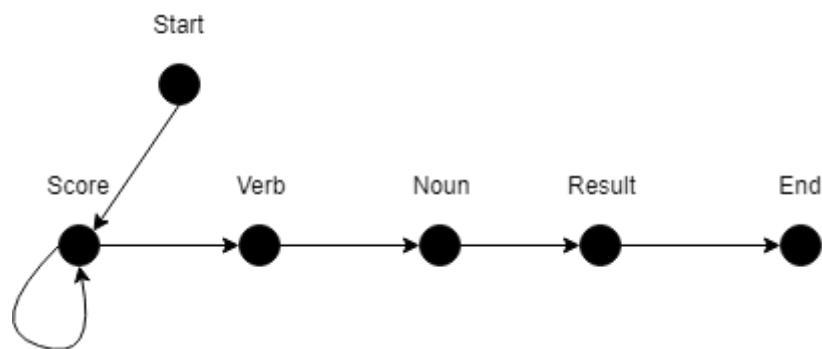
Figure 3.8 Hidden Markov Model

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved states. There are 4 states in our HMM model which is score, verb, noun, result. The transition probabilities between these states were estimated each transition occurs in the training set by the sum of all the transitions. The state probabilities were calculated from the scoring output that was reported by the text classification.

3.8 Implementation Issues and Challenges

There are few limitations founded in this project. Due to the limitation of the technology, the recommendation module may not produce the ideal result by the system. Below is the limitation of the proposed recommendation system:

3.8.1 Email Content

The writing language that user use for communication will not definite is English, it might be some other language such as Chinese, Germany, Japanese, and etc. The text mining technique in this recommendation system which also only allow the data in English to perform text mining action. Therefore, the text language that other than English will not recognize by the system.

3.8.2 Noise

The expression "noise" has negative undertones in scientific contexts. Noise is mutilation in information, that is undesirable by the perceiver of information. Noise is anything that is misleading and incidental to the first information, that isn't expected to be available in any case, however, was acquainted due with defective catching procedure. If train too much noise data into the machine learning algorithm it might affect the final result.

3.8.3 Writing Style

Each user has his/her own style in writing the message, some of them might use their own style / own create a short form of words. It will be bringing some trouble for the text mining technique. It is because the text mining technique cannot recognize all type of short form created by the millions of users

Chapter 4: Discussions

4.1 Methodology
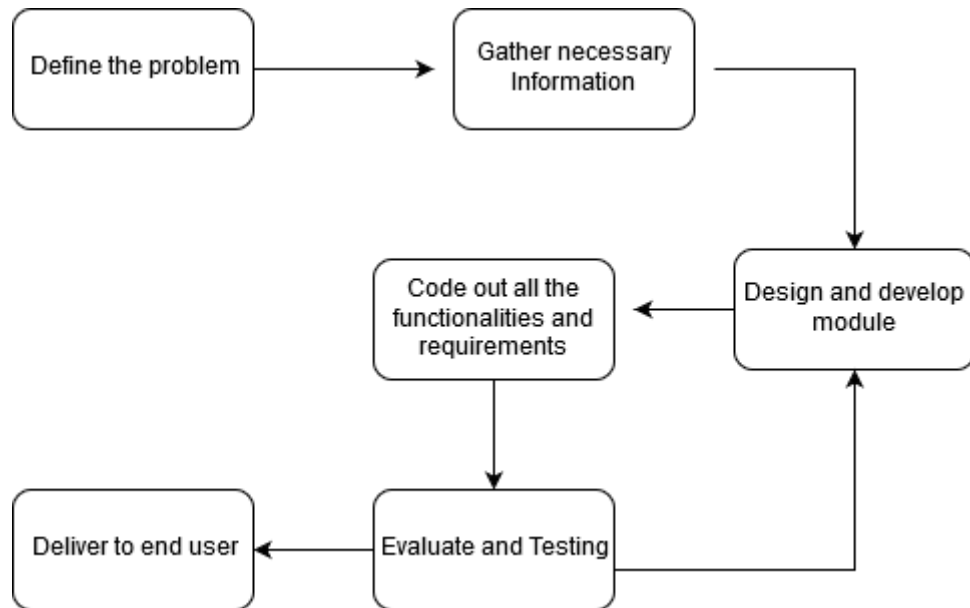
Figure 4.0 Module Life Cycle Model

Figure 4.0, shows the methodology for developing the proposed email recommendation module. All the objective and goal of this module will be identified during the planning stage. The project scope will determine all the necessary thing needed to include in the proposed module.

During the analysis phase, all the survey on technique and design will be carried out at this stage. In order to make sure that the final recommendation module had met the objective of this module. Some UML diagram also will design in order to let the developer and another user understand the design of the module, and also the module structure looks like.

In the prototyping stage, a recommendation module will be developed and all the necessary requirements and functionalities will be carrying out during this stage. This module will perform several testing such as unit testing, integration testing, system testing, and acceptance testing. Then the supervisor and the developer himself will be the tester on this module testing. Moreover, all the bugs and dissatisfaction result will bring this module back to the prototyping stage to fix the error.

Once the module is successfully testing only, this module will proceed to deployment stage and deploy the module to the end user.

4.2 Technology Involved

4.2.1 Software

i. Eclipse

Figure 4.1 Logo of Eclipse

Eclipse is an Integrated Development Environment (IDE) used in computer programming and is the most widely used Java IDE. It is a great platform that contains a base workspace and extensible plug-in system for customizing the environment.

4.2.2 Hardware

Figure 4.2 Asus Laptop

The hardware that used to develop this module includes an Asus K551LN laptop that included with specifications as below:

- Processor: Intel® Core ™ i5-4210U CPU @.70GHz 2.40GHz
- Installed Memory (RAM): 8GB SDRAM
- Operating System: Window 10 Home Single Language
- Graphics Card: NVIDIA GeForce GT840M with 4GB DDR3 VRAM
- Storage: 1TB HDD & 250 SSD

4.2.3 Programming Language

      i.   Java JDK 1.8.0

Figure 4.3 Java Logo

This module will be written in Java language which provides syntax that very easy for the coder to code in a faster way. Besides, Java also provides various libraries and Application Programming Interface (API) which can complete a complex task in a simple way.

4.2.4 Technique Involved

      i.   Text Mining

Text mining which is also called text classification. It is the process of classifying the text into predefined categories based on the text. Text mining can turn the text into useful data for future analysis. The email recommendation module also using text mining. The system will read email content as the raw data. Then the raw data will send to the text mining, in order to transform raw data into useful/meaningful data.

ii.   Machine Learning

Machine learning is a way of data analysis that automates analytical model building. Machine learning allows the computer/system to learn themselves without being programmed to complete any work. This learning ability is important to develop a smart system, it able to identifying the patterns and turning the data into forecasts. Machine learning is using in the email recommendation module to train the algorithm to provide a recommendation based on the content data. The algorithm will be trained by the data and validate by the new data, in order to improve the accuracy. Then, the system can perform more accurate in providing a recommendation.

4.2.5 Library and Service

i.   Google Natural Language API
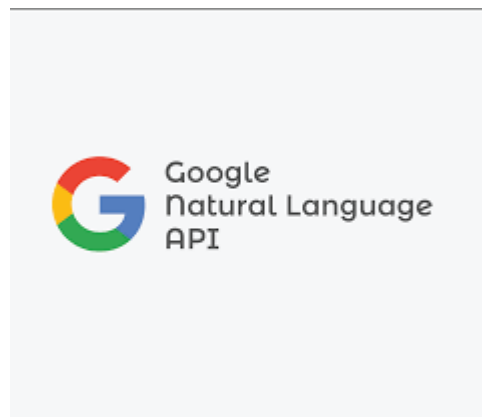
Figure 4.4 Logo of Natural Language API

Google Natural Language API playing a very important role in this module. It is a Natural Language Processing (NLP) service created by Google. In order to use this service, it must have a google cloud account before starting to build the module with this service. It is a simple and easy use service. It also allows the developer to do some customization by their own depending on their own project.

4.3 Functional Requirements

4.1.1   The system is able to retrieve the user's message from the text file.

4.1.2   The system is able to pass the message to the text mining features.

4.1.3   The system is able to analyse the message in order to find out the important keywords.

4.1.4   The system is able to pass the important keywords to match with the data.

4.1.5   The system is able to provide a recommendation based on the user's message.

4.1.6   The system is able to display the recommendation result to the user.

4.4 Assumption

In order to provide a recommendation on email content and avoid to make the module error, there is some assumption need to include before starting the module:

- The email contents should write in English.
- The email contents cannot consist of any short form of words.

Chapter 5: Implementation and Testing

5.1 System Implementation

This section will have a discussion on the system implementation that how a module to archive the output correctly and how to implement the technique on the module. This project is a recommendation related project. Google Natural Language Processing had been implemented in this project. Besides, it also implements the Hidden Markov Model (HMM) in order to increase the accuracy on providing the recommendation for the end user and also reduce the noise that appears within the documents.

5.1.1 Hidden Markov Model (HMM)

Grammar Type	Before HMM	After HMM	Number change after HMM
Verb (Number of the word)	200	235	35
Noun (Number of the word)	235	245	10

Table 5.0 Table of HMM

The training data has about 300 sentences. Before implement HMM, the result of Natural Language Processing (NLP) shows that 300 sentences, consists of 200 verbs and 235 Noun. After, implement HMM in this module, by running again the training set, it gave a new result, which consists of 235 verbs and 245 nouns within the training set. By implementing the HMM, it will increase the number of finding out the accurate verb and noun within the document. It will also increase the accuracy of searching out the verb and noun from the document.

5.2 Black-Box Testing

Test	Test Case	Expected Outcome	Actual Outcome	Result
1	Email Content Acquisition	Read the user's email content as input data	Display email content in console	Pass
2	Sentence's score	Able to display all the sentence with a score	Display sentence with a score	Pass
3	Text analysis process	Able to pass the content to API for processing	Able to display the analysis result	Pass
4	Store the specific keywords	Able to store the specific keywords into an array list	Able to store the word that had been mention into an array list	Pass
5	Matching process	Able to match the keyword with the dataset	Able to perform the matching process	Pass
6	Display result	Able to display the result to end user	Recommendation is displayed	Pass

Table 5.1 Table of Black-Box Testing

Table 5.1, show about the black-box testing result. It also shows all the necessary test case had been met the functional requirement that had been stated in the early state.

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

5.3 Performance Testing

Analysis Testing

For this testing part in this module, there are some sample sentence will be tested on the module. For a visual result, this testing only applies to the analysis text. This testing will be showing the content that user input and analytic it and display the verb that appears within the user input data. Besides, this Natural Language Processing (NPL) service is provided by the Google Cloud. The user's input will pass by the module to the NPL API for analysis, and the result return by the API will be preprocessing before showing to the user. This testing is applying on three people, which are Chong Yee Xiang, Kee Ming Wei, and Ong Jun Kai. The sample sentences are shown below:

1. How to buy an iPhone online.
2. Where can eat nearby?
3. How to drive to Singapore.
4. How you normally travel around KL.
5. What application that you normally use to purchase a ticket.
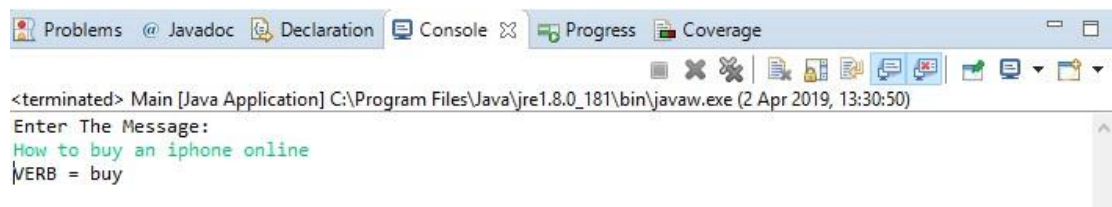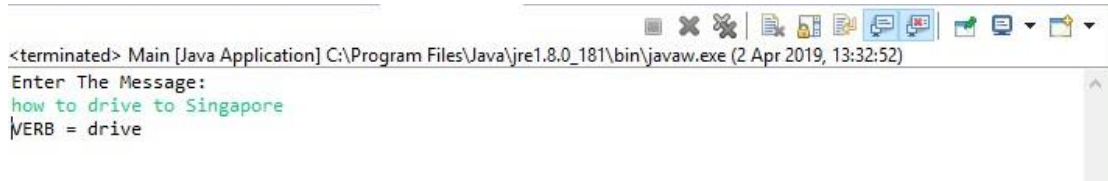
Figure 5.0 Result of Sentence 1

Figure 5.1 Result of Sentence 2

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

```
<terminated> Main [Java Application] C:\Program Files\Java\jre1.8.0_181\bin\javaw.exe (2 Apr 2019, 13:32:52)
Enter The Message:
how to drive to Singapore
VERB = drive
```

Figure 5.2 Result of Sentence 3


```
<terminated> Main [Java Application] C:\Program Files\Java\jre1.8.0_181\bin\javaw.exe (2 Apr 2019, 13:33:29)
Enter The Message:
How you normally travel around KL
VERB = travel
```

Figure 5.3 Result of Sentence 4


```
<terminated> Main [Java Application] C:\Program Files\Java\jre1.8.0_181\bin\javaw.exe (2 Apr 2019, 13:34:50)
Enter The Message:
what application that you normally use to purchase ticket
VERB = use
VERB = purchase
```

Figure 5.4 Result of Sentence 5

Recommendation Testing

For this testing part in this module, there are some sample sentence will be tested on the module. For a visual result, this testing only applies to the generated recommendation. This testing will be showing how a "verb" can provide a recommendation. This process using a matching algorithm that uses the verb that appears within the user input data and matches with a dataset that prepared early. This testing is applying on three people, which are Chong Yee Xiang, Kee Ming Wei, and Ong Jun Kai. The sample sentences are shown below:

1. Do you have any recommended place around Ipoh?
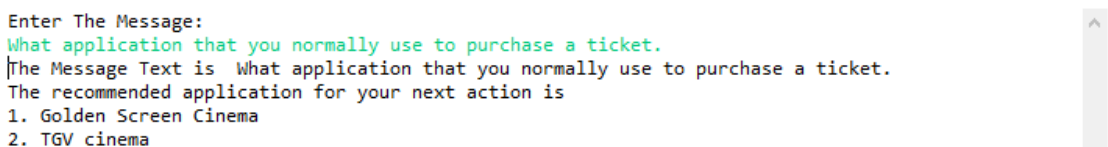2. What application you normally use to purchase a ticket.
3. How you reload your phone internet plan?
4. How you travel around KL without having a car?
5. How you avoid traffic jam while driving in KL?


```
<terminated> Main [Java Application] C:\Program Files\Java\jre1.8.0_181\bin\javaw.exe (4 Apr 2019, 10:31:19 PM)
Enter The Message:
Do you have any recommended place around ipoh?
The Message Text is Do you have any recommended place around ipoh?
The recommended application for your next action is
1. Google Maps
```

Figure 5.5 Result of Recommendation 1
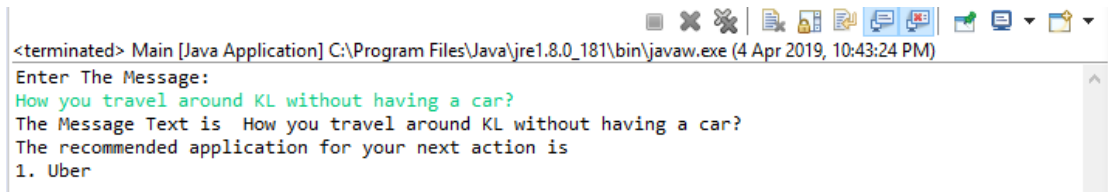

```
Enter The Message:
What application that you normally use to purchase a ticket.
The Message Text is  What application that you normally use to purchase a ticket.
The recommended application for your next action is
1. Golden Screen Cinema
2. TGV cinema
```

Figure 5.6 Result of Recommendation 2


```
<terminated> Main [Java Application] C:\Program Files\Java\jre1.8.0_181\bin\javaw.exe (4 Apr 2019, 10:41:44 PM)
Enter The Message:
How you reload your phone internet plan?
The Message Text is  How you reload your phone internet plan?
The recommended application for your next action is
1. Maybank MY
2. Touch 'n Go eWallet
```

Figure 5.7 Result of Recommendation 3

```
 <terminated> Main [Java Application] C:\Program Files\Java\jre1.8.0_181\bin\javaw.exe (4 Apr 2019, 10:43:24 PM)
Enter The Message:
How you travel around KL without having a car?
The Message Text is  How you travel around KL without having a car?
The recommended application for your next action is
1. Uber
```

Figure 5.8 Result of Recommendation 4
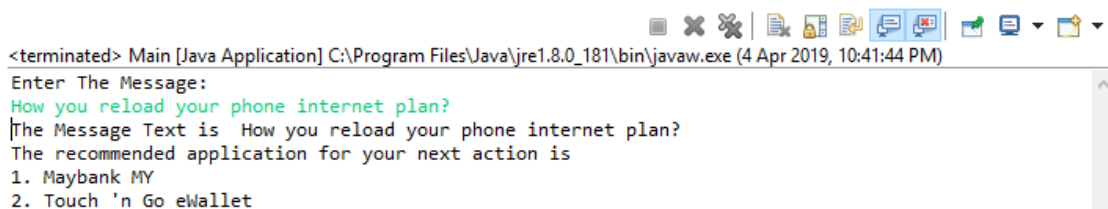


```
 <terminated> Main [Java Application] C:\Program Files\Java\jre1.8.0_181\bin\javaw.exe (4 Apr 2019, 10:45:50 PM)
Enter The Message:
How you avoid traffic jam while driving in KL.
The Message Text is  How you avoid traffic jam while driving in KL.
The recommended application for your next action is
1. Waze
```

Figure 5.9 Result of Recommendation 5

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

5.4 User Acceptance Testing

Table 5.2 below shows some recommendation result for some sample email content that had been asking some interviewee to provide their acceptance on it.

Test	Sample Email Content	Recommendation	ARR (%)
1.	How to buy an iPhone online.	-Lazada -Shopee	83.3%
2.	Where can eat nearby?	-Google Maps	86.7%
3.	How to drive to Singapore.	-Waze -Google Maps	93.3%
4.	How you normally travel around KL	-Uber	93.3%
5.	What application you normally use to purchase a ticket.	-Golden Screen Cinema -TGV cinema	90.00%

Table 5.2 Table of User Acceptance Result

Table 5.2 shows that the acceptance results in a percentage of 30 interviewees. All the test case with the result of 90% and above, show that our recommendation for email content had met the requirement to solve the test case problem. In order to show the test case recommendation, the prepared dataset consists of 20 Top daily application from the Apple Apps Store. All the interviewee are the degree student and lecturer of UTAR. The reason to choose degree student and lecturer as interviewee because most of the time, the lecturer has to use their email to communicate with their colleagues. Besides, degree student also needs to use email as their communication tool with their lecturer on appointment or inquiry. This is the reason why the interview section consists of these two groups of people.

Chapter 6: Conclusion

In short, it can be said that Email is one of the most favorable communication tools for people to exchange message or communicate nowadays. Email application had been added more and more features compare to the first generation of email which can only be sent and receive email.

The current email application may not be the most effective application. When a user needs to find an external solution based on the email message, he/she need to switch between different applications. The user also needs to think that which application is the most effective to solve the problem. An email application that works with the recommendation system to provide a recommendation based on the content of the email can provide an effective way of solving the user's problem. By using the result that recommends by the recommendation module, it can save the user's time and enhance the user' experience.

Besides, there are also some issue and challenge in this project, which is the text mining only recognize for the English. The word that other than English will be caused the system into error and cannot text mining the sentence or the certain word. Moreover, the system also cannot recognize the shortcut keyword that created by the user themselves. All the shortcut keyword will wrongly text mining and it will become data noise for the machines learning.

This project is applying two main technique which is text mining and machine learning. The content that resolves by the text mining technique generates a text file. The text file will be read by the machine learning algorithm and find the keyword to provide a recommendation based on the keyword.

This concept and the system will be fully tested before the delivery date. This module will work with a certain email application in the existing market to test the recommendation system about when the email application received a new email and will it recommend something based on the email content? This project will more focus on delivery more accurate recommendation to the users if not there is no point on developing this kind of module to help the people.

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

References

Andersson, H. (2018, July 4). *Social media apps are 'deliberately' addictive to users.* Retrieved from BBC News: https://www.bbc.com/news/technology-44640959

Benchettara, N., kanawati, R., & Rouveirol, C. (2010). SUpervised Machine Learning applied to Link Prediction In Bipartite Scocial Networks. *2010 Internation COnference on Advances in Social Networks Analysis and Mining*, 326-330.

Brindha, S., Prabha, K., & Sukumaran, S. (2016). The Comparison of Term Based Methods Using Text Mining. *International Journal of Computer Science and Mobile Computing*, 112-116.

Carbonell, J., Michalski, R., & Mitchell, T. (2014). *Machine Learning: An Artificial Intelligence Approach.* Palo Alto, Calif: Tioga Pub Co.

Dangeti, P. (2017). *Statistics for Machine Learning.* United Kingdom : Packt Publishing Ltd.

Esichaikul, V., Guha, S., & Juntapoln, C. (2010). Monitoring Email Transaction Logs by Text-Mining Email Contents. *The 3rd International Conference on Data Mining and Intelligent Information Technology Applications*, 255-258.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructed Data.* United Stated of America: Cambridge University Press.

Fujiang , S., Yu, S., & Weiping, W. (2013). Content-based Recommendation System Based on Vague Sets. *2013 Fifth International Conference on Intelligent Human-Machine Systems and Cybernetics*, 294-297.

Gilad, I. (2013, May 29). *A new inbox that puts you back in control .* Retrieved from Official Gmail Blog: https://gmail.googleblog.com/2013/05/a-new-inbox-that-puts-you-back-in.html

Lu , Z., & Ding , J. (2010). An Efficient Semantic VSM based Email Categorization Method. *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, 525-530.

Miner, G., Hill, T., Elder , J., Nisbet, B., Delen, D., & Fast, A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* USA: Academic Press.

Omar , R. A., & Tjahyanto, D. (2018). Evaluation of the Performance of Machine Learning Algorithm in Swahili-English Emails Filtering System Relative to Gmail Classifier. *2018 International COnference on Information and Communication Technology (ICOIACT)*, 266-269.

Paul, R., Neophytos , I., Mitesh, S., Peter , B., & John , R. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *Proceedings of the 1994 ACM conference on Computer supported cooperative work* , pp. 175-186 .

Skulimowski, A. M., & Kacprzyk, J. (2016). *Knowledge Information and Creativity Support Systems: Recent Trends, Advances and Solutions.* Switzerland: Springer International Publishing.

The office of Chief Statistician Malaysia. (2018, April 2). *Current Population Estimates, Malaysia, 2013-2017.* Retrieved from https://www.dosm.gov.my/v1/index.php?r=column/cone&menu_id=dDM2e nNvM09oTGtQemZPVzRTWENmZz09

Appendix

Warren, T. (2018, May 1). *Mircosoft reveals lots of new Outlook features following Gmail redesign*. Retrieved from THEVERGE: https://www.theverge.com/2018/5/1/17306162/microsoft-outlook-features-update-may-2018

Will, H., Larry , S., Mark, R., & George , F. (1995). Recommending and Evaluating Choices in a Virtual Community of Use. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* , pp. 194-201 .

Woitaszek, M., & Muhammad Shaaban. (2003). Identifying Junk Electronic Mail in Microsoft Outlook with a Support Vector Machine. *Proceedings of the 2003 Symposium on Applications and the Internet (SAINT'03)*, 166-169.

Ying Wei , C., Xin, X., & Yong Ge, S. (2012). A Collaborative Filtering Recommendation Algorithm Based on Contents' Genome. *IET International Conference on Information Science and Control Engineering 2012* .

Yuankun, M., Shujuan, J., Yongquan, L., Jianli, Z., & Yongfeng, C. (2015). A Hybrid Recommendation List Aggregation Algorithm for Group Recommendation. *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.

Zhang, Y., & Wan, Y. (2017). How to Find Valuable References? Applciation of Text Mining in Abstract Clustering. *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2017)*, 2201-2205.

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

Appendices

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

Appendix A Final Year Project 2 Biweekly Report

FINAL YEAR PROJECT WEEKLY REPORT
(Project I / Project II)

Trimester, Year: T3, Y3	Study week no.: 2
Student Name & ID: Tang Zi Cong 15ACB05224	
Supervisor: Dr. Aun Yichiet	
Project Title: Context Aware App Recommendation Using Email Semantic Analysis	

1. WORK DONE Previous work was done during FYP1.
2. WORK TO BE DONE Do review on recommendation, and understand the Natural Language Processing Model.
3. PROBLEMS ENCOUNTERED Lack of similar information and example from online.
4. SELF EVALUATION OF THE PROGRESS Slow progress, off schedule.

_____   _____
       Supervisor's signature   Student's signature

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

Appendices

FINAL YEAR PROJECT WEEKLY REPORT
(Project I / Project II)

Trimester, Year: T3, Y3	Study week no.: 5
Student Name & ID: Tang Zi Cong 15ACB05224	
Supervisor: Dr. Aun Yichiet	
Project Title: Context Aware App Recommendation Using Email Semantic Analysis	

1. WORK DONE

Finish the review and understanding of Natural Language Processing Model

2. WORK TO BE DONE

The solution on creating customization Natural Language Processing Model

3. PROBLEMS ENCOUNTERED

Lack of idea on reducing the unimportant keywords.

4. SELF EVALUATION OF THE PROGRESS

Fast progress, on schedule

_____ _____
Supervisor's signature Student's signature

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

FINAL YEAR PROJECT WEEKLY REPORT

(~~Project I~~ / Project II)

Trimester, Year: T3, Y3	Study week no.: 7
Student Name & ID: Tang Zi Cong 15ACB05224	
Supervisor: Dr. Aun Yichiet	
Project Title: Context Aware App Recommendation Using Email Semantic Analysis	

1. WORK DONE

A completed data set for recommendation is prepared to be used.

2. WORK TO BE DONE

Find the solution for generating the recommendation.

3. PROBLEMS ENCOUNTERED

Less understanding of building recommendation module.

4. SELF EVALUATION OF THE PROGRESS

Normal progress, on schedule

_____   _____
        Supervisor's signature   Student's signature

FINAL YEAR PROJECT WEEKLY REPORT
(~~Project I~~ / Project II)

Trimester, Year: T3, Y3	**Study week no.: 9**
Student Name & ID: Tang Zi Cong 15ACB05224	
Supervisor: Dr. Aun Yichiet	
Project Title: Context Aware App Recommendation Using Email Semantic Analysis	

---

1. WORK DONE

The job scope on provides recommendation and system requirement testing is done.

---

2. WORK TO BE DONE

- Documentation
- Success for generating the output recommendation

---

3. PROBLEMS ENCOUNTERED

Less understanding of software testing

---

4. SELF EVALUATION OF THE PROGRESS

Normal progress

_____   _____
      Supervisor's signature   Student's signature

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

FINAL YEAR PROJECT WEEKLY REPORT
(Project I / Project II)

Trimester, Year: T3, Y3	Study week no.: 11
Student Name & ID: Tang Zi Cong 15ACB05224	
Supervisor: Dr. Aun Yichiet	
Project Title: Context Aware App Recommendation Using Email Semantic Analysis	

1. WORK DONE
[Please write the details of the work done in the last fortnight.]

A result is generated from the system performance testing and the documentation is done.

2. WORK TO BE DONE

- Finalize documentation
- Prepare for presentation

3. PROBLEMS ENCOUNTERED

Documentation formatting issues

4. SELF EVALUATION OF THE PROGRESS

Normal progress

_____   _____

    Supervisor's signature   Student's signature

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

Appendix

Appendix B Survey Question

Survey on User Acceptance Result

UNIVERSITY TUNKU ABDUL RAHMAN
FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
BACHELOR OF INFORMATION SYSTEM (HONS) INFORMATION SYSTEM ENGINEERING

Final Year Project Title: Context Aware App Recommendation Using Email Semantic Analysis
This survey is focus on the user acceptance on the recommendation that provide by the user.

* Required

Procedure

You will be given a recommendation provided by the module based on the 5 question. You need to answer yes or not for the question.You had to think about whether the answer provide by the module, can solve the question or not. You are required to complete ALL the sections. This survey will take approximately 2 to 3 minutes to complete.

1. How to buy an iPhone online? Recommendation Application
1. Lazada 2. Shopee *

○ YES

○ No

2. Where can eat nearby? Recommendation Application 1.
Google Maps *

○ YES

○ No

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

3. How to drive to Singapore? Recommendation Application 1. Waze 2. Google Maps *

○ YES

○ No

4. How you normally travel around KL? Recommendation Application 1. Uber *

○ YES

○ No

5. What application you normally use to purchase a ticket? Recommendation Application 1. Golden Screen Cinema 2. TGV Cinema *
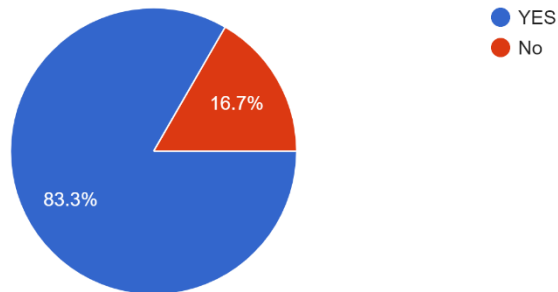
○ YES

○ No

GET LINK

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

Appendix C Survey Question Summary Result

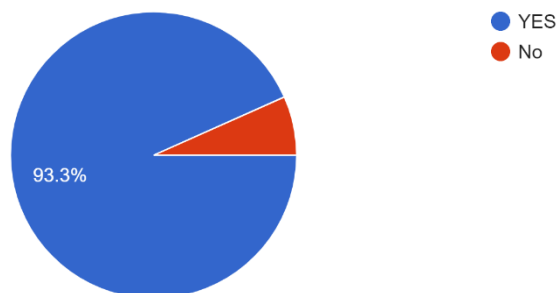1. How to buy an iPhone online? Recommend: 1. Lazada ; 2. Shopee

30 responses

2. Where can eat nearby? Recommend: 1. Google Maps

30 responses
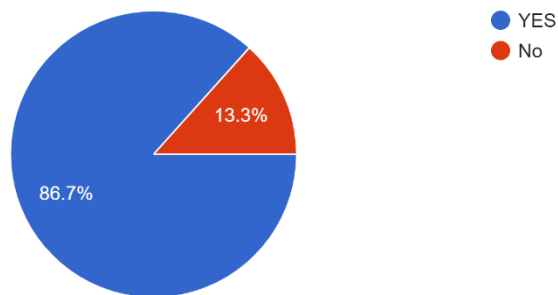
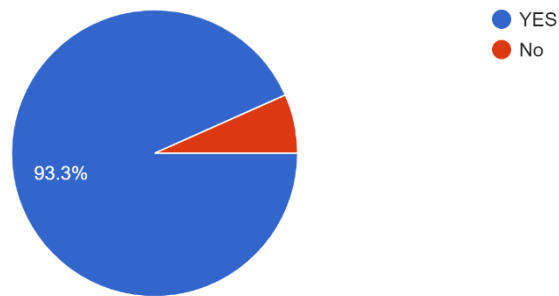3. How to drive to Singapore?Recommend: 1. Waze; 2. Google Maps

30 responses

Appendices
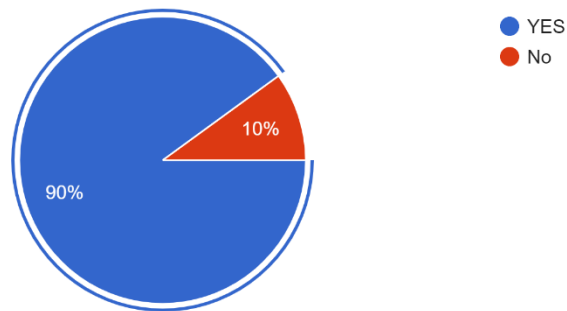
4. How you normally travel around KL?Recommend: 1. Uber
30 responses

YES
No

93.3%

5. What application you normally use to purchase a ticket? Recommend:1. Golden Screen Cinema; 2. TGV Cinema
30 responses

YES
No

10%

90%

BIS (Hons) Information Systems Engineering
Faculty of Information and Communication Technology (Kampar Campus), UTAR

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	TANG ZI CONG
ID Number(s)	15ACB05224
Programme / Course	Bachelor of Information Systems (HONS) Information Systems Engineering
Title of Final Year Project	CONTEXT AWARE APP RECOMMENDATION USING EMAIL SEMANTIC ANALYSIS

Similarity	Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)
Overall similarity index:_____% **Similarity by source** Internet Sources: _____% Publications:_____ % Student Papers:_____ %	
Number of individual sources listed of more than 3% similarity: _____	

Parameters of originality required and limits approved by UTAR are as Follows:
(i)   Overall similarity index is 20% and below, and
(ii) Matching of individual sources listed must be less than 3% each, and
(iii) Matching texts in continuous block must not exceed 8 words
Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8

Note: Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

_____   _____
Signature of Supervisor   Signature of Co-Supervisor

Name: DR. Aun Yichiet   Name:_____

Date: 08/04/2019   Date: _____

UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY (KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	15ACB05224
Student Name	TANG ZI CONG
Supervisor Name	DR. AUN YICHIET

TICK (√)	**DOCUMENT ITEMS** Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
	Front Cover
	Signed Report Status Declaration Form
	Title Page
	Signed form of the Declaration of Originality
	Acknowledgement
	Abstract
	Table of Contents
	List of Figures (if applicable)
	List of Tables (if applicable)
	List of Symbols (if applicable)
	List of Abbreviations (if applicable)
	Chapters / Content
	Bibliography (or References)
	All references in bibliography are cited in the thesis, especially in the chapter of literature review
	Appendices (if applicable)
	Poster
	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report. _____ (Signature of Student) Date:	Supervisor verification. Report with incorrect format can get 5 mark (1 grade) reduction. _____ (Signature of Supervisor) Date: