STANDALONE APPLICATION OF QUESTION CLASSIFICATION IN HIGHER EDUCATION INSTITUTIONS

BY

TAN HOW KIT

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION SYSTEMS (HONS)

BUSINESS INFORMATION SYSTEMS

Faculty of Information and Communication Technology
(Perak Campus)

JANUARY 2018

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: _____
_____
_____

Academic Session: _____

I _____

(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:

_____
_____
_____ _____
_____ Supervisor's name

Date: _____ **Date**:_____

STANDALONE APPLICATION OF QUESTION CLASSIFICATION IN HIGHER EDUCATION INSTITUTIONS

By

TAN HOW KIT

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION SYSTEMS (HONS)

BUSINESS INFORMATION SYSTEMS

Faculty of Information and Communication Technology
(Perak Campus)

JANUARY 2018

DECLARATION OF ORIGINALITY

I declare that this report entitled "**STANDALONE APPLICATION OF QUESTION CLASSIFICATION IN HIGHER EDUCATION INSTITUTIONS**" 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 : _____

Date : _____

ACKNOWLEDGEMENTS

I am really grateful because I was engaged and managed to complete this project proposal. I would like to thanks and appreciate to my supervisor, Mr Anbuselvan a/l Sangodiah who has given me an opportunity to engage in this study. The patience and advises given by him has created a motivation for me to accomplish this study. Finally, I would like to take this opportunity to thanks my family members and friends for continuous support and encouragement throughout the entire process.

ABSTRACT

Assessing students through examination is always a universal test practiced in educational institutions. Designing a good examination question will help educators to better evaluate the understanding level of students. Classifying questions into Bloom Taxonomy categories will improve the overall quality of assessment as Bloom Taxonomy was invented to promote higher levels of thinking in education and to measure cognitive level of learners. However, mistakes might be made when labeling or classifying exam questions into category of BT work was done manually and the educators need a strong understanding and experience in Bloom Taxonomy in order to do the classification task in exam questions. This project proposed a question classification tool which will categorize exam questions into appropriate category of Bloom Taxonomy automatically using machine learning. Besides that, natural language processing techniques such as tokenization, stop-word removal, stemming and lemmatization played important role in text pre-processing stage before machine learning was used. It was used to transform the input questions into computer understandable structure to facilitate later processing. In this project, supervised machine-learning model (Support Vector Machine) was adopted by training it with a data set consisting of questions predefined with labels or categories of BT to classify the unseen exam questions into appropriate level or category of BT. For comparison purposes in terms of accuracy in classifying questions, other machine learning models such as Neural Network, Naive Bayes, Decision Tree had been used to compared with SVM. In summary, the deliverable of the project will benefit educators in labeling exam questions in accordance to BT automatically in higher education institutions.

Keywords: Question Classification, Bloom's Taxonomy, Educational Data Mining, Natural Language Processing, Machine Learning

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

DM	Data Mining
NLP	Natural Language Processing
BT	Bloom's Taxonomy
EDM	Educational Data Mining
KDD	Knowledge Discovery in Database
QA	Question Answering
IR	Information Retrieval
SVM	Support Vector Machine
FS	Feature Selection
NB	Naïve Bayes
KNN	K-Nearest Neighbor
DT	Decision Tree
BoW	Bag-of-Words
CSV	Comma-Separated Values
ML	Machine Learning

Chapter 1: Introduction

1.1 Problem Statement

In the conventional educational model, people often only studied and learned the theories of a subject without giving much understanding. This is because the questions in examination usually are theory-based and therefore students only need to memorize those theories in order to pass a subject. Understanding the cognitive domain from Bloom Taxonomy is important because it can help educator to well design and classify the examination or practical questions in order to enhance the overall learning quality. However, there is a problem of inefficient in manual classification. Despite classifying questions manually could be effective; inefficiency of manual classification is occurring because the manual process is quite time-consuming and can lead to mistakes being made (Mitchell 2014). This issue has negatively impacted on those who may want to use the classified information due to the questions are unable to process on time. A possible cause of this issue is people will slow down the process when they get tired. Thus, a study which investigates automated classification technique could remedy this situation.

Besides, there is also a problem of inconsistency of labeling in classifying questions. Despite human can make judgment better than a machine does, inaccuracy of results in classifying questions manually is still occurring due to the deteriorating quality of decisions made by individuals after a long period decision making (Tierney 2011). This issue has negatively impacted on those individuals that may want to use the distilled information because of the inaccuracy of labeling result. A possible cause of this issue probably is people tend to make mistake when they get tired. Thus, a study which investigates inconsistency of labeling by using classification technique could remedy this situation.

1.2 Background Information

Data mining, also known as Knowledge Discovery in Database, is a practice that analyzing a large amount of data among huge data repository and then interprets them into useful information associated with the techniques between artificial intelligence, machine learning, statistics, and database systems (Meseguer et al. 2015). According to Bagga & Singh (2012), data mining applications have been extensively applied in field of Medical science to understand the mapping relations between human DNA sequence variation and disease susceptibility; Financial Data Analysis to enable people to make better decisions according to market analysis; Retail Data Mining to identifying customer behavior; Telecommunication Data Mining to spot patterns, deceptive actions and helps to maximize the resources usage in order to increase the overall services condition. Figure 1.1.1 shows the trend of data mining from years of 2013 to 2018. According to Google Trends, the popularity of data mining is always above average and sometimes reaching a peak value around the world.

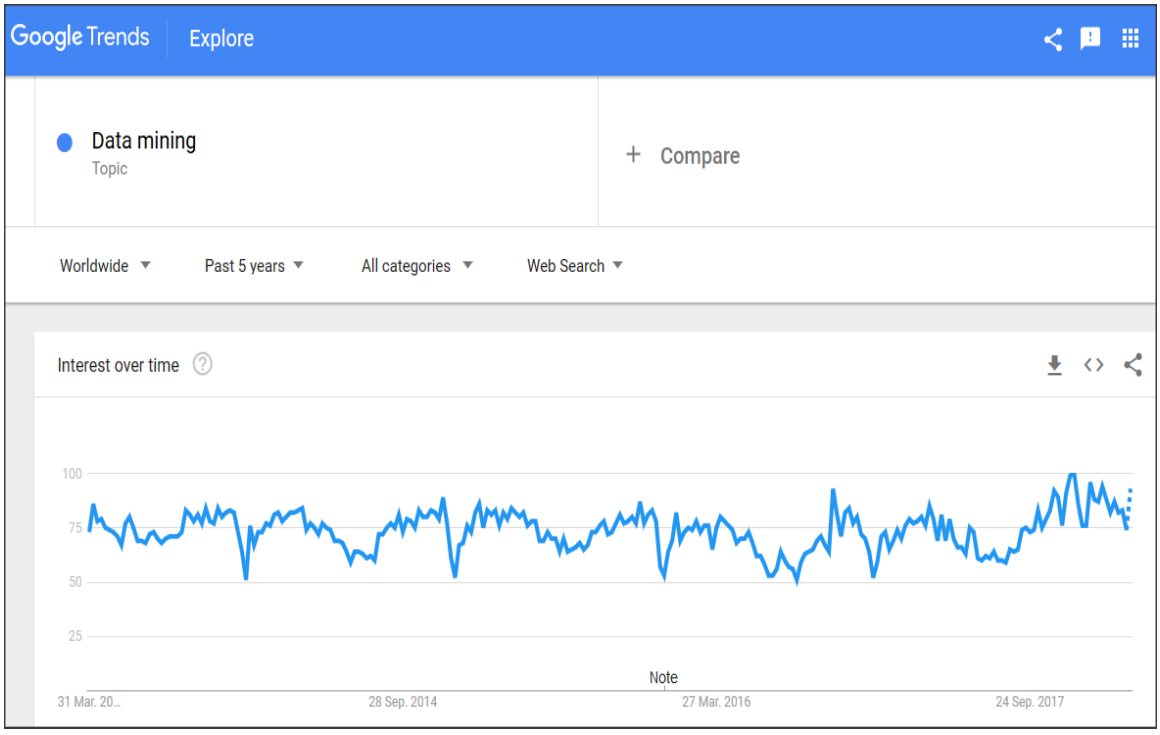


Figure 1.5.1 Data Mining's Trend between years 2013 to 2018 (Google Trend, n.d.)

Educational Data Mining will be the focal point in this paper. Seal, Marzak & Behja (2013) says that educational data mining as a rising discipline, focusing on establishing new practices for searching the exclusive and huge data that derive from educational context to develop an improved understanding of scholars and the pattern which they study in. According to Baker (2010), educational data mining are disparate from others knowledge discovery in database methods because it needs to clearly utilize the numerous degree of purposeful hierarchy in educational data. Educational Data Mining divided into following general categories: relationship mining, discovery with models, prediction, clustering, and distillation of data for human judgment (Baker 2010). In the area of distillation of data for human judgment, data is distilled for two key objectives which are identification and classification (Baker 2010). For classification, data is distilled to support the development of prediction model. In previous researches, Xu and Reynolds (2012) had implement data mining techniques to evaluate students' written responses to a teacher leadership dilemma. Through this exercise, they were able to provide insight to understand the linguistic structure based on the unique concepts generated from the responses. Besides, He (2013) applied both data mining and text mining approach to analyze the online query and also conversations that documented by a live video streaming system. This exercise found the likeness along with differences in the students' patterns and participation's subject between the online query and online conversations. Moreover, Akcapinar (2015) used text mining technique to automated feedback in order to reduce the plagiarism behavior in the online assignment. The study was conducted by 59 participations who were participated the Computer Hardware class and the result has successfully reduced the ratio of plagiarism.

Text classification is one of the popular topics for researchers nowadays because peoples are living in a big data world. Text classification is a subfield of text mining which classifying text records into one or more predefined category (Manning & Schutze 1999). With good use of text classification in education, the learning and teaching process will be analyzed and therefore bring an improvement in the overall academic performance. Moreover, question classification is a process of analyzing a question and the system will automatically label the question to the predefined answer type. According to Sangodiah, Ahmad & Ahmad (2014), question classification is more challenging as contrasted to document classification in getting a satisfactory

accuracy. This is due to very little information occurred in question classification and that might not be sufficient to effectively classify the questions as opposed to document classification. Question classification system was also treated as a component of Question Answering systems and Information Retrieval system too (Metzler & Croft 2004). There are some past researches of question classification in educational field. Omar et al. (2012) had used rule-based approach and proposed an automated analysis of examination questions and then categories to the correct level regarding Bloom Taxonomy by using Natural Language Processing approaches to determine important keywords in the questions. Similar work done by Haris (2015) has using hybrid ability of rules and statistical approach to determine the divisions of questions regarding Bloom Taxonomy. The combination of rules and N-gram had performed well to categorize the questions as these methods take advantages to overcome each of their weakness by each other's strength.

Bloom taxonomy was developed in years 1956, by a group of the educational psychologist for the purpose of classifying the level of learning and understanding according to the levels of taxonomy (Bloom et al. 1956). According to Bloom et al. (1956), there are 3 domains in Bloom taxonomy: Cognitive, Affective, and also Psychomotor. Question classification incorporating to the Cognitive domain of Bloom taxonomy is able to classify the question into the 6 categories which are Knowledge, Comprehension, Application, Analysis, Synthesis and also Evaluation (Bloom, et al. 1956).

To develop a questions classification tool, this study compares several classification techniques in question classification in order to find an appropriate classification technique to classify question which can achieve a reasonable accuracy.

1.3 Project Objectives

I. To investigate existing question classification work in the context of BT.

In this study, several techniques and accuracies of question classifiers are studied in order to get a better insight and idea to develop a question classification tool with the desired accuracy by assessing the past experiences made by the researchers.

II. To use appropriate NLP techniques and feature type to build question classifier model

Natural language processing techniques such as tokenization, stop-word removal, stemming and lemmatization played important role in text pre-processing stage before machine learning was used. It was used to transform the input questions into computer understandable structure to facilitate later processing. Also, BoW was use as a feature extraction technique in this project.

III. To develop a question classification tool to classify questions in accordance to BT

The primary objective of this study is to develop a question classification tool that will analyzes the text-based questions and then categorizes them based on BT cognitive domain and comes with at least 70% of accuracy by using NLP and ML techniques. Classifying the examination or practical questions set based on Bloom Taxonomy's cognitive domain could help educators better design the questions and improve the overall learning objectives of students effectively as Bloom Taxonomy was developed to promote higher levels of thinking in education rather than just rote learning (Clark 1999).

IV. To evaluate the question classification tool

In this project, supervised machine-learning model (Support Vector Machine) was adopted by training it with a data set consisting of questions predefined with labels or categories of BT to classify the unseen exam questions into appropriate level or category of BT. For comparison purposes in terms of accuracy in classifying questions, other machine learning models such as Neural Network, Naive Bayes, Decision Tree had been used to compared with SVM.

1.4 Project Scope

In this project, a question classification tool was constructed and developed. Support Vector Machine, Decision Trees, Naïve Bayes, and k-nearest neighbors classification techniques were studied and compared in order to find an appropriate classification technique to classify questions with the desired accuracy. Text-based documents such as practical, tutorial, or examination questions will be used as input to analyze and classify in this study. Besides, this question classifier is primarily focusing on business domain with various courses and subjects questions. Moreover, this question classification tool is designed for educators in order to allow them to identify the questions based on the cognitive domain from Bloom Taxonomy and therefore better design the practical or examination questions for students in order to achieve an overall improved academic performance. Furthermore, this question classification tool is built using Python programming and based on Text Mining, Natural Language Processing, and Question Classification in order to classify questions into predefined groups. Nevertheless, all of the questions are categories based on Bloom Taxonomy. In this project, a questions classification tool was delivered with the ability of question categorizing according to Bloom's Taxonomy Cognitive domain.

1.5 Impact and Contribution

This project will propose a question classification tool which can benefit educators in the educational environment. With this question classification tool, educators can easily evaluate the unstructured examination or practical questions and then better design the questions based on Bloom Taxonomy's cognitive domain for the purpose of improving the overall learning objectives of students. With the designed questions, educators are able to evaluate the understanding level of students and therefore can design and provide appropriate teaching methods to each of them in order to improve overall teaching and learning process effectively and efficiently.

1.5.1 Significant

This project is designed to develop a question classification tool with an automating processing ability. Classifying a set of questions manually will be inefficient and time-consuming because people need take some time to handle a huge workload (Mitchell 2014). Besides, inaccuracy result in classifying questions manually is also occurring due to deterioration quality of decisions made by an individual after a long period of decision making (Tierney 2011). Hence, this project will introduce an automated question classification tool with the desired accuracy to overcome the problems that may happen in manual classification. Besides, the question classification tool might improve the overall learning and teaching quality in education environment as educators can well-structured the examination questions accordance to BT level and to better access the understanding level of students.

Chapter 2: Literature Review

2.1 Data Mining

Data mining is an interdisciplinary field in between of artificial intelligence, machine learning, statistics, and database systems (Meseguer et al. 2015). Data mining is a technique of discovering and uncovering patterns in large datasets and then interprets them into an understandable structure. According to Fayyad et al. (1997), data mining is also a data analysis and discovery algorithms step in knowledge discovery in the database - KDD. Figure 2.1 shows 6 common types of data mining methods with their corresponding short description.

Types	Description
Classification	Learning function that classifies a fata item into one of several predefined classes. (Weiss & Kulikowski 1991).
Regression	Learning function that maps a data item to a real-valued prediction variable.
Clustering	Common descriptive task of grouping a set of objects based on their similarity characteristics.
Summarization	Involves methods for finding a compact description for a subset of data.
Dependency modeling	Finding a model that describes significant dependencies between variables.
Change and deviation detection	Detect the most significant changes in the data from previous measured values (Berndt & Clifford 1996).

Figure 2.1 Common types of Data mining methods (Fayyad et al. 1997)

2.1.1 Text Mining

According to Radovanovic & Ivanovic (2008), text mining has a strong connection with natural language processing, data mining, machine learning, information retrieval, and knowledge management. Text mining is derived from data mining and it discovers and extracts helpful knowledge from large-scale text data through analyzing and searching for impressive patterns (Feldman & Sanger 2007). One of the popular text mining techniques is text categorization or text classification. Text classification is a technique of classifying a text document into predefined categories with some similar keyword or patterns. Before text classification works, the document needs to transform first into a representation suitable for learning algorithm and classification task (Doleck et al. 2015). Aggarwal & Zhai (2012) said there are many existing techniques have been designed for text classification such as Decision Trees, Pattern (Rule)-based Classifiers, Support Vector Machine Classifiers, Neural Network Classifiers, Bayesian (Generative) Classifiers, et cetera (Dang et al. 2016). Figure 2.1.1 shows the process of text classification.

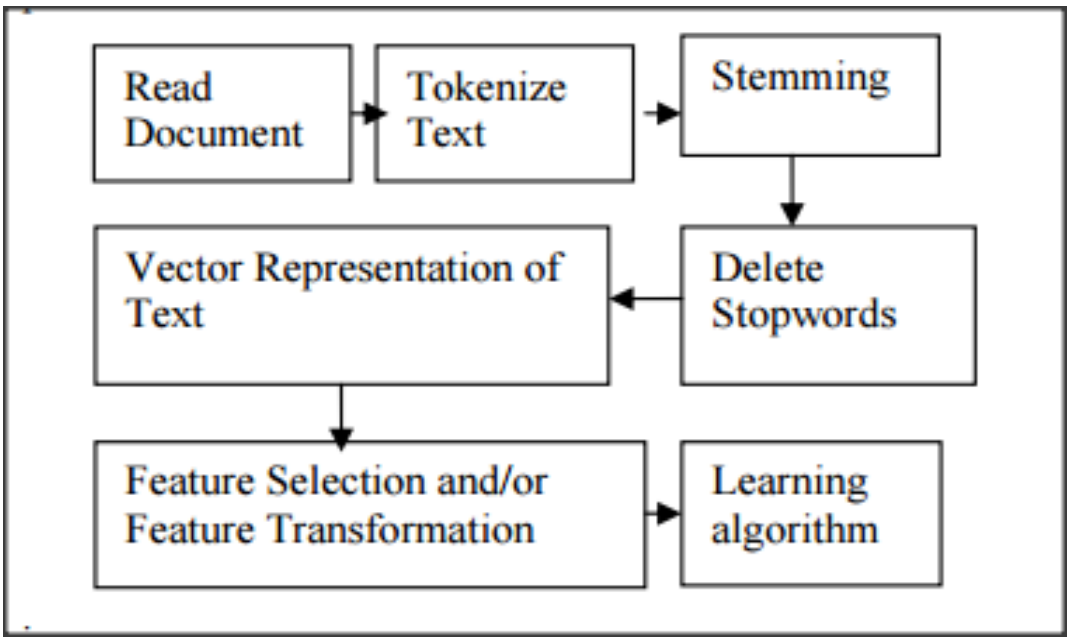


Figure 2.1.1 Process of Text Classification (Radovanovic & Ivanovic 2008)

2.2 Bloom Taxonomy

Bloom taxonomy was developed in years 1956 by a group of educational psychologist to promote higher levels of thinking in education rather than just rote learning (Clark 1999). The objective of Bloom taxonomy is to classify the level of understanding according to its corresponding level of taxonomy. According to Bloom et al. (1956), there are 3 domains in Bloom taxonomy, which are Cognitive, Affective, and Psychomotor. Cognitive domain deals with knowledge or mental skills, Affective domain deals with internal feelings and emotions, and Psychomotor domain deal with the manipulative or motor-skill area. Cognitive domain will be the main concerned in this study because it deals with knowledge area which is closely related to question classifications. There are 6 major classes of the cognitive domain in Bloom Taxonomy, starting from the simplest Knowledge, followed by Comprehension, Application, Analysis, Synthesis, and finally the complex one, Evaluation (Bloom et al. 1956). In the year 2001, the model of the taxonomy was revised by Anderson and a group of cognitive psychologist. According to Anderson et al. (2001), the revised model modified the name in 6 major classes from noun to verb forms and rearranged the sequence of them, which becomes Remembering, follow by Understanding, Applying, Analyzing, Evaluating, and the last, Creating. Figure 2.2.1 shows the figure of comparing the original taxonomy with the revised taxonomy.

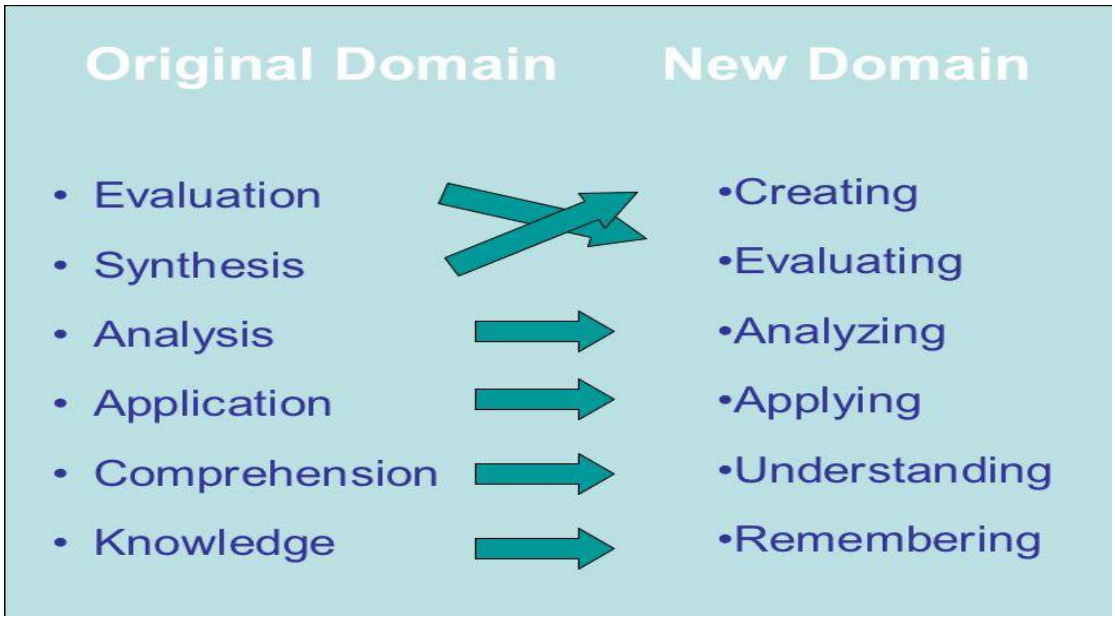


Figure 2.2.1 Compare of original cognitive domain with revised version (Clark 2015)

Figure 2.2.2 shows each of the original categories of cognitive domain starting from low level to high level, and Figure 2.2.3 shows each of the revised version of categories which is starting from low level to high level too, as well as illustrating some keyword and simple examples that can be used as the reference on question classifications.

Category	Keywords	Example
Knowledge (Recall information)	Arrange, defines, describes, identifies, labels, lists, knows, names, matches, recalls, recognizes, selects, states	Define a term.
Comprehension (Understand the meaning)	Converts, distinguish, estimate, explain, translates, summarizes, interpret, defend, infer	Explain the steps for performing a complicate task.
Application (Apply what was learned in a new situation)	Apply, compute, construct, demonstrates, discovers, operate, predict, solve, use	Use a manual to calculate an employee's vacation time.
Analysis (Distinguish, classifies, and relates the assumptions of a statement)	Analyze, compare, diagrams, differentiate, distinguish, identify, illustrate, discriminate	Gather information from a department and select the required tasks for training.
Synthesis (Put parts together to form a new whole)	Categorize, compile, combine, design, compose, revise, create, generate, rearrange	Design a machine to perform specific task.
Evaluation (Ability to make judgment about the value of ideas)	Appraise, compare, conclude, defend, evaluate, justify, criticize, interpret, summarize	Explain and justify a new budget.

Figure 2.2.2 Six categories of Cognitive Domain - Original (Clark 1999)

Category	Keywords	Example
Remembering (Recall previous learned material)	Define, describe, know, label, list, state, select	Recite safety rules.
Understanding (Comprehend of instructions and problems)	Convert, defend, distinguish, estimate, explain, infer, interpret	Explain the step for performing a complicate task.
Applying (Apply a concept in a new situation)	Apply, compute, construct, discover, solve, predict, demonstrate, produces	Use a manual to calculate an employee's vacation time.
Analyzing (Distinguish, classifies, and relates the assumptions of a statement)	Analyze, compare, contrast, diagrams, differentiate, distinguish, discriminate, identify	Gather information from a department and select the required tasks for training.
Evaluating (Ability to make judgment about the value of ideas)	Appraise, compare, conclude, defend, describe, evaluate, explain, justify, summarize, relate	Explain and justify a new budget.
Creating (Build a structure or pattern from various elements.)	Categorize, create, generate, rearrange, reconstruct, reorganize, revise, rewrite, compose	Design a machine to perform specific task.

Table 2.2.3 Six categories of revised Cognitive Domain (Clark 2015)

Bloom Taxonomy had been widely spread in the educational environment to improve the quality of learning. In past research, bloom taxonomy had applied in a computer science class for testing and evaluating the programming questions (Scott 2003); Comparing the difficulty level of programming and networking course by computing Bloom Rating (Oliver et al. 2004); Applied the revised Bloom's taxonomy to classify assignment together with test questions in an Introduction to Linux subject (Johnson et al. 2012); and others field to improve the learning experiences.

2.3 Natural Language Processing

Natural Language Processing (NLP), is a technique of automated or semi-automated processing of the human language (Copestake 2004). NLP is a subdomain of artificial intelligence and linguistics that focuses on making machines comprehend the statements or words drafted in the human language (Chopra et al. 2013). There are several forms of NLP in used today included dialog or speech systems, document classification, search and retrieval, textual analysis, question answering and information retrieval, etcetera (Eslick & Liu 2005). According to Chorpa et al. (2013), NLP involved 5 general steps starting from Morphological and Lexical Analysis, followed by Syntactic Analysis, Semantic Analysis, Discourse Integration, and finally, Pragmatic Analysis. Table 2.3.1 shows the steps involved in NLP with their corresponding descriptions.

Phases	**Description**
Morphological and Lexical Analysis	Analyze and identify the structure of words and then divide them into words, sentences, and paragraphs.
Syntactic Analysis (Parsing)	Analyze of words in grammatical sentences and transform them to shows the relationship among the words.
Semantic Analysis	Draw the exact meaning from text and check for its meaningfulness by mapping syntactic structure and objects in task domain.
Discourse Integration	Meaning of a sentences depends on the precede meaning of sentences, and also brings the meaning for the following sentences.
Pragmatic Analysis	Words will be re-interpreted on it real meaning which require real world knowledge because it involves deriving those aspects of language.

Table 2.3.1 Five general steps in NLP (Chopra et al. 2013)

2.4 Feature Selection

In machine learning, feature selection also known as variable selection or attributes selection. Feature selection is a technique about automated picking of attributes within the given data. For example, columns in tabular form, which are remarkably important for the predictive modeling issue that is working on (Brownlee 2014). According to Brownlee (2014), feature selection is likely performed as a filter; it attempts to decrease the number of attributes in the dataset by including and excluding the attributes that exist in the data without changing them. Guyon & Elisseeff (2003) said feature selection methods aid to improve the prediction performance of the predictors by selecting the features that will mostly achieve a satisfactory or improved accuracy meanwhile requiring minimal data. Moreover, feature selection technique can be applied to recognize and eliminate undesirable, unnecessary, and duplicated attributes against the data that not only have no significance to the predictive model accuracy result but also may reduce the accuracy of the prediction model.

2.4.1 Bag-of-Words Model

Most of the machine learning techniques and algorithms required well defined fixed-length inputs and outputs, therefore there is a problem existed which was modeling the raw text that is cluttered. Machine learning algorithms are unable to work with raw text straightly; the raw text must be transformed into number, or precisely, vectors of numbers (Brownlee 2017). According to Brownlee (2017), the process of transforming those raw text-based data into number also known as feature extraction or feature encoding. Moreover, one of the most well-known and simple techniques of feature extraction that deal with text-based documents be known as Bag-of-Words model. Bag-of-Words model is a technique of eliciting features from text data in order to facilitate the process in modeling, for example with machine learning algorithms. The process of Bag-of-Words is pretty simple and adaptable, and may also be applied in countless of methods for extracting features against text-based documents. According to Goldberg (2017), Bag-of-Words is a denotation of text which defines the existence of words inside a document that involved two elements. The first is a vocabulary of recognized words, and the second is a gauge of the

existence of recognized words. It's known as "bag" of words since whichever information regard to the arrangement or organization of words within the text document is abandoned. BoW model is only focused on if known words exist in the text document, not the position of words in the text document. BoW model may be either simple or complicated, the complicatedness falls both in determining how to construct the vocabulary of recognized words (or tokens), and how to weigh the existence of recognized words.

Below shows the example of modeling two simple text-based documents using bag-of-words.

1. Jack likes to play badminton. Jackson likes badminton too.
2. Jack also likes to play basketball.

Based on the above example, a list was composed as follow for each sentence. The sentences are chopped into a single word, also known as tokens. In this way, all word or token is known as a "gram". Building a vocabulary of single-word is called unigram.

1. "Jack", "likes", "to", "play", "badminton", "Jackson", "likes", "badminton", "too"
2. "Jack", "also", "likes", "to", "play", "basketball".

The next step is to construct a list of every word in the model vocabulary, and the list should not contain any duplicated word or token. The unique words extracted from the example are:

"Jack", "likes", "to", "play", "badminton", "Jackson", "too", "also", "basketball"

Afterward is to score the words in each sentence, also known as term frequency or term weighting. The purpose of this is to convert each sentence of free text in the direction of a vector that may work as input or output against a machine learning model. The easiest scoring approach is to record the existence of words as a Boolean form, 0 for missing, 1 for existing. The scoring of example would look like follows:

1. "**Jack**"= 1, "**likes**"= 2, "**to**"= 1, "**play**"= 1, "**badminton**"= 2, "**Jackson**"= 1, "**too**"= 1, "**also**"= 0, "**basketball**"= 0.

2. "**Jack**"= 1, "**likes**"= 1, "**to**"= 1, "**play**"= 1, "**badminton**"= 0, "**Jackson**"= 0, "**too**"= 0, "**also**"= 1, "**basketball**"= 1.

And last for a binary vector, which would look likes follows:

1. Jack likes to play badminton. Jackson likes badminton too.
   = [1, 2, 1, 1, 2, 1, 1, 0, 0]

2. Jack also likes to play basketball.
   = [1, 1, 1, 1, 0, 0, 0, 1, 1]

2.5 Question Classification

Question classification is a process of analyzing a question and the system will automatically label the question based on the proposed answer group. According to Panicket et al. (2012), earlier approaches for automatic document classifier consisted of manually building. The advantages of these approaches were it required rules manually defined and when the classifier is transplanted to a completely different domain, the domain expert needs to intervene and the work has to be done in its entirety. Machine Learning approach is currently applied to overcome the defect of rule-based classification. In this approach, a set of pre-defined questions is fed to the classifier. This action will act as the training example for the classifier. Based on these examples, the classifier will classify the future questions. There is much existing text classifier approaches included probabilistic, decision tree, decision rule, regression-based, neural network, support vector machine, and etcetera (Alpaydin 2010).

Although question classification functions in the same method as document classification, question classification is more challenging as opposed to document classification in getting a satisfactory accuracy because there exist only a little information in question classification, and that might not be sufficient to effectively classify the questions as compared to document classification (Sangodiah, Ahmad & Ahmad 2014). According to Metzler & Croft (2004), question classification was employed as part of question answering and information retrieval systems. Question answering is a process of fetching answers to the questions composed in natural language from a group of documents, where the answers are usually a limited segment of text extract from the corpus. And information retrieval is a task of returning relevant documents to a particular natural language query (Jurafsky & Martin 2000).

There are some past researches of classifying question according to Bloom taxonomy. Omar et al. (2012) proposed an automated analysis of exam question by using rule-based approach. The study was using written final examination questions in Programming subjects and fed with 70 training examination questions set and 30 questions to test it. First of all, they implement stopwords removal to process the questions to increase the readability. Next, each processed word will be tagged by a tagger to determine significant keywords which might be important in categorizing. After the cleaning process, some rules will be implemented to determine the questions'

category. However, the accuracy of the developed rules was not good due to most of the training questions start with a verb. Based on Bloom's Taxonomy, the verb 'Write' can be categorized into Knowledge and Synthesis. Hence, they used category weighting which assigns weight to the conflicting categories to increase the accuracy of categorizing. In this case, questions' classifying from subject matter experts (SMEs) is used to calculate the assigned weight.

Similar work has done by Haris & Omar (2015) by using the hybrid ability of rules and statistical approach. Haris & Omar (2015) has developed 64 rules and covers 6 categories of Bloom's Taxonomy in the study. The developed rules will try to determine the syntactic structure from the input questions and then categorize it into suitable categories. If all the defined rules do not match with the syntactic structure, N-gram will handle the categorization process by calculating the frequency and the degree of similarity in test questions and then compare to the training set. The combination of rules and statistical approach has performed well in question categorization because both methods take the advantages of own to overcome each other's weakness.

Osman & Alattab (2013) has done a study of analyzing classroom questions according to Bloom Taxonomy Cognitive Level. In this study, they perform four feature selection approaches which are Term Frequency, Mutual Information, Information Gain, and Chi-Square to filter out the important words from the questions. Besides, they also do implement four machine learning techniques which are k-Nearest Neighbour, Naïve Bayes, Support Vector Machine, and Rocchio Algorithm to train the system to classify the questions. The result of this study shows that Rocchio Algorithm has achieved a better accuracy, particularly work with Information Gain feature selection approach, mainly because it can utilize the selected terms better than others. Support Vector Machine has a comparable performance with Rocchio Algorithm in particular when working with Term Frequency and Chi-Square feature selection approach. Moreover, Naïve Bayes and k-Nearest Neighbour have collaborated well with Term Frequency approach to attain a satisfactory result. In a nutshell, the use of the combination of rule and statistical approach seems better to achieve a higher accuracy in questions classification.

Chapter 3: Methodology

3.1 Training Classifier

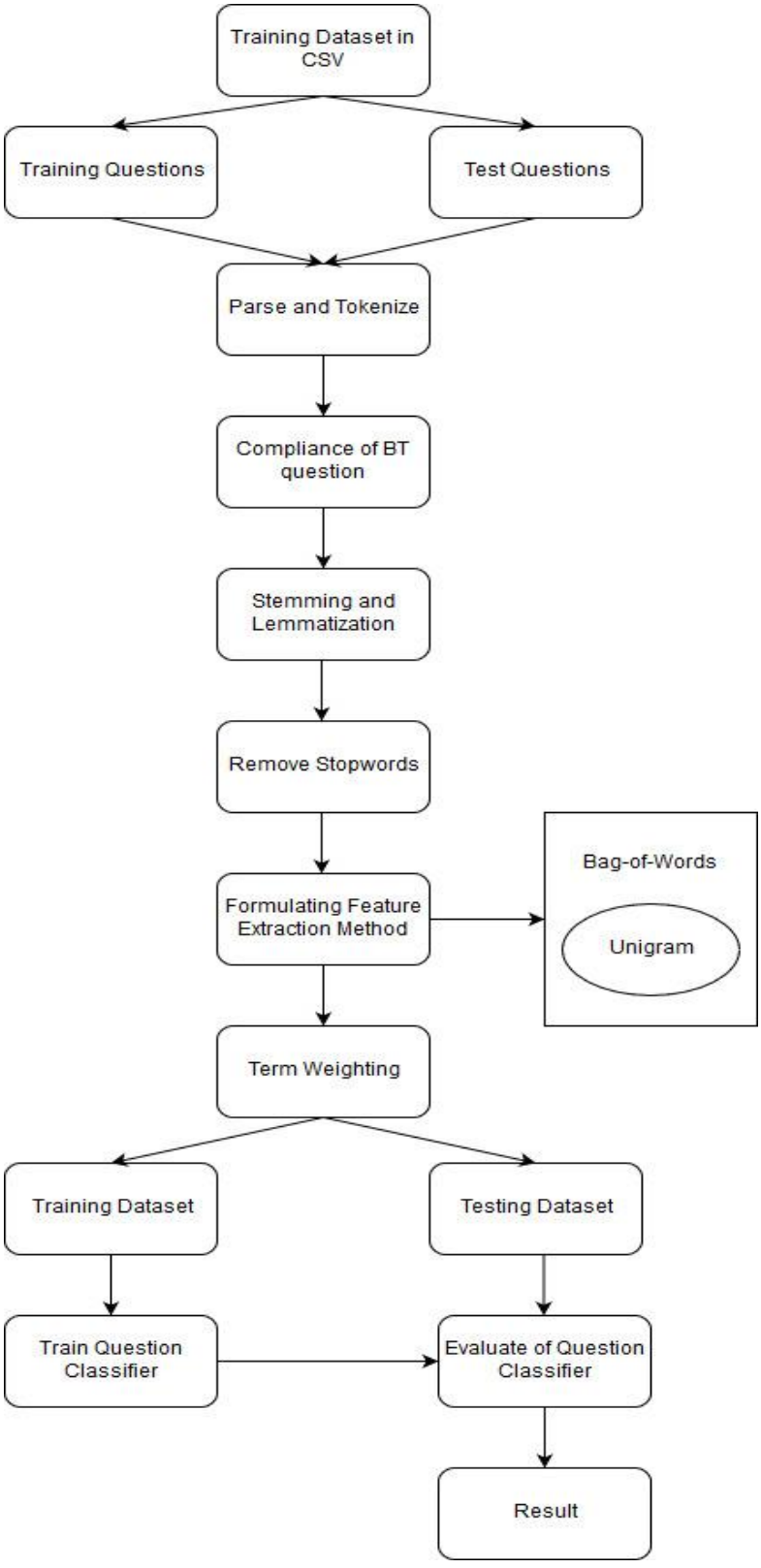


Figure 3.1.1 Training Classifier

Dataset

A set of training data that stored in CSV file format will first be fed into the system in order to let the machine learn the patterns and also train the classifier. A total number of 163 questions along with its predefined answer that regarding on business domain with various courses and subjects will be used in this project and train the classifier. All the answers in training dataset has been labeled and verified by education experts.

Text-Preprocessing

Then, text preprocessing will take part to deal with the training dataset and split them into training questions for machine learning and also testing questions for evaluating the machine accuracy. Next, the program will transform both of the training and also testing questions into machine understandable format to facilitate later works. In the first stage of text preprocessing, the program will remove all the punctuation characters that exist in the training dataset because those punctuation characters do not have any value for the later process but will decrease the classifiers' accuracy. In the second stage, the program will tokenize and chopped every question into a separate pieces or words and then store into a list. Next, stemming will kick in to transform back every single word into their root form. For example, "fishing", "fished", "fisher" will be converted into their root form "fish" after stemming operate. Then, Lemmatization will be used to make sure there are only nouns and verbs exist in the features set. And last, stopword removal will be used to expel those unnecessary words such as "an", "and", "so", "that" and so on within the features set.

Feature Extraction

After that, feature extraction method, Bag-of-Words model was used to create the features set. A list was composed for every training questions that stored each word or tokens that have been processed in the text-preprocessing stage. The approach that has been used to create the vocabulary is also known as unigram. Next, a list of all the words in the model vocabulary was formulated. The list was then converted into a set

in order to make sure there have no duplicated word or token within the list (every word or token is unique within the list).

Term-Weighting

Afterwards, the machine will score the words in each training questions with the features set. The purpose of this is to turn each sentences of free text into a vector that can work as input or output for a machine learning model. The scoring approach that has been adopted is to mark the presence of words as a Boolean value, 0 for absent, 1 for present.

Features Set / Questions	above	define	accept	commerce	Bloom's Taxonomy Level
Define Electronic Commerce	0	1	0	1	**REMEMBER**

Table 3.1.2 Sample of Processed Training Data

Evaluate the Classifier

In this stage, four different classification techniques such as Support Vector Machine, Naïve Bayes, K-Nearest Neighbor, and Decision Tree were chosen to do a comparison with their accuracy. 10-fold cross validation method was used to evaluate the performance and accuracy among the classifiers.

Storing Classifier

And last, the trained classifier will be stored into a ".pickle" file so users no need to train the classifier every time when they run the question classification tool and therefore saving the processing time and computer resources.

3.2 System Design

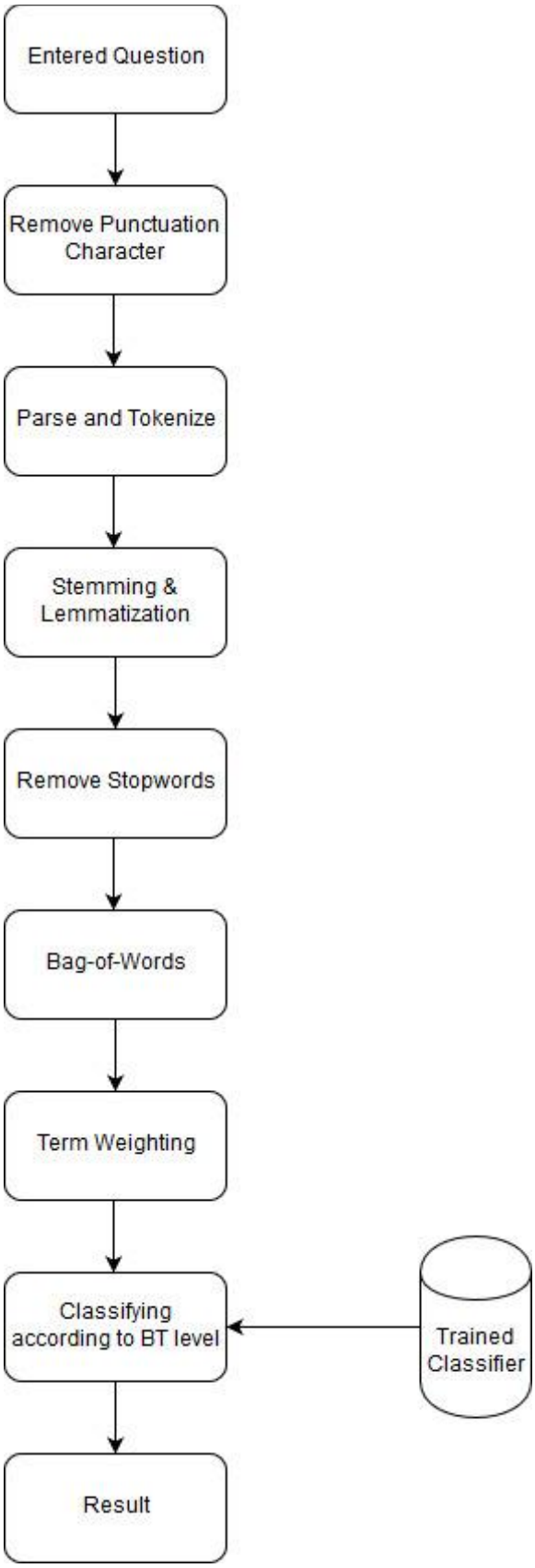


Figure 3.2.1 – Question Classification Tool Design

Figure 3.2.1 shows the system design of question classification tool in this study. First of all, the program will prompt and receive an input from users. Next, the question classification tool will perform text pre-processing function in order to convert the question given by users into a machine-understandable format to facilitate the later process. Several text-pre-processing approaches will take part in here such as punctuation character removal, tokenization, stemming, lemmatization, and also stopword removal. After that, the machine will perform BoW to compare the processed question to the stored features set, and remove those words that do not exist in the features set. Then, the processed words or question will do the term weighting to compare the features set and determine whether how many processed word from the questions appear in the features set. Afterward, the trained and stored classifier in the earlier stage will be call out and input the processed question into it. And last, the classifier will automatically label and determine whether the question that given by users is belong to which category of Bloom's Taxonomy cognitive domain.

Features Set / Questions	above	define	accept	commerce	Bloom's Taxonomy Level
Define Electronic Commerce	0	1	0	1	

Table 3.2.2 Sample of BoW and Term Weighting

Chapter 4: Evaluation

4.1 10-fold Cross-Validation Method

Support Vector Machine

Multiple Scores when k =	k-List	Accuracy
3	[0.64285714, 0.51851852, 0.62264151]	0.595 (+/- 0.11)
4	[0.56818182, 0.70731707, 0.5, 0.68421053]	0.615 (+/- 0.17)
5	[0.57142857, 0.73529412, 0.60606061, 0.5483871, 0.73333333]	0.639 (+/- 0.16)
6	[0.56666667, 0.75, 0.64285714, 0.57692308, 0.69230769, 0.64]	0.645 (+/- 0.13)
7	[0.57692308, 0.69230769, 0.69565217, 0.65217391, 0.56521739, 0.52380952, 0.80952381]	0.645 (+-/ 0.18)
8	[0.60869565, 0.5, 0.77272727, 0.71428571, 0.52380952, 0.47368421, 0.61111111, 0.82352941]	0.628 (+/- 0.24)
9	[0.59090909, 0.45, 0.9, 0.66666667, 0.66666667, 0.52941176, 0.625, 0.625, 0.8125]	0.652 (+/- 0.26)
10	[0.68421053, 0.38888889, 0.77777778, 0.625, 0.6875, 0.75, 0.625, 0.66666667, 0.53333333, 0.78571429]	0.652 (+/- 0.23)

Table 4.1.1 10-fold Cross-Validation (Support Vector Machine)

Decision Tree

Multiple Scores when k =	k-List	Accuracy
3	[0.51785714, 0.59259259, 0.60377358]	0.571 (+/- 0.08)
4	[0.63636364, 0.80487805 0.65, 0.60526316]	0.674 (+/- 0.15)
5	[0.68571429, 0.70588235, 0.66666667, 0.67741935, 0.6]	0.667 (+/- 0.07)
6	[0.7, 0.71428571, 0.67857143, 0.61538462, 0.80769231, 0.64]	0.693 (+/- 0.12)
7	[0.76923077, 0.65384615, 0.69565217, 0.69565217, 0.73913043, 0.52380952, 0.71428571]	0.685 (+-/ 0.15)
8	[0.7826087, 0.54545455, 0.77272727, 0.57142857, 0.57142857, 0.84210526, 0.38888889, 0.70588235]	0.648 (+/- 0.29)
9	[0.81818182, 0.55, 0.9, 0.72222222, 0.55555556, 0.76470588, 0.6875, 0.4375, 0.75]	0.687 (+/- 0.28)
10	[0.73684211, 0.72222222, 0.83333333, 0.625, 0.6875, 0.75, 0.8125, 0.66666667, 0.53333333, 0.71428571]	0.708 (+/- 0.17)

Table 4.1.2 10-fold Cross-Validation (Decision Tree)

Naïve Bayes

Multiple Scores when k =	k-List	Accuracy
3	[0.55357143, 0.35185185, 0.43396226]	0.446 (+/- 0.17)
4	[0.5, 0.56097561, 0.45, 0.47368421]	0.496 (+/- 0.08)
5	[0.51428571, 0.58823529, 0.60606061, 0.41935484, 0.5]	0.526 (+/- 0.13)
6	[0.53333333, 0.60714286, 0.46428571, 0.5, 0.5, 0.52]	0.521 (+/- 0.09)
7	[0.57692308, 0.57692308, 0.39130435, 0.7826087, 0.43478261, 0.47619048, 0.61904762]	0.551 (+-/ 0.24)
8	[0.65217391, 0.5, 0.63636364, 0.61904762, 0.52380952, 0.42105263, 0.33333333, 0.52941176]	0.527 (+/- 0.21)
9	[0.68181818, 0.45, 0.7, 0.5, 0.77777778, 0.41176471, 0.4375, 0.375, 0.5]	0.537 (+/- 0.27)
10	[0.63157895, 0.44444444, 0.55555556, 0.4375, 0.5, 0.8125, 0.4375, 0.4, 0.46666667, 0.5]	0.519 (+/- 0.23)

Table 4.1.3 10-fold Cross-Validation (Naïve Bayes)

K-Nearest Neighbour

Multiple Scores when k =	k-List	Accuracy
3	[0.58928571, 0.62962963, 0.50943396]	0.576 (+/- 0.10)
4	[0.54545455, 0.68292683, 0.625, 0.60526316]	0.615 (+/- 0.10)
5	[0.62857143, 0.52941176, 0.54545455, 0.61290323, 0.53333333]	0.570 (+/- 0.08)
6	[0.6, 0.60714286, 0.67857143, 0.61538462, 0.69230769, 0.56]	0.626 (+/- 0.09)
7	[0.65384615, 0.53846154, 0.60869565, 0.60869565, 0.65217391, 0.61904762, 0.61904762]	0.614 (+-/ 0.07)
8	[0.73913043, 0.5, 0.63636364, 0.52380952, 0.61904762, 0.57894737, 0.5, 0.64705882]	0.593 (+/- 0.16)
9	[0.72727273, 0.5, 0.65, 0.72222222, 0.44444444, 0.70588235, 0.625, 0.5625, 0.625]	0.618 (+/- 0.19)
10	[0.78947368, 0.55555556, 0.55555556, 0.625, 0.625, 0.5625, 0.75, 0.73333333, 0.6, 0.57142857]	0.637 (+/- 0.17)

Table 4.1.4 10-fold Cross-Validation (K-Nearest Neighbour)

Comparing Result

Classifier	Means in k-List	Accuracy (Final mean value)	Standard Deviation
Support Vector Machine	[0.595, 0.615, 0.639, 0.645, 0.645, 0.628, 0.652, 0.652]	**0.633875**	0.40179751562499993
Decision Tree	[0.571, 0.674, 0.667, 0.693, 0.685, 0.648, 0.687, 0.708]	**0.666625**	0.44438889062500003
Na ïve Bayes	[0.446, 0.496, 0.526, 0.521, 0.551, 0.527, 0.537, 0.519]	**0.515375**	0.26561139062500005
K-Nearest Neighbour	[0.576, 0.615, 0.57, 0.626, 0.614, 0.593, 0.618, 0.637]	**0.6061249999999999**	0.3673875156249999

Table 4.1.5 10-fold Cross-Validation (Comparing Classifiers)

4.2 Evaluate with Real Data type

Support Vector Machine

No.	Questions	Predicted Answer	Correct Answer
1	Apply Porter's five competitive forces analysis to examine the summer job industry for your uncle.	CREATE	**APPLY**
2	Discuss the extent to which Value Chain Analysis can be applied in the logistics sector	**APPLY**	APPLY
3	Information Systems value is determined by the strong relationships among THREE (3) major components. Name them.	**REMEMBER**	REMEMBER
4	Briefly explain any TWO (2) observations of information technology trend using Moore's Law.	**UNDERSTAND**	UNDERSTAND
5	Relate cycle-time reduction to improved performance. Justify your answer.	**EVALUATE**	EVALUATE
6	Prepare a research proposal on a study that you have to conduct on the purchasing behaviour of teenagers in the Klang Valley.	**CREATE**	CREATE
7	Compare FOUR (4) point of views of entrepreneurs with FOUR (4) for managers the way they look at the things.	**ANALYZE**	ANALYZE
8	Define electronic commerce.	UNDERSTAND	**REMEMBER**
9	Define the term project in the context of project management.	**REMEMBER**	REMEMBER
10	Argue critically with relevant examples.	UNDERSTAND	**EVALUATE**
Total Correct Answer		7/10	70%

Table 4.2.1 Support Vector Machine Classifier Result

Decision Tree

No.	Questions	Predicted Answer	Correct Answer
1	Apply Porter's five competitive forces analysis to examine the summer job industry for your uncle.	UNDERSTAND	**APPLY**
2	Discuss the extent to which Value Chain Analysis can be applied in the logistics sector	EVALUATE	**APPLY**
3	Information Systems value is determined by the strong relationships among THREE (3) major components. Name them.	UNDERSTAND	**REMEMBER**
4	Briefly explain any TWO (2) observations of information technology trend using Moore's Law.	**UNDERSTAND**	UNDERSTAND
5	Relate cycle-time reduction to improved performance. Justify your answer.	UNDERSTAND	**EVALUATE**
6	Prepare a research proposal on a study that you have to conduct on the purchasing behaviour of teenagers in the Klang Valley.	UNDERSTAND	**CREATE**
7	Compare FOUR (4) point of views of entrepreneurs with FOUR (4) for managers the way they look at the things.	UNDERSTAND	**ANALYZE**
8	Define electronic commerce	UNDERSTAND	**REMEMBER**
9	Define the term project in the context of project management.	UNDERSTAND	**REMEMBER**
10	Argue critically with relevant examples.	UNDERSTAND	**EVALUATE**
Total Correct Answer		1/10	10%

Table 4.2.2 Decision Tree Classifier Result

Naïve Bayes

No.	Questions	Predicted Answer	Correct Answer
1	Apply Porter's five competitive forces analysis to examine the summer job industry for your uncle.	**APPLY**	APPLY
2	Discuss the extent to which Value Chain Analysis can be applied in the logistics sector	**APPLY**	APPLY
3	Information Systems value is determined by the strong relationships among THREE (3) major components. Name them.	**REMEMBER**	REMEMBER
4	Briefly explain any TWO (2) observations of information technology trend using Moore's Law.	ANALYZE	**UNDERSTAND**
5	Relate cycle-time reduction to improved performance. Justify your answer.	**EVALUATE**	EVALUATE
6	Prepare a research proposal on a study that you have to conduct on the purchasing behaviour of teenagers in the Klang Valley.	**CREATE**	CREATE
7	Compare FOUR (4) point of views of entrepreneurs with FOUR (4) for managers the way they look at the things.	**ANALYZE**	ANALYZE
8	Define electronic commerce	APPLY	**REMEMBER**
9	Define the term project in the context of project management.	**REMEMBER**	REMEMBER
10	Argue critically with relevant examples.	APPLY	**EVALUATE**
Total Correct Answer		7/10	70%

Table 4.2.3 Naïve Bayes Classifier Result

K-Nearest Neighbour

No.	Questions	Predicted Answer	Correct Answer
1	Apply Porter's five competitive forces analysis to examine the summer job industry for your uncle.	UNDERSTAND	**APPLY**
2	Discuss the extent to which Value Chain Analysis can be applied in the logistics sector	**APPLY**	APPLY
3	Information Systems value is determined by the strong relationships among THREE (3) major components. Name them.	**REMEMBER**	REMEMBER
4	Briefly explain any TWO (2) observations of information technology trend using Moore's Law.	APPLY	**UNDERSTAND**
5	Relate cycle-time reduction to improved performance. Justify your answer.	REMEMBER	**EVALUATE**
6	Prepare a research proposal on a study that you have to conduct on the purchasing behaviour of teenagers in the Klang Valley.	EVALUATE	**CREATE**
7	Compare FOUR (4) point of views of entrepreneurs with FOUR (4) for managers the way they look at the things.	**ANALYZE**	ANALYZE
8	Define electronic commerce	UNDERSTAND	**REMEMBER**
9	Define the term project in the context of project management.	UNDERSTAND	**REMEMBER**
10	Argue critically with relevant examples.	UNDERSTAND	**EVALUATE**
Total Correct Answer		3/10	30%

Table 4.2.4 K-Nearest Neighbour Classifier Result

For the purpose of achieving a satisfactory accuracy for the question classification tool in this study, Support Vector Machine, Decision Tree, Naïve Bayes, and also K-Nearest Neighbour classifier techniques were studied and compared through 10-fold cross-validation method and also real data type. A total number of 10 real data type act as testing questions regarding on business domain with various courses and subjects will be entered into the according classifier in order to determine which classification techniques can achieve the highest accuracy among others.

Table 4.1.5 shows the table regarding of the comparison of accuracy through 10-fold cross-validation method among four classifiers. Based on the table, Decision Tree has achieved 66.66% accuracy, which was the highest among others, followed by SVM which has 63.875% accuracy. Meanwhile although Decision Tree has achieved the highest accuracy with cross-validation method, it has been achieved the lowest accuracy when it comes to classifying with real data type.

Table 4.2.1 – Table 4.2.4 shows the tables regarding of a number of 10 real testing questions and its original answers along with predicted answers among 4 different classification techniques. Based on the table above, Decision Tree classification has achieved the lowest accuracy among others three which only correctly predicted 1 question out of 10. Next, K-Nearest Neighbour classification technique has achieved a slightly better result compared to Decision Tree classification which has correctly predicted 3 questions out of 10. However, K-Nearest Neighbour classifications were also not in the consideration to adopt due to the accuracy was not even hit 50%. Moreover, Support Vector Machine and also Naïve Bayes classification have achieved a satisfactory accuracy which both of them has correctly predicted 7 questions out of 10 testing questions. Due to both of the classification techniques have achieved the same and satisfy accuracy, however, SVM has achieved a better accuracy compare to Naïve Bayes in cross-validation method. Therefore, SVM will be chosen to develop the question classification tool in this study.

4.3 Graphical User Interface

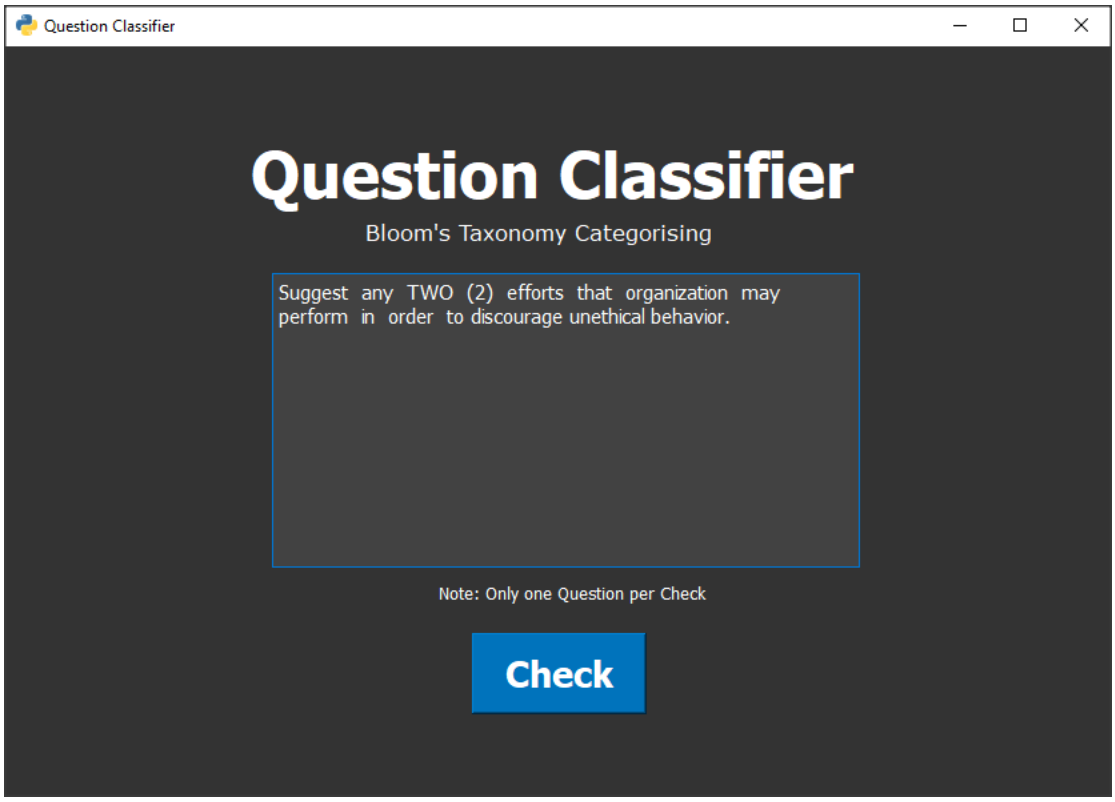


Figure 4.3.1 – Question Classifier Home Page

Figure 4.3.1 shows the graphical user interface of the home page of proposed question classification tool. In this page, the program allows users to enter their question into the designated text box that places in the middle of the program with blue colour border. However, the program does only allow users to enter one question per check. After that, users just need to click the check button in order to view which category of Bloom's Taxonomy cognitive domain is the question belong to. If the users want to exit the system, they can just simply click the "X" mark that place in the top right corner of the program.

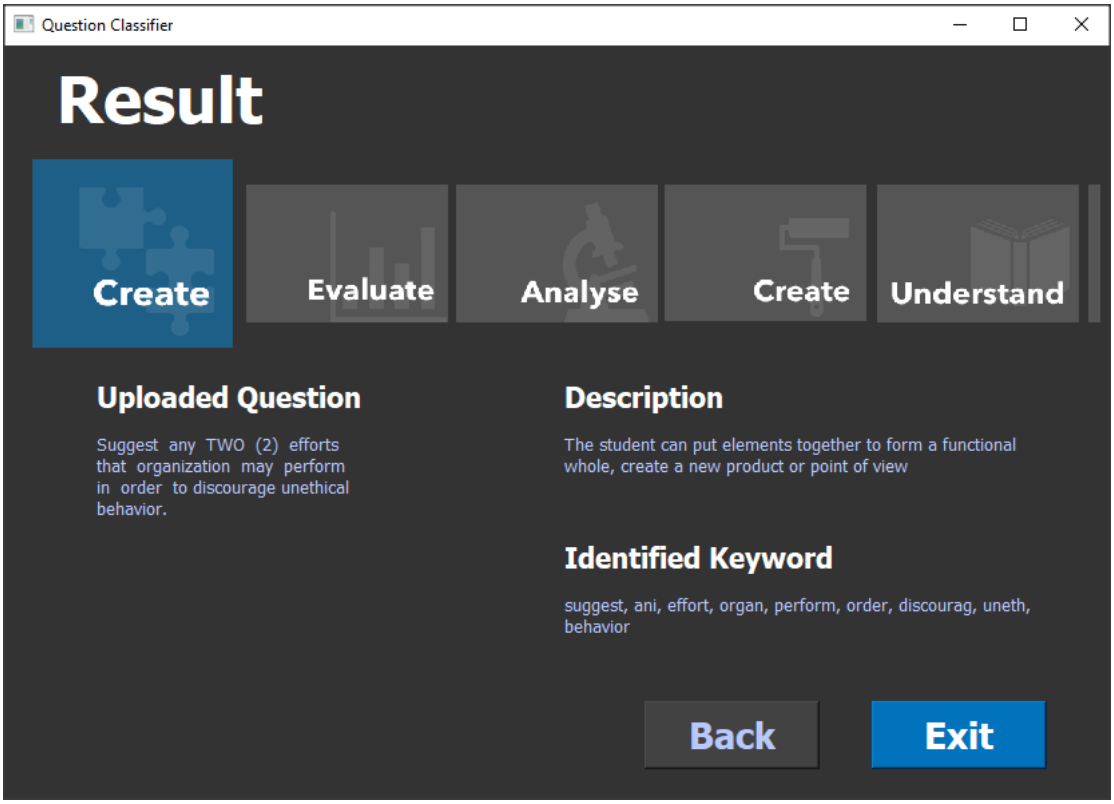


Figure 4.3.2 – Question Classifier Result Page

Figure 4.3.2 shows the result page of the proposed question classifier. In this page, the program will tell the users which level of Bloom's Taxonomy cognitive domain does the entered question belong to. The predicted level will be shown on top of this page. Next, users can view the uploaded questions in the middle left on this page, the description of the predicted bloom's taxonomy level, and also the keywords that identified from the inputted question. If the users want to check the level of another question, they can click the back button in order to navigate back them to the home page. Else if the users want to exit the program, they can just simply click the provided exit button on the bottom right part of the program or the "X" mark that place in the top right side of the program.

Chapter 5: Conclusion

This study has presented a question classification tool that comes with 70% accuracy and which might benefit to educators in educational environments. The proposed question classification tool will categorize the input questions into according Bloom Taxonomy's cognitive domain automatically. Understanding the cognitive domain from Bloom Taxonomy might be important because it helps educator to design better and classify the examination or practical questions in order to improve the overall quality of learning. To develop the classification tool, a study of Data Mining, Text Mining, Bloom Taxonomy, Natural Language Processing, and Question Classification has been done in order to get an insight or overview in the question classification. Several question classification techniques have also been studied and compared in this paper.

In this study, Natural Language Processing has played an important role. Text pre-processing techniques such as Tokenization, Stemming, Lemmatization, and Stopword Removal has used to do the cleaning of the raw question into computer understandable structure in order to facilitate later processing. Four machine learning classification techniques (Support Vector Machine, Decision Tree, Naïve Bayes, and K-Nearest Neighbor) has also been studied and compared in order to identify the most suitable classifier that comes with the highest accuracy among others. After the comparison, Support Vector Machine has achieved a satisfactory result and therefore decided to use to develop the rules in order to match and categorize questions into according Bloom Taxonomy's category. In order to teach the machine, a number of 163 training questions along with its answer were first fed to the system to let it study the patterns.

This automated question classification tool might overcome the problem occurred when work was done manually. However, there might be a problem of accuracy because there exists only a very little information in examination questions and also due to the limited number of training data set. Therefore, that might not enough to be in classifying process to get a higher accuracy. In order to overcome this problem, a large training data set is required. Hence, this study compared several classification techniques in question classification and finds an appropriate classification technique with the desired accuracy.

5.1 Implementation Issue and Challenge

This study has developed a standalone desktop question classification tool that can classify the text-based examination or practical questions according to the six Bloom Taxonomy Cognitive levels. However, question classification could be challenging as compared to document classification in getting a reasonable accuracy because in the examination or practical questions, there exist only a very little information and that may not be enough to effectively classify the questions as opposed to document classification (Sangodiah, Ahmad & Ahmad 2014). Besides, some of the keywords from the questions might fall into two different categories of Bloom Taxonomy Cognitive level. For example, the verb 'Write' can be categorized into Knowledge and Synthesis level at the same time (Omar et al. 2012). Moreover, a large training set is required to train the system in order to achieve a higher accuracy. Furthermore, some final examination questions might use images as their question or description. In this case, the proposed question classification tool will not able to process the questions in the images form. Lastly, this program was only able to handle the questions that are in English language but not others.

5.2 Future Work

There are few further improvements that can be made on this standalone desktop question classification tool. First of all, the proposed classifier was designed and can just only operate in a desktop application based. Therefore, a further improvement can be made on this classifier in order to let it operate on a universal platform such as mobile devices, web-based, and so on. Besides, the proposed classifier can only classify those questions that are in English based language. Therefore, an improvement can be made on this existing classifier in order to let it understand and handling others languages too. Moreover, the accuracy of this classifier can be further improved also. Although the accuracy of this classifier has achieved the project objective, however, a higher percentage of the accuracy is possible as long as there is a sufficiently large amount of training dataset available. The larger the training data set, the higher the classifier's accuracy. Furthermore, the currently proposed question classifier was only allowed users to enter one question per check. An improvement might be made based on this issue to let users enter multiple questions at a time or even let the users import Microsoft Excel to it. Lastly, the proposed question classifier was only capable of dealing with text-based data. Hence, a further improvement can be made such as analysing graphical-based item or a Portable Document Format file and then identifying which level of Bloom's Taxonomy cognitive domain do they belong to.

Bibliography

Aggarwal, C & Zhai, C 2012, 'A Survey of Text Classification Algorithms', *Mining Text Data,* pp. 163-222.

Akcapinar, G 2015, 'How automated feedback through text mining changes plagiaristic behavior in online assignments', *Computers & Education*, vol. 87, no. C, pp. 123-130.

Alpaydin, E 2010, *Introduction to Machine Learning*. MIT Press, Cambridge.

Alwan, M 2015, *What is System Development Life Cycle?*. Available from: <https://airbrake.io/blog/insight/what-is-system-development-life-cycle>. [26 November 2016]

Anderson, L, Krathwohl, D, Airasian, P, Cruikshank, K, Mayer, R, Pintrich, P, Raths, J & Wittrock, M 2001, *Taxonomy for Learning, Teaching, and Assessing, A: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Edition*, Pearson, New Jersey.

Azevedo, A & Santos, M 2008, 'KDD, SEMMA and CRISP-DM: A Parallel Overview', in *IADIS European Conference Data Mining,* 24-26 July, Amsterdam, Netherlands, pp. 182-185.

Bagga, S & Singh, G.N. 2012, 'Applications of Data Mining', *International Journal for Science and Emerging Technologies with Latest Trends,* vol. 19, no. 23, pp. 19-22. Available from: < http://www.ijsett.com/images/P5.pdf>. [19 November 2016]

Baker, R 2010, 'Data Mining for Education', *International Encyclopedia of Education*, vol. 7, pp. 112-118.

Berndt, D & Clifford, J 1996, *Finding Patterns in Time Series: A Dynamic Programming Approach*, AAAI Press, Menlo Park.

Bloom, B, Engelhart, M, Furst, E, Hill, W & Krathwohl, D 1956, *Taxanomy of educational objectives: Handbook 1: Cognitive domain*, David McKAY, New York.

Brownlee, J 2014, *An Introduction to Feature Selection*. Available from <https://machinelearningmastery.com/an-introduction-to-feature-selection/>. [11 April 2018].

Brownlee, J 2017, *A Gentle Introduction to the Bag-of-Words Model*. Available from < https://machinelearningmastery.com/gentle-introduction-bag-words-model/>. [11 April 2018].

Chapman, P, Clinton, J, Kerber, R, Khabaza, T, Reinartz, T, Shearer, C & Rudiger, W 2000, *CRISP-DM 1.0*, SPSS Inc., Chicago. Available from: <https://www.the-modeling-agency.com/crisp-dm.pdf>. [25 November 2016]

Chopra, A, Prashar, A & Sain, C 2013, 'Natural Language Processing', *International Journal of Technology Enhancements and Emerging Engineering Research*, vol. 1, no. 4, pp. 131-133.

Clark, D 1999, *Bloom's Taxonomy: The Original Cognitive Domain*. Available from: <http://www.nwlink.com/~donclark/hrd/Bloom/original_cognitive_version.html>. [24 November 2016].

Clark, D 2015, *Bloom's Taxonomy of Learning Domains*. Available from: <http://www.nwlink.com/~donclark/hrd/bloom.html>. [22 November 2016]

Copestake, A 2004, *Natural Language Processing*, lecturer notes distributed in Computer Laboratory at University of Cambridge, England on 18 April 2004.

Dang, N, Prieta, F, Corchado, J & Moreno, M 2016, 'Framework for Retrieving Relevant Contents Related to Fashion from Online Social Network Data', *Advances in Intelligent Systems and Computing*, vol. 473, pp. 335-347.

Doleck, T, Basnet, R, Poitras, E & Lajoie, S 2015, 'Mining learner – system interaction data: implications for modeling learner behaviors and improving overlay models', *Journal of Computers in Education*, vol. 2, no. 4, pp. 421-447.

Eslick, I & Liu H 2005, *Lingutils: A Natural Language Toolkit for Common Lisp*. Available from: < https://common-lisp.net/project/langutils/LISP2005-langutils.pdf >. [27 November 2016]

Fayyad, U, Piatetsky, G & Smyth, P 1997, *From Data Mining to Knowledge Discovery in Database*, AAAI, Providence.

Feldman, R & Sanger, J 2007, *The Text Mining Handbook*, Cambridge University Press, New York.

Goldberg, Y 2017, *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers, California.

Guyon, I & Elisseeff, A 2003, 'An Introduction to Variable and Feature Selection', *Journal of Machine Learning Research 3,* pp. 1157-1182.

Haris, S & Omar, N 2015, 'Bloom's Taxonomy Question Categorization using Rules and N-gram Approach', *Journal of Theoretical and Applied Information Technology,* vol. 76, no. 3, pp. 401-407.

He, W 2013, 'Examining students' online interaction in a live video streaming environment using data mining and text mining', *Computers in Human Behavior*, vol. 29, no. 1, pp. 90-102.

Ikonomakis, M, Kotsiantis, S & Tampakas, V 2005, 'Text Classification Using Machine Learning Techniques', *WSEAS TRANSACTIONS on COMPUTERS,* vol. 4, no. 8, pp. 966-974.

International Educational Data Mining Society, n.d., *Educational Data Mining*. Avalilable from: <http://www.educationaldatamining.org>. [19 November 2016].

Joachims, T n.d., 'Text Categorization with Support Vector Machines: Learning with Many Relevant Features', *Cornell CIS*. Available from: < http://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf>. [22 November 2016]

Jurafsky, D & Martin, J 2000, *Speech and Language Processing*. Prentice-Hall, New Jersey.

Li, X & Roth, D 2002, 'Learning Question Classifiers', *Proceedings of the 19th international conference on Computational linguistics*, vol. 1, pp. 1-7. Available from: < http://dl.acm.org/citation.cfm?id=1072378 >. [19 November 2016]

Manning, C & Schutze H 1999, Foundations of Statistical Natural Language Processing, 2nd edn., The MIT Press, Cambridge.

Meseguer, J, Calafate, C, Cano, J, Manzoni, P 2015, 'Assessing the Impact of Driving Behavior on Instantaneous Fuel Consumption', *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 443 – 448.

Metzler, D & Croft, B 2004, 'Analysis of Statistical Question Classification for Fact -based Questions', *Journal of Information Retrieval*, vol. 8, no. 3, pp. 481-504.

Mitchell, E 2014, *6 Indicators Your Business Processes are Inefficient*. Available from: <http://www.eclgrp.com/blog/6-indicators-your-business-processes-are-inefficient>. [21 November 2016].

Oliver, D, Dobele, T, Greber, M & Roberts, T 2004, 'This course has a Bloom Rating of 3.9', *Proceedings of Sixth Australasian Conference on Computing Education*, vol. 30, pp. 227-231.

Omar, N, Haris, S, Hassan, R, Arshad, H, Rahmat, M, Zainal, NF & Zulkifli, R 2012 'Automated Analysis of Exam Questions According to Bloom's Taxonomy', *Procedia – Social and Behavioral Sciences*, vol. 59, pp. 297-303.

Osman, A & Alattab, A 2013, 'Educational Data Mining: A Case Study of Teacher's Classroom Questions', in *13th International Conference on Intellient Systems Design and Applications*, 8-10 December, Bangi, Malaysia, pp. 92-97.

Radovanovic, M & Ivanovic, M 2008, 'Text Mining: Approaches and Applications', *Novi sad journal of mathematics*, vol. 38, no. 3, pp 227-234.

Sael, N, Marzak, A & Behja, H 2013, 'Web Usage Mining Data Preprocessing and Multi Level Analysis on Moodle', *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*.

Sangodiah, A, Ahmad, R & Ahmad, W 2014, 'A Review in Feature Extraction Approach in Question Classification Using Support Vector Machine', *2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pp. 459-464.

SAS n.d., *SEMMA*. Available from: <http://faculty.smu.edu/tfomby/eco5385_eco6380/data/SPSS/SAS%20_%20SEMMA.pdf>. [25 November 2016]

SAS n.d., *What can I do with SAS*. Available from: <https://support.sas.com/software/products/university-edition/faq/SAS_whatis.htm>. [25 November 2016]

Scott, T 2003, 'Bloom's taxonomy applied to testing in computer science classes', *Journal of Computing Science in Colleges*, vol. 19, no. 1, pp. 267-274.

Shearer, C 2000, 'The CRISP-DM Model: The New Blueprint for Data Mining', *Journal of Data Warehousing*, vol. 5, no. 4, pp. 14-15.

SIGKDD 2006, *Data Mining Curriculum*. Available from: <http://www.kdd.org/curriculum/index.html>. [18 November 2016]

Tierney, J 2011. *Do You Suffer From Decision Fatigue?*. Available from: <http://www.nytimes.com/2011/08/21/magazine/do-you-suffer-from-decision-fatigue.html>. [20 November 2016].

Weiss, S & Kulikowski, C 1991, *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems*, Morgan Kaufmann, San Francisco.

Xu, Y & Reynolds, N 2012, 'Using Text Mining Techniques to Analyze Students' Written Responses to a Teacher Leadership Dilemma', *International Journal of Computer Theory and Engineering*, vol. 4, no. 4, pp. 575-578.

Appendix A: Training Dataset

NO	QUESTION	LEVEL	BT LEVEL
1	Suggest any TWO (2) efforts that organization may perform in order to discourage unethical behavior.	HOTS	CREATE
2	Discuss how Resource Based View can be implemented in organizations to reinforce financial standing.	MOTS	APPLY
3	Define factors of production.	LOTS	REMEMBER
4	Briefly describe FIVE (5) general types of factors of production.	LOTS	UNDERSTAND
5	In your opinion, provide any TWO (2) reasons why the actions of one firm can significantly affect the sales of the other firms in an oligopoly market.	LOTS	UNDERSTAND
6	Define electronic commerce.	LOTS	REMEMBER
7	Explain THREE (3) broad categories of e-commerce with examples.	LOTS	UNDERSTAND
8	Discuss any THREE (3) ways by which an organization can benefit from e-commerce.	LOTS	UNDERSTAND
9	Advise Smith on any FIVE (5) types of information which should be included in the website.	HOTS	CREATE
10	Suggest any THREE (3) specific ways with elaborations on how this new business may utilize the Internet technology to achieve cost efficiency.	HOTS	CREATE
11	Differentiate between data and information using examples.	MOTS	ANALYZE
12	State any FOUR (4) characteristics of valuable information.	LOTS	REMEMBER
13	Explain the differences between volatile and non-volatile storage.	LOTS	UNDERSTAND
14	Suggest THREE (3) specific types of application software that may help to support James's business activities.	HOTS	CREATE
15	List and describe THREE (3) accounting cycles in an	LOTS	REMEMBER

	Accounting Information System.		
16	Explain how SWOT and Porter's techniques is applied in the manufacturing industries.	MOTS	APPLY
17	State FOUR (4) basic business activities that are performed in the revenue cycle.	LOTS	REMEMBER
18	List any THREE (3) practical approaches for payment collection in revenue cycle.	LOTS	REMEMBER
19	Compare and contrast Manufacturing Resource Planning (MRP-II) and Lean Manufacturing.	MOTS	ANALYZE
20	As a manager in an Information Technology firm, you are required to suggest and elaborate any THREE (3) actions that organizations can take as part of the responsibility towards the society.	HOTS	CREATE
21	Illustrate each of these situations with TWO (2) real-life examples.	LOTS	UNDERSTAND
22	Explain the differences between ethical dilemma and ethical lapse.	LOTS	UNDERSTAND
23	Suggest and justify ONE (1) strategy that your company may consider for marketing its products internationally.	HOTS	EVALUATE
24	Differentiate between Market Economy and Planned Economy.	MOTS	ANALYZE
25	Provide each of Market Economy and Planned Economy with an example.	LOTS	UNDERSTAND
26	Differentiate between a wholesaler and a retailer.	MOTS	ANALYZE
27	In your opinion, identify the relationship between factors of and a country's economic system.	LOTS	UNDERSTAND
28	Briefly describe THREE (3) types of business process that traditionally organized around functional areas of business. Provide an example of each.	LOTS	REMEMBER
29	Propose and justify ONE (1) international market-entry strategy that you may consider to market your products internationally.	HOTS	EVALUATE
30	Construct a basic diagram or figure to demonstrate your understandings on what Information System is.	LOTS	UNDERSTAND

31	State a main difference between a customer and a contact.	LOTS	UNDERSTAND
32	Identify ONE (1) most likely error of the following quotation.	LOTS	REMEMBER
33	Determine which sub-stage that the deal is currently in for the following cases.	MOTS	APPLY
34	Three sales sub-stages: qualification, proposition and negotiation are included in both OpenERP and SugarCRM, before a deal is won. Give a brief description of these sub-stages.	LOTS	UNDERSTAND
35	State your recommended solution with justification.	HOTS	EVALUATE
36	Explain TWO (2) ways on how monopoly harms consumers and hurts the economy. Justify and illustrate your answers.	HOTS	EVALUATE
37	Suggest and elaborate any TWO (2) efforts or methods that your company may use in order to gain competitive advantage over your competitors.	HOTS	CREATE
38	Justify your answers with elaborations by providing any TWO (2) reasons.	HOTS	EVALUATE
39	Differentiate between data and information using a real life example.	MOTS	ANALYZE
40	Briefly discuss on the difference between volatile and non-volatile storage.	LOTS	UNDERSTAND
41	Briefly discuss any THREE (3) specific internet technologies or tools that are commonly found in providing support for business activities.	LOTS	UNDERSTAND
42	Provide a real life example on conflict of interest situation.	LOTS	UNDERSTAND
43	Illustrate your understanding for oligopoly competition and monopolistic competition using ONE (1) real life example for each.	LOTS	UNDERSTAND
44	Briefly discuss any TWO (2) consequences for whistle-blowing activity.	LOTS	UNDERSTAND
45	List any TWO (2) real life examples of the organization that implement the clicks-and-bricks B2C E-Commerce.	LOTS	REMEMBER

46	Provide THREE (3) examples of input devices.	LOTS	REMEMBER
47	Identify and discuss THREE (3) issues on employee productivity problems that companies face.	LOTS	UNDERSTAND
48	Briefly provide any TWO (2) real life situations to illustrate on ethical dilemma.	LOTS	UNDERSTAND
49	Propose an e-commerce classification model that suits the needs of John's business. Justify your answer.	HOTS	EVALUATE
50	Explain the concept of clicks-and-bricks model in e-commerce.	LOTS	UNDERSTAND
51	Explain each inventory control method and suggest TWO (2) types of products for each method.	HOTS	CREATE
52	Explain Greimas square tool and discuss how it can be applied the aviation industry.	MOTS	APPLY
53	List THREE (3) advantages and TWO (2) disadvantages codification.	LOTS	REMEMBER
54	Identify and discuss THREE (3) issues on employee productivity problems that companies face.	LOTS	UNDERSTAND
55	Provide TWO (2) possible reasons why AirAsia decide to offer this new online service.	LOTS	UNDERSTAND
56	Identify THREE (3) possible threats you should anticipate for this Internet-based business.	LOTS	UNDERSTAND
57	Suggest THREE (3) methods Game2 can take to prevent damage to its Web sites and continuing operations. Justify your answer.	HOTS	EVALUATE
58	Explain the concept of Customer Relationship Management (CRM).	LOTS	UNDERSTAND
59	Apply Porter's five competitive forces analysis to examine the summer job industry for your uncle.	MOTS	APPLY
60	Discover FOUR (4) characteristic of IT infrastructure to be considered.	HOTS	CREATE
61	Concisely define what SWOT analysis is and state why it is frequently producing conflicting views.	LOTS	UNDERSTAND
62	Formulate the profit for the company.	HOTS	CREATE
63	Discuss the extent to which Value Chain Analysis can be applied in the logistics sector.	MOTS	APPLY

64	Information systems value is determined by the strong relationships among THREE (3) major components. Name them.	LOTS	REMEMBER
65	Cloud computing evolved from earlier technologies. State them.	LOTS	REMEMBER
66	Briefly explain any TWO (2) observations of information technology trend using Moore's Law.	LOTS	UNDERSTAND
67	Relate cycle-time reduction to improved performance. Justify your answer.	HOTS	EVALUATE
68	Define privacy.	LOTS	REMEMBER
69	In your own words, describe Supply Chain Management (SCM).	LOTS	UNDERSTAND
70	In your opinion, discuss how modern supply chain management systems facilitate a pull-based model.	LOTS	UNDERSTAND
71	Define the term Radio Frequency Identification (RFID).	LOTS	REMEMBER
72	Explain how RFID provides value to businesses.	LOTS	UNDERSTAND
73	Use Porter's five competitive forces to analyze FiredUp, Inc.	MOTS	APPLY
74	Explain how GST approach can be deployed in the context of Accounting Information System.	MOTS	APPLY
75	Base on the case study given, summarize THREE (3) ISM security problems.	LOTS	UNDERSTAND
76	Classify information systems based on the support IS can provide.	LOTS	UNDERSTAND
77	Explain how Referral Networks approach can be used as a tool to increase number of customers.	MOTS	APPLY
78	Draw the diagrams for a centralized database and a distributed database.	LOTS	UNDERSTAND
79	Explain how the future of social media can be explained by Web 3.0.	LOTS	UNDERSTAND
80	Define knowledge and provide an example of knowledge produced by a vehicle management system.	LOTS	REMEMBER
81	Identify PayPal's business model.	LOTS	REMEMBER

82	Scaling methods can improve the performance of the web site. Propose to Jane which scaling method suit her current business needs. Justify your answer.	HOTS	EVALUATE
83	Define CRM and discuss how the CRM can be implemented in the business environment.	MOTS	APPLY
84	Examine how e-commerce facilitates Value Chain and Supply Chain.	MOTS	ANALYZE
85	Explain how auction is related to dynamic pricing.	LOTS	UNDERSTAND
86	Compare and contrast Address Verification System (AVS) and Card verification number (CVN).	MOTS	ANALYZE
87	Provide THREE (3) justification on how would the university community be encouraged to place orders and become loyal customers.	HOTS	EVALUATE
88	Personalization is becoming an important element in Electronic Commerce. Explain TWO (2) techniques that can be used to learn about consumer behavior and how it can be used to facilitate customer service.	LOTS	UNDERSTAND
89	Provide TWO (2) suggestions on how a Web presence help Columbiana's government.	HOTS	CREATE
90	Briefly discuss how TQM could be applied m the management of the newspapers' editorials.	MOTS	APPLY
91	Based on your observation of the local newspaper industry, discuss and state your stance on whether gender affects professional news selection.	HOTS	EVALUATE
92	Discuss how far you agree with this statement with emphasis on circulation management of a newspaper company.	HOTS	EVALUATE
93	Argue critically with relevant examples.	HOTS	EVALUATE
94	Critically discuss the quote above and explain FIVE(5) ways the interest groups can be more effective with relevant examples.	MOTS	ANALYZE
95	Critically analyze and discuss the problems with secondary data. Provide relevant examples to support your answer.	MOTS	ANALYZE
96	Describe the term 'urbanization' and discuss how urbanization influences the development of marketing	LOTS	UNDERSTAND

	and advertising strategies.		
97	Critically discuss how demographic characteristics would affect a product development in a foreign market.	MOTS	ANALYZE
98	Analyze and discuss the challenges and opportunities that social media creates.	MOTS	ANALYZE
99	Critically analyze the image above and discuss the strategic decisions associated with the appeals of the advertisement.	MOTS	ANALYZE
100	Show how NLP techniques can be deployed in business.	MOTS	APPLY
101	Suggest THREE (3) ethical considerations that ANA advertisers should take into account in designing commercials.	HOTS	CREATE
102	Create a poster that promote jeans to local senior citizen which appeal directly to a multiracial citizen in Malaysia.	HOTS	CREATE
103	Select a local 'HOMEGROWN' brand and discuss its international marketing mix.	LOTS	UNDERSTAND
104	Critically discuss FOUR (4) main dimensions that used to classify countries based on Hofstede's Cultural Dimensions Theory.	MOTS	ANALYZE
105	Compare and contrast between low context cultures and high context cultures.	MOTS	ANALYZE
106	Critically discuss the THREE (3) options in handling the international media planning or buying.	MOTS	ANALYZE
107	Design an ideal newspaper organisational chart. Then describe the functions of the respective departments and how they relate to each other ?C horizontally and vertically.	HOTS	CREATE
108	Demonstrate email and social media approaches to create effective marketing plan.	MOTS	APPLY
109	Examine the two roles and show how newspaper managers can successfully draw a balance between their economic role and social role.	MOTS	ANALYZE
110	Suggest a house-style and editorial policies for your	HOTS	CREATE

	newly introduced newspaper targeted at young readers aged between 15 to 25 years.		
111	Examine how conventional newspaper can continue to maintain its circulation with the easy access to alternative information on the web.	MOTS	ANALYZE
112	As an editor-in-chief, suggest how you can motivate your experienced journalists to remain with your paper.	HOTS	CREATE
113	Discuss the factors that you think may have affected the quality of newspapers in Malaysia and identify the key elements for a successful newspaper .	MOTS	ANALYZE
114	List any five departments and discuss the relations between these departments and the importance of depending on each other.	MOTS	ANALYZE
115	PEST and SWOT are popular strategy tools. Discuss how the TWO(2) tools can be applied in manufacturing business environment.	MOTS	APPLY
116	Evaluate the performance of Malaysian newspapers and comment on what you think they should do to win back readers, particularly those who show preference for online news.	HOTS	EVALUATE
117	Prepare a research proposal on a study that you have to conduct on the purchasing behaviour of teenagers in the Klang Valley.	HOTS	CREATE
118	Evaluate the need for ethics in research.	HOTS	EVALUATE
119	Explain the case study research method and distinguish between single and multiple case study design.	MOTS	ANALYZE
120	Argue the case for conducting experimental research involving humans and propose guidelines to ensure that the dignity and welfare of the subjects are maintained.	HOTS	EVALUATE
121	Differentiate between qualitative research and quantitative research in terms of their goal and design.	MOTS	ANALYZE
122	Develop a questionnaire based on at least FOUR	HOTS	CREATE

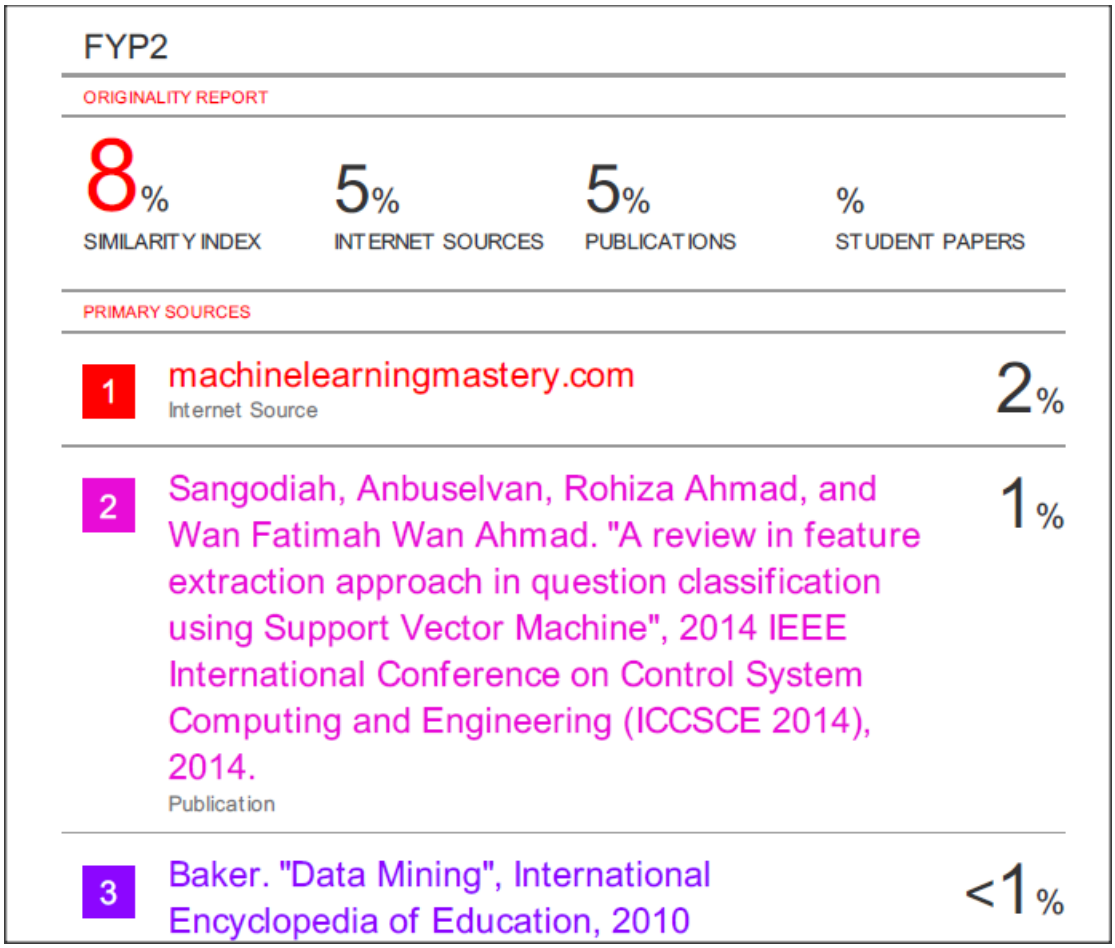
	open-ended and FOUR closed-ended questions for the interviews.		
123	Analyze the purpose of having a Purchasing Department in an organization.	MOTS	ANALYZE
124	Distinguish Paradox of the Commons from Tragedy of the Commons.	MOTS	ANALYZE
125	Evaluate the three specific effects caused by the applications of information technology on the nature of competition.	HOTS	EVALUATE
126	Advise the five steps, as recommended by Porter and Millar, that can be applied by business managers in taking advantage of information technology.	HOTS	CREATE
127	Winer (2001) proposes a 7-component Customer Relationship Management (CRM) Model as a framework to implement a successful CRM program. Evaluate each of the seven components of the CRM Model.	HOTS	EVALUATE
128	Davenport (2000) discusses four emerging trends to be considered in designing future enterprise systems. Analyze each emerging trend, its relevance and importance to the designing of future enterprise systems.	MOTS	ANALYZE
129	Apply SWOT analysis in education environment and discuss its impact in the revenue analysis.	MOTS	APPLY
130	Alavi and Leidner (2001) examine how information technology (IT) can be a key enabler in supporting four organizational knowledge management (KM) processes. Critically assess the role of IT in supporting the four KM processes.	HOTS	EVALUATE
131	PERCEPTION STUDIES based on SURVEY METHOD are useful and common in both academic and market research. Critically evaluate the strategies used in data collection that can increase the response rate when conducting a survey.	HOTS	EVALUATE
132	Critically review how a shift in the mental model with a different set of assumptions could change	HOTS	EVALUATE

	Yahoo.		
133	With reference to the case, evaluate the similarities and differences between open mindedness and mental models.	HOTS	EVALUATE
134	Leaders often face a dilemma because of a perceived conflict between the realm of business and the realm of ethics. Compare the characteristics of leaders who follow a strict rational self-interest approach with those who take an ethical approach.	MOTS	ANALYZE
135	Evaluate the appropriateness of drugs advertising by the pharmaceutical firms.	HOTS	EVALUATE
136	Distinguish clearly between power and authority.	MOTS	ANALYZE
137	Evaluate the types of power Benjamin Fang is exercising in Sun Plantations, and explain how this is being eroded by the changes taking place in the environment.	HOTS	EVALUATE
138	Discuss the extent to which the existence of a conflict between a company's objectives is acceptable.	HOTS	EVALUATE
139	Discuss and show the impact of CRM strategies in the telecommunication sector.	MOTS	APPLY
140	Discuss the extent to which it is feasible for a company to 'operationalise' its social responsibility Spirations, that is, whether it is possible to bring these considerations into strategic decision-making in a programmed or systematic way.	HOTS	EVALUATE
141	Identify and recommend any appropriate amendments to the proposal presented in order to rectify the errors made in estimating the company's current cost of capital and current value. For each of the revisions, explain the reasons why it should be amended.	HOTS	EVALUATE
142	Recommend with justification, a value or range of values to the Board of Directors for the acquisition of ESL Berhad.	HOTS	EVALUATE
143	Compare how the Net Present Value (NPV), internal rate of return (IRR) method and modified internal rate ofreturn (MIRR) method is different.	MOTS	ANALYZE

144	Analyze the product-market options that are available to Associated Meats Sdn Bhd.	MOTS	ANALYZE
145	Use philanthropy approach to manage corporate social responsibility effectively.	MOTS	APPLY
146	Critically review the strengths, weaknesses, opportunities and threats of Associated Meats Sdn Bhd in light of the forecast trends and developments.	HOTS	EVALUATE
147	Critically appraise the five competitive forces encased within Porter's "Five Forces" model within the context of a profit-oriented organisation and discuss the threat posed to the firm by each of these forces.	HOTS	EVALUATE
148	Evaluate the nature and effect of significant "entry barriers" on the formulation of a strategic plan for a business which is already established in the industry.	HOTS	EVALUATE
149	Synthesize any FIVE (5) factors that should be considered when devising a safe system to undertake a task in an organization.	HOTS	CREATE
150	Synthesize the need for ventilation as an engineering control.	HOTS	CREATE
151	Assess the reason for hazard communication and the mode of communication practice for chemical safety at worksite.	HOTS	EVALUATE
152	List FIVE (5) functions of Department of Safety and Health as outline in OSHA 1994.	LOTS	REMEMBER
153	Define the term project in the context of project management.	LOTS	REMEMBER
154	Define standard operating procedures as highlighted.	LOTS	REMEMBER
155	Show in a step by step manner how TQM can be implemented successfully in education sector.	MOTS	APPLY
156	Compare FOUR (4) point of views of entrepreneurs with FOUR (4) for managers the way they look at the things.	MOTS	ANALYZE
157	Being a technopreneur means dealing with innovation. Define what innovation is and also propose FOUR (4) types of innovations you can have	HOTS	CREATE

	in your organization.		
158	Presenting the business plan is very crucial to convince potential consumers and financers. List any SEVEN (7) suggestions for good presentation of your business plan.	LOTS	REMEMBER
159	List the advantages and disadvantages of Public Offering.	LOTS	REMEMBER
160	Show your calculations for all THREE (3) options and then discuss which one of the options you prefer.	MOTS	ANALYZE
161	Currently the product life cycle for Apple iPod is in the 'growth stage'. Provide evidence to support your answer.	HOTS	EVALUATE
162	Define brand audit.	LOTS	REMEMBER
163	Explain how Porter's concept can be used in education domain.	MOTS	APPLY

Plagiarism Check Result



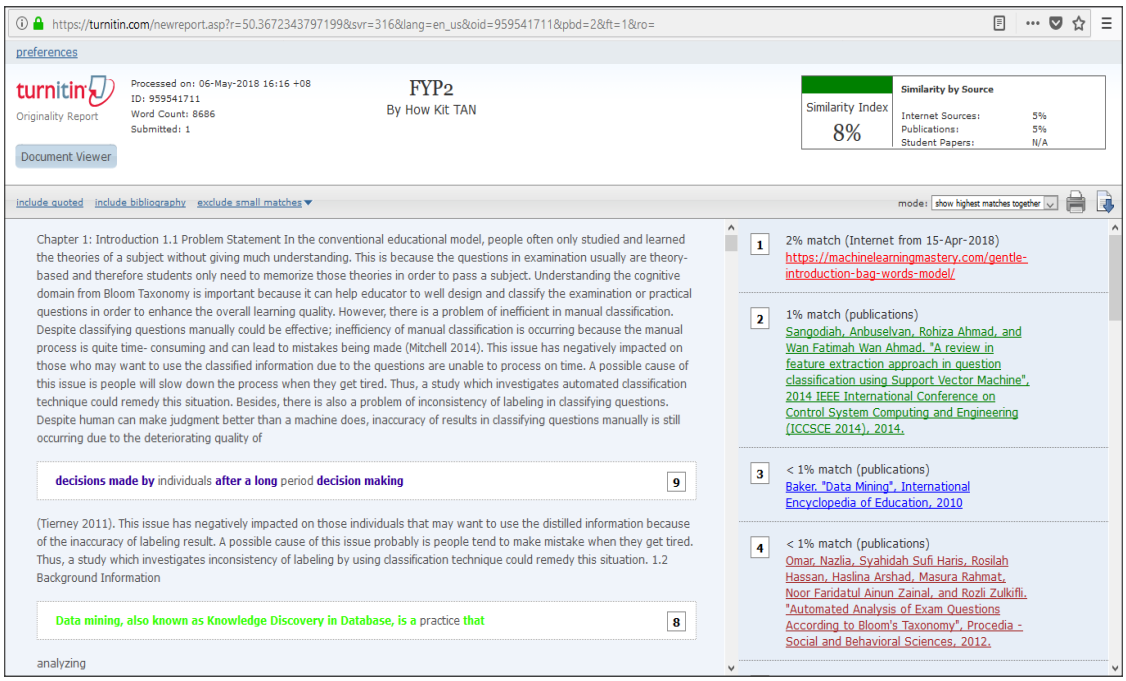
FYP2

ORIGINALITY REPORT

8%	5%	5%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	machinelearningmastery.com Internet Source	2%
2	Sangodiah, Anbuselvan, Rohiza Ahmad, and Wan Fatimah Wan Ahmad. "A review in feature extraction approach in question classification using Support Vector Machine", 2014 IEEE International Conference on Control System Computing and Engineering (ICCSCE 2014), 2014. Publication	1%
3	Baker. "Data Mining", International Encyclopedia of Education, 2010	<1%



https://turnitin.com/newreport.asp?r=50.3672343797199&svr=316&lang=en_us&oid=959541711&pbd=2&ft=1&ro=

preferences

turnitin Originality Report

Processed on: 06-May-2018 16:16 +08
ID: 959541711
Word Count: 8686
Submitted: 1

FYP2
By How Kit TAN

Similarity Index 8%

Similarity by Source
Internet Sources: 5%
Publications: 5%
Student Papers: N/A

Document Viewer

include quoted include bibliography exclude small matches

mode: show highest matches together

Chapter 1: Introduction 1.1 Problem Statement In the conventional educational model, people often only studied and learned the theories of a subject without giving much understanding. This is because the questions in examination usually are theory-based and therefore students only need to memorize those theories in order to pass a subject. Understanding the cognitive domain from Bloom Taxonomy is important because it can help educator to well design and classify the examination or practical questions in order to enhance the overall learning quality. However, there is a problem of inefficient in manual classification. Despite classifying questions manually could be effective; inefficiency of manual classification is occurring because the manual process is quite time- consuming and can lead to mistakes being made (Mitchell 2014). This issue has negatively impacted on those who may want to use the classified information due to the questions are unable to process on time. A possible cause of this issue is people will slow down the process when they get tired. Thus, a study which investigates automated classification technique could remedy this situation. Besides, there is also a problem of inconsistency of labeling in classifying questions. Despite human can make judgment better than a machine does, inaccuracy of results in classifying questions manually is still occurring due to the deteriorating quality of

decisions made by individuals after a long period decision making 9

(Tierney 2011). This issue has negatively impacted on those individuals that may want to use the distilled information because of the inaccuracy of labeling result. A possible cause of this issue probably is people tend to make mistake when they get tired. Thus, a study which investigates inconsistency of labeling by using classification technique could remedy this situation. 1.2 Background Information

Data mining, also known as Knowledge Discovery in Database, is a practice that 8

analyzing

1 2% match (Internet from 15-Apr-2018)
https://machinelearningmastery.com/gentle-introduction-bag-words-model/

2 1% match (publications)
Sangodiah, Anbuselvan, Rohiza Ahmad, and Wan Fatimah Wan Ahmad. "A review in feature extraction approach in question classification using Support Vector Machine", 2014 IEEE International Conference on Control System Computing and Engineering (ICCSCE 2014), 2014.

3 < 1% match (publications)
Baker. "Data Mining", International Encyclopedia of Education, 2010

4 < 1% match (publications)
Omar, Nazlia, Syahidah Sufi Haris, Rosilah Hassan, Haslina Arshad, Masura Rahmat, Noor Faridatul Ainun Zainal, and Rozli Zulkifli. "Automated Analysis of Exam Questions According to Bloom's Taxonomy", Procedia - Social and Behavioral Sciences, 2012.