KULLIYYAH OF INFORMATION AND

COMMUNICATION TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

FINAL YEAR PROJECT REPORT

NOTE ASSIST : TEXT SUMMARIZATION USING NLP

AHMAD ZULFAHMI BIN HARUM 1626867

HAZIQ ISKANDAR BIN SURIANI 1628259

SUPERVISED BY

DR. SURIANI BINTI SULAIMAN

AUGUST 2020

SEMESTER 2 2019/2020

FINAL YEAR PROJECT REPORT

NOTE ASSIST - TEXT SUMMARIZATION USING NLP

by

AHMAD ZULFAHMI BIN HARUM 1626867

HAZIQ ISKANDAR BIN SURIANI 1628259

SUPERVISED BY

DR. SURIANI BINTI SULAIMAN

In partial fulfillment of the requirements for the Bachelor of Computer Science Department of Computer Science Kulliyyah of Information and Communication Technology International Islamic University Malaysia

AUGUST 2019

Semester 2 2019/2020

ACKNOWLEDGEMENTS

In the name of Allah, the most merciful and the most gracious.

First of all, praise and thanks to Allah and foremost blessing enabled us to complete our final year project during the time given. With the help and the guidance of Allah S.W.T., we are grateful that we were given the strength to finish this project successfully. May Him always strengthen our faith in Him and endows us with knowledge and wisdom.

We would like to express our deepest gratitude, appreciation and indebtedness to our supervisor, Dr. Suriani Binti Sulaiman, for the willingness to supervise us throughout this project. When we were at the verge of failing the project, she showed us the brightside and continued to have hope in us. We are sincerely grateful to have a dedicated supervisor like her to help us with her guidance, knowledge and experience.

A special thanks to Madam Norzariyah Binti Yahya, for coordinating us for the projects regarding the needs and format of the documentation during the CSC 4904 class. With her words of encouragement and guidance, she showed us the potential that we have in ourselves.

Not to forget our family, our lecturers and friends who help us with the project and never lose hope on us by giving continuous moral support without any hesitancy. Our sincere thanks to all those who directly or indirectly assisted us to complete this project.

ABSTRACT

Students are often seen to be easily distracted in lecture rooms or halls. Some have moderate capabilities to memorize and understand topics fast. Besides, language barriers between students and lecturers on campus is a common issue with students from various language backgrounds. This project develops a note-taking web application with the support of Artificial Intelligence (AI), speech recognition and Natural Language Processing (NLP). This web application works by capturing a lecturer's speech during class into texts. The saved note can then be summarized using the TextRank algorithm to get an informative summary. The objectives of this project is to ensure that important points are being captured during the delivery of lectures, provide a semi-automated jotting platform where users can take control over their notes and assist in the note taking process especially for people with disabilities such as the hearing-impaired. By using the Summa python module for Text Summarization, YAKE! Module for Keyword Extraction and HTML5 Web Speech Kit for Speech Recognition, Note Assist helps students in managing their study notes. Our analysis shows that Note Assist are able to provide an acceptable quality of note summary.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
NLP	Natural Language Processing
YAKE!	Yet Another Keyword Extraction!

CHAPTER ONE

INTRODUCTION

1.1   BACKGROUND

Taking notes is one of the ways to memorize things. Note has been a crucial component in learning, organizing and keeping track of your schedule but it may be a fussy task to the students because jotting notes takes time and not all human beings can multitask for example writing and listening at the same time. Thus, taking note virtually using today's technology widens accessibility and eases the task.

In a big lecture hall that is occupied with at least 50 students, focusing on the lecture might be challenging as we assumed that the lecturer is not using a microphone. This can be related during the time of our study, we often see students easily distracted as they cannot give their full attention

The accessibility in notes software can be done through smartphones or laptops. Furthermore, the notes can be stored in a cloud storage for a secure backup. Improvements can be observed from time to time for note taking software in which a person's speech a person talks can be converted directly into texts. This can be done using one of the Artificial Intelligence (AI) subfields known as speech recognition. Speech recognition enables the interpretation of phrases, words and sentences in human speech which are identified by the computer. The implementation of the speech recognition enhances the use of note-taking software. In addition, users can straight away extract the important points from the speech in textual form.

This project describes the use of speech recognition to detect speech from audio or video and convert it into text. The developed software known as "Note Assist" can give a good contribution in users' lives and assist them in taking the notes. Other than that, for the users to be able to access any platforms such as mobile, laptops or desktops, we are making the software in a web-based system. Moreover, to get the best detection of the speech, users are recommended to use microphone or wired/wireless earphones to avoid noises from the environment.

1.2   PROBLEM STATEMENT

Note has been a crucial component in learning, organizing and keeping track of important points from the may be a fussy task to the students and accessibility is sometimes limited to the learners especially to persons with disabilities.

In a big lecture, focusing on the speaker might be challenging as we assume that few challenges may be a barrier to a comprehensive note taking process. Problems like language barriers, noise pollution and paying attention to tasks: taking notes and listening to the lecture. This can be related during the time of our study, we often see students easily distracted as they can not give their full attention. This method can contribute in giving an effective lesson to the audience from a speaker.

1.3   PROJECT OBJECTIVES

1)   To provide a cloud-based note taking platforms with high accessibility.

The system is able to be accessed by users in an online browser platform anywhere and anytime with an internet connection.

2)   To create a semi-automated jotting platform.

The system enables users to organize notes according to their preferences with a maximised control of the system.

3)   To provide automated summarization of the notes.

The system is able to provide a simple and shorter summarization of the notes with the same comprehension of the notes.

4)   To assist in the note taking process by the people with disabilities.

The system is able to ease the process of note taking by the people with disabilities such as deaf, dyslexic and mute.

1.4   PROJECT SCOPE

Note Assist is a website that helps users in note taking process on their lessons, forums, speeches and such. As a website, it will require an internet connection to enable the access to it. The website will allow users to write notes, transcribe a live speech to note, transcribe a recorded audio to note, summarize a long note, summarize an article from another website, summarize speeches from media files, organize notes in folder organization and suggest a category for the notes. For the development, it uses a python framework, Flask, and is integrated with web development languages: HTML, CSS and JavaScripts., and use HTML5 for speech recognition. It uses a python library to analyze and summarize articles and also for the keyword extraction. It also uses SQLite3 as a database.

1.5 SIGNIFICANCE OF THE PROJECT

Note Assist could help anyone with their lessons by giving them the opportunity to pay full attention to the lessons without the hassle of jotting down notes. With undivided attention, users can fully comprehend the lessons, speech or forums they are participating in. As a website, it provides a cross platform access to all users with any devices that available with the Chrome browser

This system is useful to break the language barriers between the two parties; students and teachers, through the multilanguage based platform. Moreover, it is more efficient in helping students to focus on the teaching lessons than dividing their concentrations to two different tasks; focusing on the lessons and jotting notes. This

enables students to give their thorough concentration to comprehend the teaching. Hence, they will not miss any of the important points in the teachers lesson sessions.

Furthermore, this project also targeted specifically students with disabilities such as deaf people may have problems on hearing the speeches, an automated speech engine will help them to capture the words by teachers. Hence, this project might be a game changer to a group of people that need special attention like them. In addition, the software can be used in any corporation that held a conference meeting to also break the language barriers through the translator features in the software.

CHAPTER TWO
REVIEW OF PREVIOUS WORKS

This section introduces the review of existing note-taking applications and also the previously available speech recognition engine and summarization algorithm. A study on similar applications, which implicate applications that have the same characteristics as Note Assist. A total of eleven applications has been studied to acquire the requirements for this project has been used as a reference for the development process for Note Assist Application. The table below depicts the features that are offered by the applications:

	SPEECH RECOGNITION	OFFLINE MODE	WIDGET SUPPORT	MULTI LANGUAGE (SPEECH)	TRANSLATIONS	CATEGORISATION	LIVE SYNCHRONIZATION	SUMMARISATION	WORDS COUNT	VOICE RECORDED	GRAMMAR CHECK	EXPORT/IMPORT	REMINDERS	SHARING OPTION	BACKUP/RESTORE	WORDS TAGGING	MULTIPLATFORM	SAVING OPTION	MODIFIABLE (DELETE, EDIT)	HANDWRITTEN NOTES	COLLABORATION	CLOUD STORAGE	KEYBOARD WRITTEN	VOCABLURY CHECK
GOOGLE KEEP	✓	✓	✓	☐	☐	✓	☐	☐	☐	✓	☐	☐	✓	✓	☐	☐	✓	✓	✓	✓	✓	✓	✓	☐
EVERNOTE	☐	✓	✓	☐	☐	✓	☐	☐	☐	✓	☐	☐	✓	✓	✓	✓	✓	✓	✓	✓	☐	✓	✓	☐
MS ONENOTE	✓	✓	☐	☐	✓	☐	☐	☐	☐	✓	☐	✓	☐	✓	☐	✓	✓	✓	✓	✓	✓	☐	✓	☐
MILANOTE	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	✓	☐	✓	☐	☐	✓	✓	✓	☐	✓	✓	✓	☐
DROPBOX PAPER	☐	☐	☐	☐	✓	☐	☐	☐	✓	☐	☐	☐	☐	✓	☐	☐	✓	✓	✓	☐	✓	✓	✓	☐
SIMPLENOTE	☐	✓	☐	☐	☐	✓	☐	☐	☐	☐	☐	☐	☐	✓	☐	✓	✓	✓	✓	☐	✓	☐	✓	☐
SQUID	☐	✓	✓	☐	☐	☐	☐	☐	☐	☐	☐	✓	☐	☐	✓	☐	☐	✓	✓	✓	☐	☐	☐	☐
ZOHO NOTEBOOK	☐	✓	✓	☐	☐	☐	☐	☐	☐	✓	☐	☐	✓	✓	☐	✓	✓	✓	✓	✓	☐	☐	✓	☐
VOICE NOTEBOOK	✓	☐	✓	☐	✓	☐	☐	☐	✓	☐	☐	✓	☐	✓	☐	☐	☐	✓	✓	☐	☐	☐	✓	☐
VOICE NOTES	✓	✓	☐	☐	✓	✓	☐	☐	✓	✓	✓	✓	☐	☐	☐	☐	✓	✓	☐	☐	☐	☐	✓	☐
SPEECHNOTES	✓	✓	✓	☐	✓	☐	☐	☐	✓	☐	☐	✓	☐	✓	✓	☐	☐	✓	✓	☐	☐	☐	✓	☐

Table 2.0 Software Features Comparisons

2.1 Top Three Existing Note Taking Applications

1) Google Keep

A note-taking service developed by Google that is available on the web, and mobile apps for the Android and iOS mobile operating systems.It offers a variety of tools for taking notes, including text, lists, images, and audio. Users can set reminders, which are integrated with Google Now. It is a cloud-based service that will be synchronized all through the cloud in many platforms attached to an account. Belows are few depictions of Google Keep:
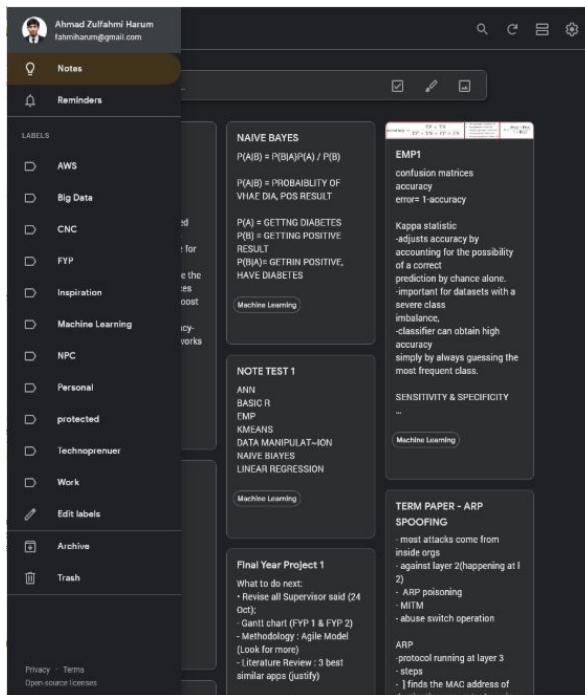
Figure 2.1.1.1 Google Keep Desktop   Figure 2.1.1.2 Google Keep Mobile
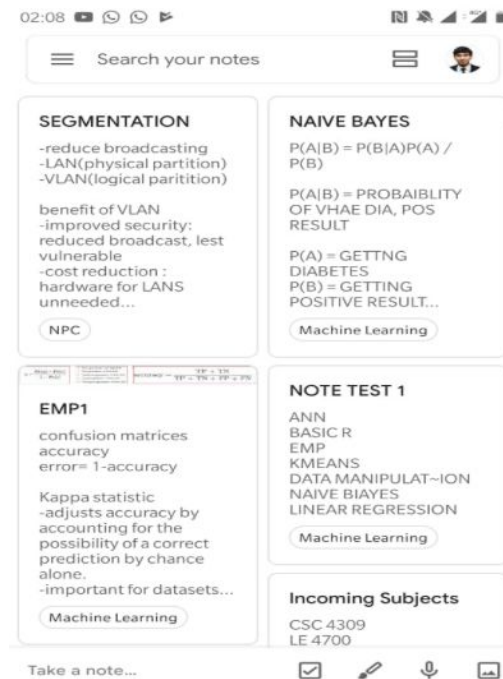
Figure 2.1.1.1 shows the desktop version of Google Keep. Google Keep also has introduced a dark mode feature to give an eye relief to the user compared to the light mode that may harm the user's eyes after a long day of staring at the electronic lights. It is convenient for users to straight away type in their notes to Keep as it keeps the jotting platform at a place that is easy to find. After that, users can add labels to the notes that are taken whenever they want. Regardless, before, in the middle or after taking down the note. In the jot taking of Keep desktop, users can include several types of note; word, picture and writing. But in the mobile platform, Figure 2.1.1.2., they enhanced several features in it due to sophisticated devices that led to a variety of data input.

The advantages of using Keep as your daily note driver is it is synchronized in a personalized Google account. Which is the main company for mobile, email, search

engines and services, makes it accessible anywhere and fully utilized in an operating system developed by Google; Android.

The backdraw of using the Keep, it is less organized compared to other note takers. The only mechanism that segregates the section is only singular labeling. Users might find it hard to search through the date or any other segregating elements. Luckily , they have a search panel for users to look for keywords.

Keep might have been one of the best note drivers and is the software that is used as a reference model. Keep does not have the features to be synchronous in a real time and simplifying note. That is one of the features that will be proposed to the note assist.

2) Evernote

Evernote is an app designed for notetaking, organizing, task management, and archiving. It is developed by the Evernote Corporation. The app allows users to create notes, which can be text, drawings, photographs, or saved web content. Notes are stored in notebooks and can be tagged, annotated, edited, searched, given attachments, and exported. It is cross-platform, for Android, iOS, macOS, and Microsoft Windows. It is free to use with monthly usage limits, and offers paid plans for expanded or lifted limits.

The great features for this app is a clipper from the website. Meaning, any useful article user found on the website can be clipped into Evernote and read later in offline mode. In addition, the interface of Evernote either in mobile or computer applications are neat and simple. The word tagging are very useful features in helping keeping notes

organized.
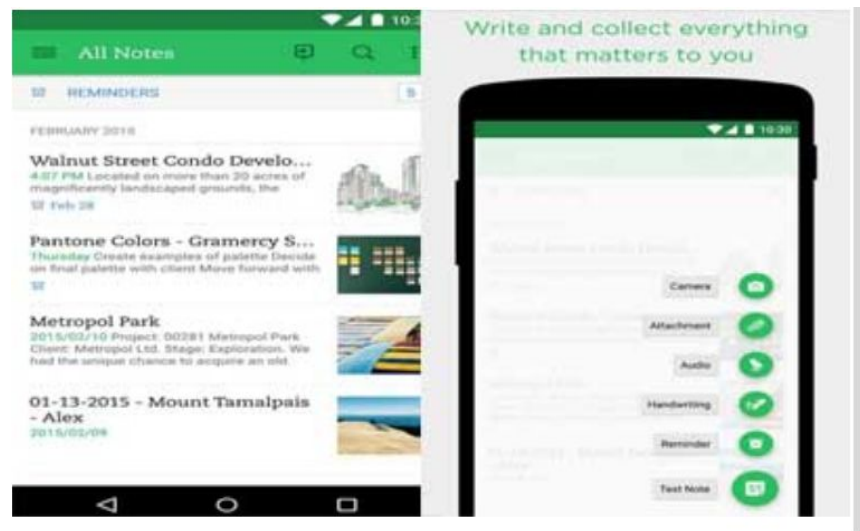
Figure 2.1.2.1 Evernote Mobile Version

The disadvantages of Evernote is the desktop version does not contain any feature for password to keep the notes from strangers but can be used in mobile phones. Other than that, there is no speech recognition in these apps. If this feature exists, this app will be very flexible and give freedom for users to use the notes.
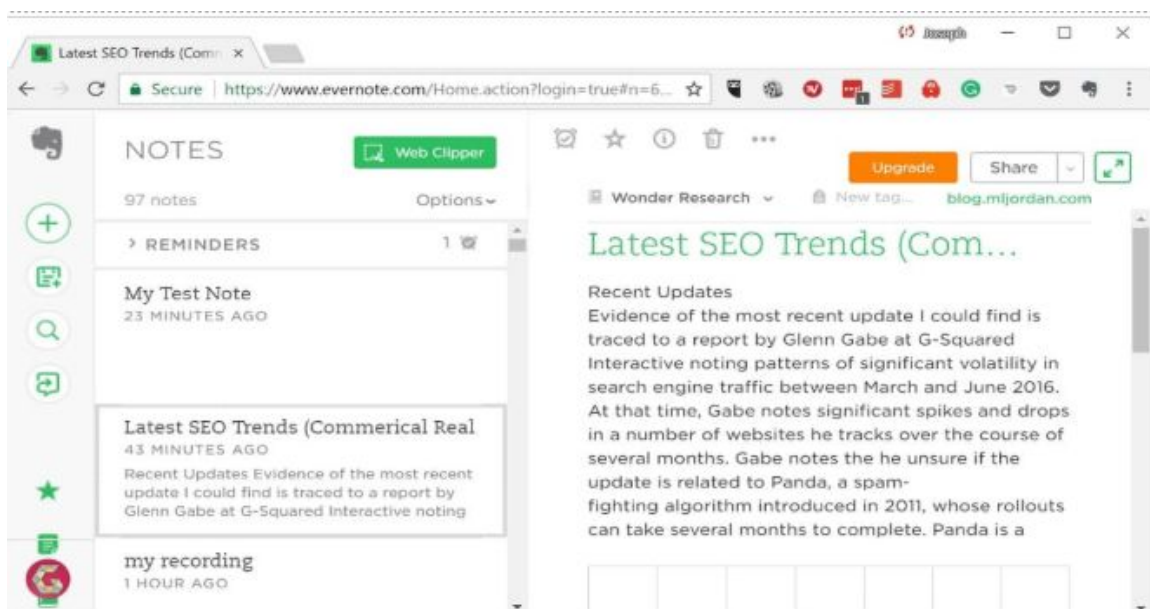
Figure 2.1.2.2 Evernote Desktop Version

3) Voice Notes

Voice Notes by Gawk is a note-taking app with voice recognition where your speech converts to text with one touch only. It can remind you later at the time you set where it integrated with calendars in your phone. Like others, it automatically saves notes into your storage files and easy backup to the cloud.

There are few advantages of Voice Notes; the interface of Voice Notes is easy to access and minimal. When we need to create a note, just click the create button and it will automatically start to detect your voice. These apps also allow the user to set reminders after creating the note. For example, "Feed cat" and there will be a calendar button to choose the date and time. Besides, the note created can be categorised accordingly. Voice Notes got the feature to export/import files which were able to save the files outside the apps or do backup. Below are the screenshot of the apps:
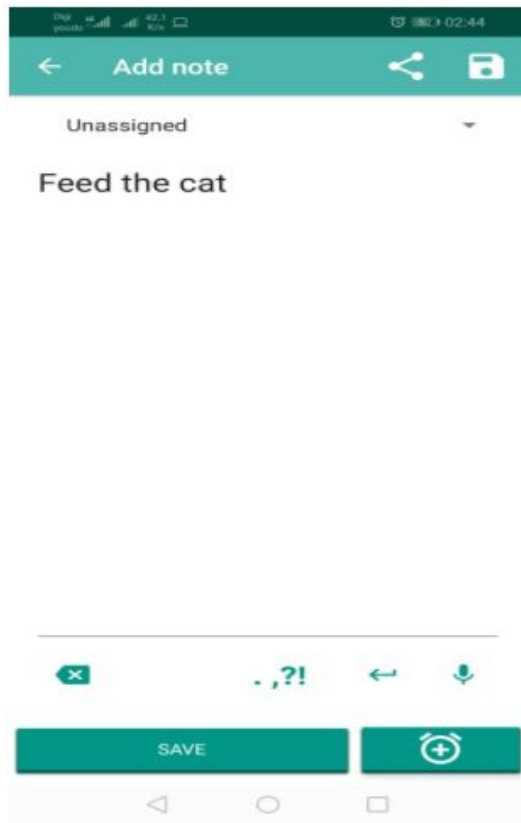
Figure 2.1.3.1 Voice Note Mobile   Figure 2.1.3.2 Voice Note Calendars

The drawback of this app is there is no automatic backup to the cloud. Other than that,

Voice Notes does not have a multiplatform such as web-based for users to access on

any devices.

2.2   Discussion of Review Previous Work

These applications have features that will ease the adaption of the note taking process.Furthermore, the features are:

1. Voice Note

An audio recording feature that will enable users to capture an audio and store it in the application for reference purposes.

2. Speech Recognition

Converting speech to text features that has the ability to capture audio and process it to text. This feature is one that will eliminate the process of note taking through user interactions.

3. Categorisation

The ability to categorise notes according to its subject from the study of the content of the note and suggesting users to label the note under the keyword.

These features are found in the best three applications mentioned. This literature also has produced another output, three most applications has that are crucial for note-taking adaptation are;

1. Sharing

Such feature is produced to enable users to collaborate with related parties to have an updated content of the note subject. It has saved time for delegating tasks through the net rather than the conventional meeting to combine every part from each party.

2. Offline mode

Most of the applications are available without the internet connectivity which enable users to store the notes even when there is an internet connectivity problem.

3. Multi-platform

Feature that enables users to access any platform that is available at the moment; website, through the laptops or personal computers, and mobile applications.

This project also aimed to produce a text summarization features. Few studies on the related algorithm has been done to optimize the algorithm for the mentioned feature.

Thus, all these three apps got their own pros and cons. The pros of these apps can be implemented in the system adaptation in our project while the cons of each app can be overcome by our upcoming software; Note Assist. The next feature can be added is text summarisation and grammar checker.

2.3 Summarization Modules, Keyword Extraction and Speech Recognition

A. Summarization Modules

Many summarization modules are created as the world becomes more advanced and more scientific studies are being done. During the development of the project, the python summarization modules used are Summa Summarization, Gensim Summarization, and Spacy Summarization. These modules serve as a summarization generator for the article or text input from users. Besides, they implemented the TextRank algorithm but come with different flavours of text pre-processing approaches such as removing different list of stopwords, performing POS tagging, and implementing tokenization. The test of each module has been carried out to compare the results of the best quality of summary (ie., the most important information in a document or article) for this web application.

a) Summa

Summa is a Python module that implements an algorithm which is a graph-based ranking model for text processing. The algorithm is called TextRank in which the implementation can result in text summarization and keyword extraction. Besides,

Summma requires two Python libraries for scientific computing which are, NumPy and Scipy. These two libraries will be installed along with Summa.

b)   Gensim Summarization

Gensim is an open-source Python library created by Radim Rehurek. A lot of quality documentation that can be accessed from its main website (www.radimrehurek.com). Gensim can provide a technique to extract underlying topics from rich documents. Other than that, summarization can also be implemented through Gensim where the concept algorithm used is the same as Summa, which is TextRank.

c)   Spacy Summarization

Spacy, also one of the popular open-source libraries, mainly focuses on the advancement of NLP in Python. It helps with the process of a large volume of text along with understanding the text itself. A lot of features and capabilities are taken into account such as Part-of-Speech (POS) Tagging,   lemmatization, text classifier and many more. For this development, Spacy is used as a summarizer. The summarizer implements the TextRank algorithm.

B.   HTML5 Web Speech Kit

Speech recognition is a common mechanism used in daily online routine. From subtitling movies to transcribing live speeches, it is used widely in most of today's

applications for several features. The ability of this feature to input raw audio data and process the audio into a text has made it relevant in most of the platforms. From time to time, the technology has been improvised to be able to capture audio and predict the word precisely.

HTML5 has been offering its very own speech recognition system. It is offered through Javascript and can be directly transferred to the front-end without depending on the back-end. This engine has tackled the hassle to process audio from the client side without having multiple requests from the server . However, Web Speech Kit has a limitation in its compatibility with the browsers. Only a few browsers are supporting Web Speech Kit, like Chrome, Opera and Firefox that have a different implementation method.

C.   Keyword Extraction module

Keyword extraction is an important component that goes along with the summarization. Python libraries are also provided with supervised and unsupervised approaches for keyword extraction. Supervised approach is less effective because it needs to train on a particular document while unsupervised approaches is more independent in which it does not rely on external corpus or dictionaries. Few existing unsupervised keyword extractor modules can be found such as RAKE, TextRank, TF.IDF, KP-Miner, and MultipartiteRank but all of these can not surpass the 'Yet Another Keyword Extractor!' (YAKE!) module.

YAKE! Module feels the gap for the existing keyword extractor modules where it provides multilingual online for single documents tool. Besides, the main features of this module are unsupervised approach, corpus-independent, domain and language independent, and single-document support. Thus, this module gives the meaningful keywords and suits best to do as reference for a single document. How the YAKE! module work will be discussed more under Chapter Three.

CHAPTER THREE

METHODOLOGY

For this project, the methodology used is Agile Model. This approach helps in software development where there will produce ongoing release cycles. Agile is a developing cycle that enables developers to update the development as the requirement changes. Suits for this project that the requirement may change from time to time. Below is the diagram of the agile model.

Agile model
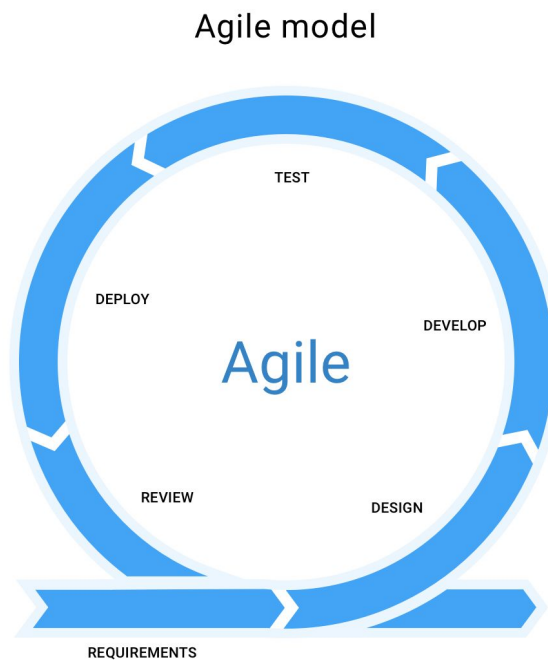
Figure 3.0 Agile Cycle

The methodology applied to this project is Agile system development cycle. Agile is put to frame as it is a cycle that focuses on customer collaboration. Note Assist is a product that is used for many human interactions; speaker and audience, teachers and

students or boss and secretary. Optimization of Note Assist needs the review in the customer perspective in every single feature that is available in application.

Starting from requirements stages, this project has collected the requirements and discovered insight of the project. Later, based on the acquired requirements , the designing process started. In this stage, the elements used in the development of this system drivers will be decided at this stage; what framework, database and user interface.

Then comes developing stages, the designs that are fabricated in the designing stages will be put to work here. Series of debugging to uncover the bugs and fixing it. In the testing stage, few tests will be run to the application, usability testing, UX testing and such. This is to discover inconvenience and elements that disturbs the features.

Deploy stage, once the application is ready for end users, it will be deployed to testers and open its access to special end-users segments. In the last stage, reviews are acquired through the UX testing in the previous stage. All discovered facts, knowledge or notes to improve the application will be brought to the initial stage, Requirement stage, and the process is started from the initial until the outcomes of the project satisfies all user requirements.

In Conclusion, Agile is very flexible in adapting changes to the applications. That is the factor it is suitable for Note Assist developmentation. From time to time, functionalities can be implemented stage by stage. Through this, a note taking application that meets all user requirements.

3.1   DESIGN

3.1.1   Home Page
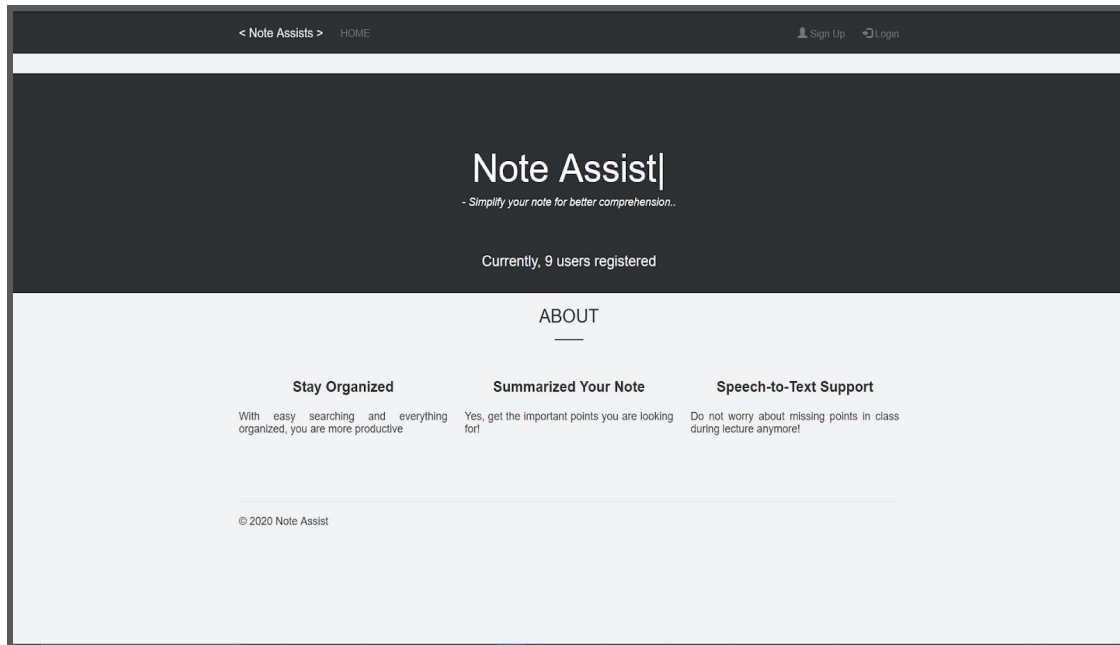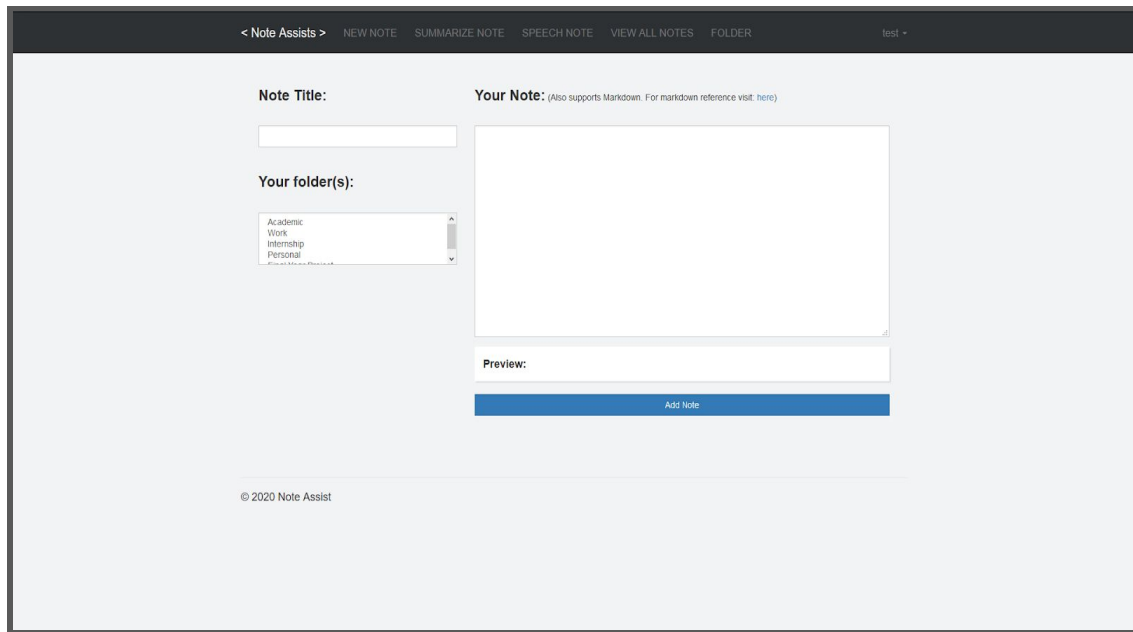
Figure 3.1.1 Home Page

  Figure 3.1.1 is the actual depiction of the home page for Note Assist. In this page, the user will have to sign in to access the notes stored and features: saving the notes and converting speech notes to text. Next is Add Note (figure 3.1.2) and Speech Note (figure 3.1.4) pages.

3.1.2 Add Note Page

Figure 3.1.2 Add Note Page

Both pages enable users to take note using various channels;–audio, media and typing. These features can be accessed through the page from Homepage (Figure 3.1.1), Add Note page (Figure 3.1.2).

3.1.3   Summarize Page

Figure 3.1.3 Summarize Page

Next is the Summarization Page (Figure 3.1.3) which will make use of Natural Language Processing technique to provide a summary for a text, an article or a speech. Few options have been provided to ease the process for users to use the features.

3.1.4 Speech Page

Figure 3.1.4 Speech Note Page

Speech Note page (Figure 3.1.4). Speech notes have another feature that enables the web application to convert speech in the form of audio file wav to text.

3.1.5 Folder Page
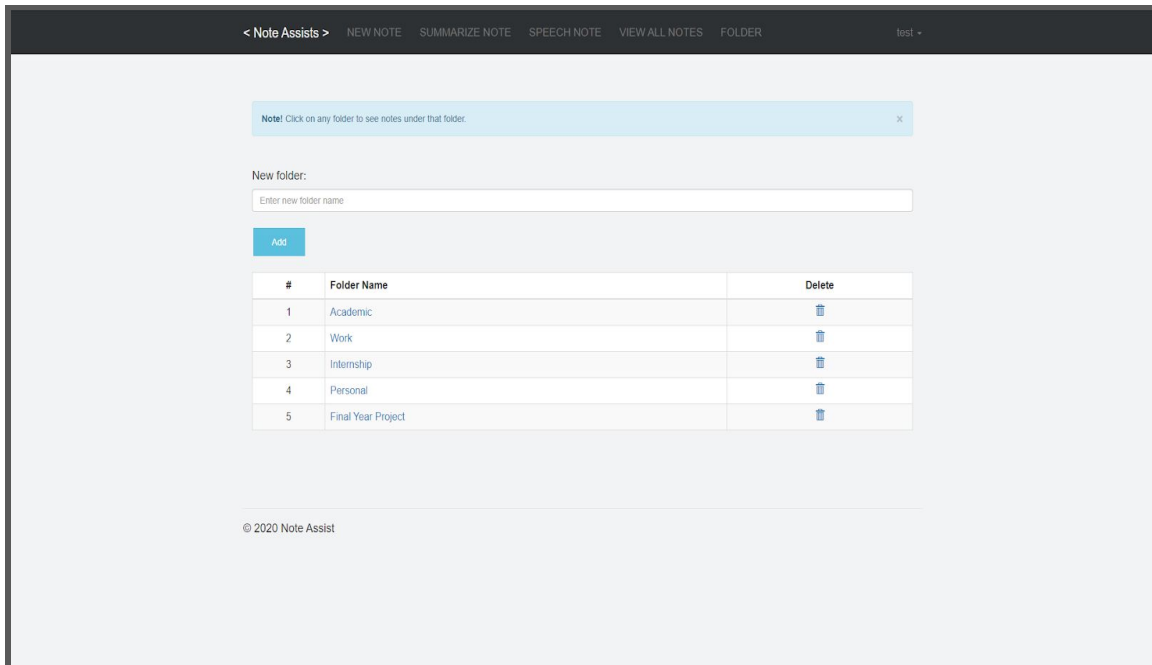
Figure 3.1.5 Speech Note Page

Figure 3.1.5 shows the tag page that allows users to browse notes according to the folder assigned.

3.1.6   View Note Page
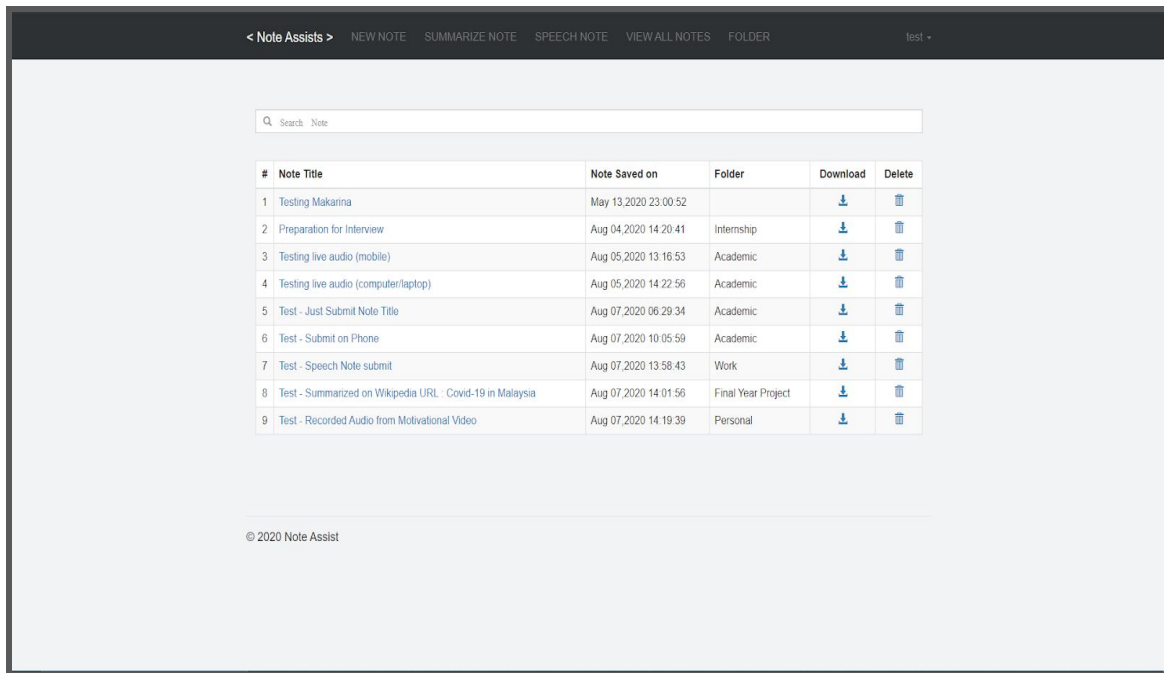
Figure 3.1.6 View Note Page

The *View Note Page* (Figure 4.5) is the dashboard for the users to access saved notes. Tagging feature is provided to tackle the hassle process of finding notes in a particular topic.

3.2   IMPLEMENTATION

3.2.1 Summarization and Keyword Extraction Algorithm

A. Summarization Technique

The text summarization has two types; extractive text summarization and abstractive text summarization.

a)   Extractive Text Summarization

Extractive summarization relies on extracting several parts from the text such as sentences, words and phrases. Later, gather these parts to create a summary.

b)   Abstractive Text Summarization

Abstractive summarization is more advanced in NLP in which this algorithm generates an entirely new summary and differs from the original text. It uses the encoder-decoder method.

For this project, Note Assist will be using the Extractive Text Summarization algorithm because it is more reliable and effective in extracting sentences and phrases from the piece of text without changing the meaning of the sentences. The specific algorithm for this extractive text summarization is called TextRank algorithm. Using the TextRank algorithm, the summary can be reduced into based on the ratio proportion of text, lines of the text or word counts. Following is the flow of the TextRank algorithm
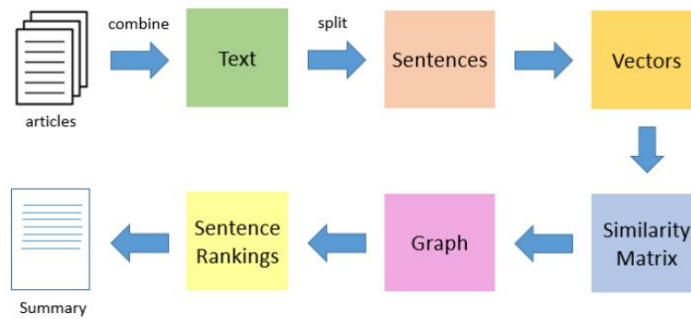
Figure 3.2.1.1 Flow of TextRank Algorithm

TextRank pipeline:

1. Concatenate all the text contained in the text.

2. Split the text into individual sentences.

3. Find vector representation for each and every sentence

4. Similarities between sentence vectors are then calculated and stored in a matrix.

5. Sentences as vertices and similarity score as edges are then converted into a graph for sentence rank calculation

6. A certain number of top-ranked sentences from the final summary.

The above summarization pipeline shows that it is an ideal technique to meet the objectives of the project where the TextRank algorithm allows the user to get the information needed in which the summary consists of the top-ranked sentences based on the calculated vectors and scores.

B. Keyword Extraction Technique

YAKE! Module has six main components to make the wonder happen and it allows generating keywords in different languages.
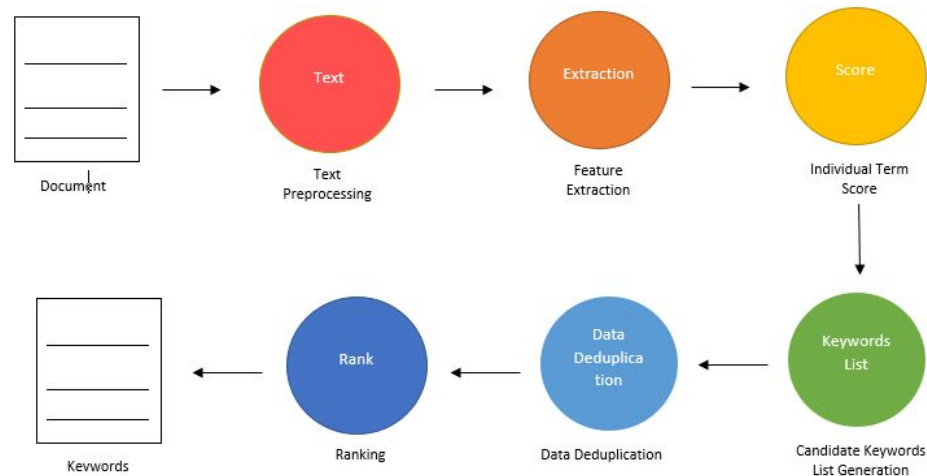
Figure 3.2.1.2 Flow of Keyword Extraction Pipeline

YAKE! (Keyword Extraction) Pipeline:

1. Text pre-processing

2. Feature Extraction

3. Individual Terms Score

4. Candidate Keywords List Generation

5. Data Deduplication

6. Ranking

First, the document will undergo text pre-processing where the purpose is to get cleaned individual terms. Thus, all unnecessary character and unmeaningful stopwords

will be removed such as brackets, period, white spaces, comma, and determiner from stopwords.

Second, the individual terms then are being processed for feature extraction. Feature extraction falls into five feature sets to determine the characteristics of each individual term. The sets are (a) Casing, (b) Word Positional, (c) Word Frequency, (d) Word Relatedness to Context, and (e) Word DifSentence. The Casing will get the aspect of every word. Word Positional determines the words that often occur at the beginning of a document and assumes that the relevant keywords will focus at the beginning of the document. Word Frequency will specify the frequency of every word. The fourth feature, Word Relatedness to Context, determines the number of different terms that occur to the left side of the extracted (candidate) word. The candidate word is meaningless when the number of the different terms often occurs with the candidate word. Finally, Word DifSentence measures how regularly a candidate word shows up inside various sentences.

The third component, Individual Terms Score, is all the features from the previous step that will be combined and computes scores for each term. Fourth, the weight of the scores of each term will generate the keywords based on the N-grams (unigram, bigram, and trigram). Fifth, any duplication of the candidates coming from the previous step will be eliminated using the Levenshtein distance and finally, the ranking will be based on the lower the score of candidate words, the more meaningful the word will be and formed by the N-grams.

Thus, by implementing these six main components, meaningful keywords can be generated from the YAKE! Module.

3.2.2 Speech-To-Text Algorithms

a) Python Speech Recognition

Python has been supporting speech recognition since long and it provides its very own speech recognition module that is easily accessed through python language. It grabs audio data in signal and later processes using a few modules. The process is simplified as below in figure.
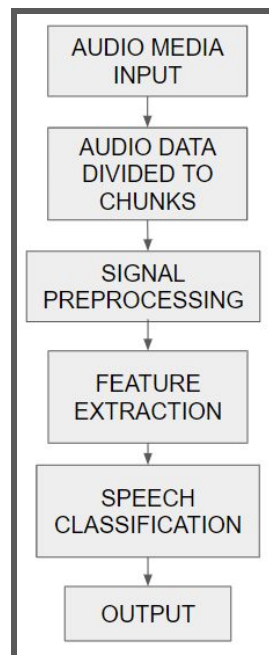
3.2.2.1 Speech Recognition Process

From the first stage, the audio input data will be captured and divided into chunks. This is to process captured data in a prompt manner and preprocess it chunks by chunks to

give a faster processing speed and output. Later, the data will be encoded and decoded in signal preprocessing and sent to the feature extraction process to recognize the words and classify the relevant word. Lastly it will output a transcription of the recognized speech. However, the drawbacks of this module are the dependency to PyAudio that disabled the ability to capture audio signal from cloud-based microphone input. The inability of this module to provide a synchronous and continuous data departure and arrival has also caused constraints to this project.

b) Web Speech Recognition

Web Speech is an HTML5 API that provides media access to client browsers and it's very own speech recognition system. It provides server-based services using the implementation of JavaScripts alongside HTML blocks. The process involves a complicated communication process between clients and servers in receiving the processed audio data. The process is simplified in figure 3.2.2.2.
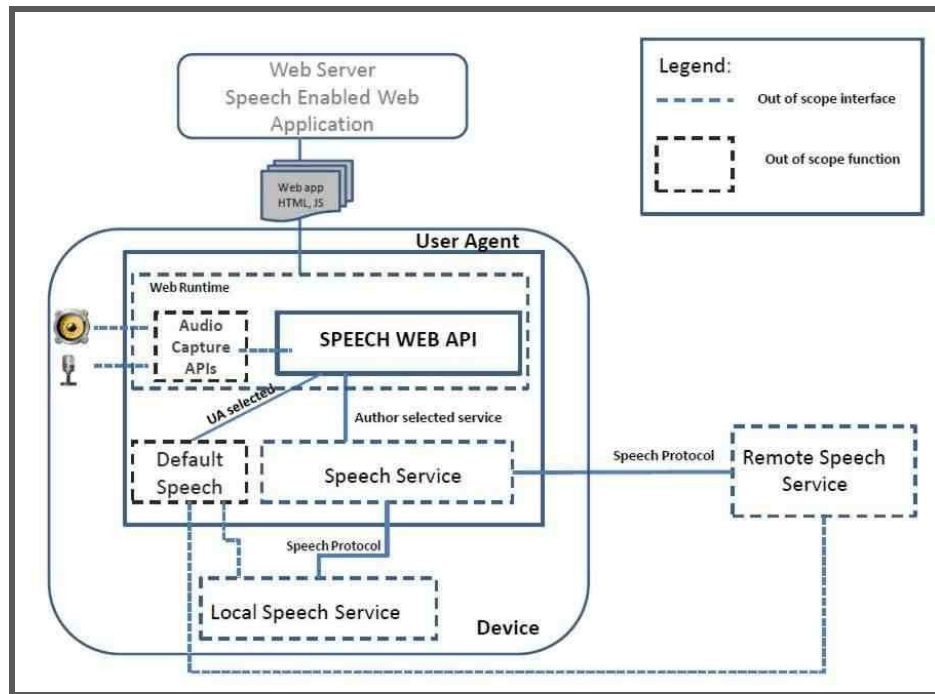
Figure 3.2.2.2 Web Speech Recognition Process

The process mostly runs at the client side from getting the audio signal to encoding and decoding. Later,it will be locally processed through the implemented api at the browser side. Then speech data will be sent to the server and redirected to the client side for the transcribed speech. The setbacks of this API are the accuracy of the transcribed speech and low security level at encrypting the audio data.

3.2.3 Flask Web Framework

Flask is a web framework that is based on python language. It is a common web application framework that is used by many developers. The framework maps the file and the page in the website and manages the requests between server and clients. It is considered as a microframework because of its independence towards particular tools or libraries.

3.2.4 Database

Note Assist implemented SQLite3 which is one of the database engines provided by Python. All the information from Note Assist such as username, notes will be stored in the database.

3.2.5 PythonAnywhere

PythonAnywhere is a web-hosting server that can focus on python projects. It is easy to configure and deploy any python web application. Note Assists is deployed under PythonAnywhere and can be accessed using the link `https://noteassist.pythonanywhere.com`

3.3   TESTING / EVALUATION

As the developed system has been done, it needs to be tested again and against the requirements to ensure the system has met the required objectives of the project. In this last phase of the testing, it was being conducted virtually online as the system has been deployed in the cloud to be accessed by the users and time constraint. We collected the feedback and responses from the tester using Google Form.

The testing has been done by a group of personnels that is inclusive from the target users. The invited tester comes from two lectures and five students to undergo the system testing. However, only one lecturer and three students accepted to conduct the testing.

The feedback collected from the tester ranges from the article objective, the summary relevancy, keyword match, main idea, word rearrangements and conveniences of the system. The results are depicted as follows.

1. Testing Summary.

Testing Summary
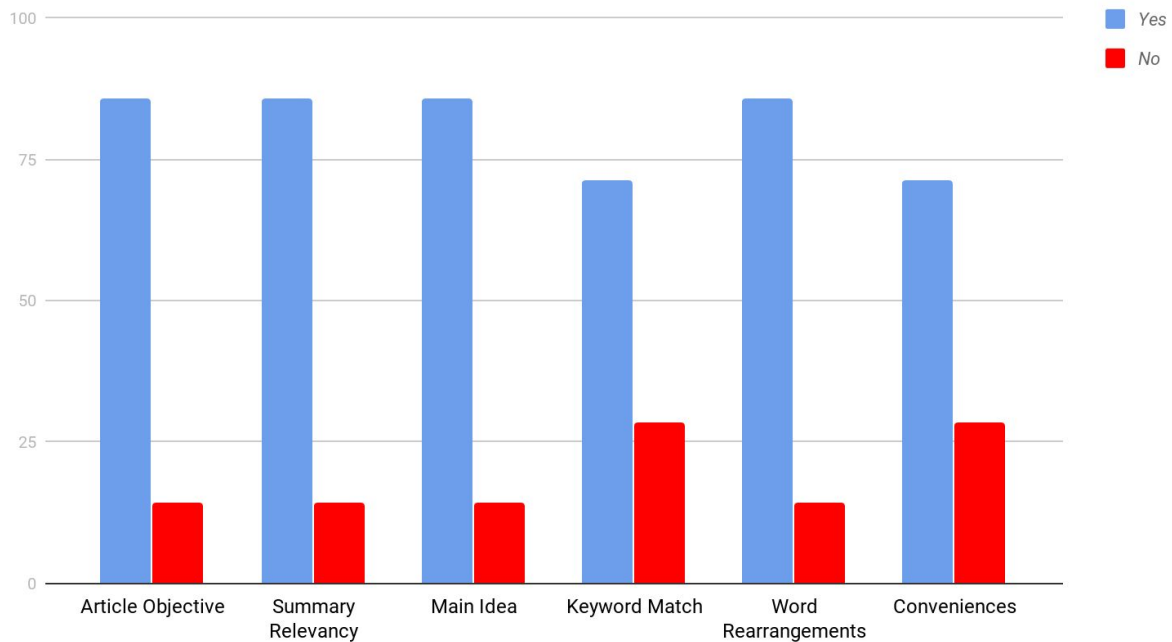
Figure 3.3.1 Testing Summary Chart.

From figure 3.3.1, it simplifies the summary of the testing conducted. This testing tasked the testers to test the summarization and speech recognition function on the website deployed in the cloud. The response reflected the article objective, summary relevancy,main idea and word rearrangements has a 85.7% of confidence. On the other hand, 72% of the testers agreed the keywords extracted from the article were accurate and the system is convenient to use.

2. Feedbacks

Figure 3.3.2.a Testers Feedback.
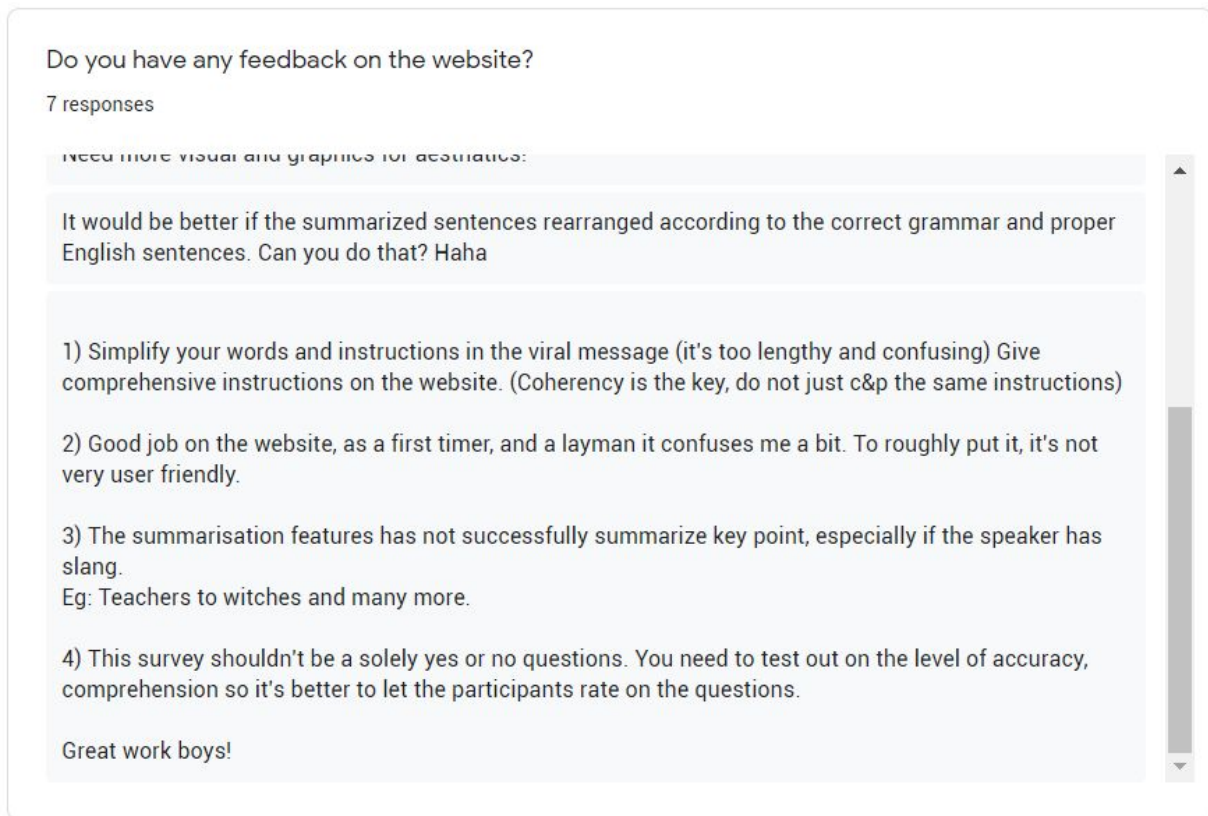
Figure 3.3.2.b Testers Feedback.

The testing has been conducted and the responses are inclined to show the requirements of the project are satisfied. The feedback was also collected and is shown in Figure 3.3.2a and b. The highlighted feedback concerned on the accuracy of the algorithm and also the user experience of the website. This will be taken note for future works.

CHAPTER FOUR

ANALYSIS OF THE RESULTS
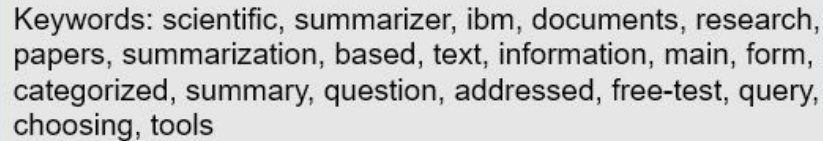
4.0 Results of Text Summarization and Keyword Extraction

The research, analysis and testing that are done on this project has brought good results. For the implementation of summarization modules, Summa and Gensim Summarization outperform Spacy Summarization in term of word preprocessing, sentences order and scores. Even though Gensim Summarization is on par with Summa, Summa still takes the lead where it provides information that is adequate to create a good summary and easy to configure specified percentage to cover for the summary from the original text or article. The summarization should cover 30% to 40% of the original text plus the summary at least consists of adequate and important information from the article. This unique feature is what makes Note Assist different from other similar note-taking applications. Thus, the Summa module is implemented in Note Assist.

Original Text	Summarized Text
[Word count : 210 words]	[Word count: 74 words]
The main question addressed by the research is how scientific documents can be summarized based on a given information need, either form of a free-test query or by choosing categorized values. The research work is relevant because nowadays there are a lot of scientific documents published as the world becomes more advanced and more scientific studies are being done. Summarization of scientific papers can give researchers an adequate amount of information in order to reduce their workload. There are many tools for text summarization but these tools only focus on more simple text and documents, unlike scientific papers that are more complex. The paper is well written and the text used is clear and easy to read. The paper provides a figure on how IBM Science Summarizer works. It also provides a clear image of the UI that IBM Summarizer used. During the human evaluation of IBM Summarizer, the researcher asked authors from the NLP community to evaluate the system. They were given two types of summaries which is a section-based summary and section-agnostic summary which means summarization without section. They calculate the average score given by the authors for each summary type across all papers. Based on the evaluation, the quality of the section based outperforms the section-agnostic summaries.	The main question addressed by the research is how scientific documents can be summarized based on a given information need, either form of a free-test query or by choosing categorized values. Summarization of scientific papers can give researchers an adequate amount of information in order to reduce their workload. There are many tools for text summarization but these tools only focus on more simple text and documents, unlike scientific papers that are more complex.

Figure 4.0.1 Text Summarization With Ratio 0.3

Figure above shows the text summarization generated from the implementation of Summa module in Note Assist. The original text is from the part of contents in a *'A Summarization System for Scientific Documents' Paper Review* And the summarized text is relevant with the original text. It meets the requirement to be considered as a summary which is short yet still contains significant information.

Keywords: scientific, summarizer, ibm, documents, research, papers, summarization, based, text, information, main, form, categorized, summary, question, addressed, free-test, query, choosing, tools

Figure 4.0.2 Keywords of the Original Text

Furthermore, the keywords will be generated from the original text along with the summarized text. Figure above shows that generated keywords are relevant with the summarized text and original text where all the keywords are meaningful and can be used as reference for users to look for desired keywords in the text.

4.1 Results of Transcribed Text Summarization

The Speech Note page in Note Assist provides users with the flexibility in using speech recognition functions such as users are able to convert audio files to text or capturing real-time speech. Note Assist doesn't have the ability to automatically summarize the speech but it can achieve the same goal through manually converting the audio file to text first and summarizing it through the Summarize Note page.

Audio-To-Text	Summarized Text
[Words count : 443]	[Words count: 143]
Symptoms of covid-19 before proceeding please note that this general overview is compiled for initial self-assessment only and may vary but each individual if you are not feeling well you should immediately consult a medical practitioner to have an accurate diagnosis and proper treatment. Building 135 patients from uniontown hospital m56 patients from wuhan pulmonary hospital the symptoms of broken down into. Dayton. Symptoms of similar to the common cold with a mild sore throat and neither having a fever no feeling tired patients can still consume food and drink as usual. The patient's throat painful body temperature bring it around 36.5 degrees c levittown, other symptoms like diarrhea. Cosmos series. Seems like feeling weak and joint pain start to manifest the patient may show a temperature reading between 36.5 degrees to 37° celsius. 6. The patient show a temperature reading above 37.2 degrees c the second most common symptom dry cough also of his. Breathing difficulty mayekar occasionally most patients in this stage are easily feeling tired of the symptoms remain about the same. Among the top five key indications of covid-19 according to the final report of the initial outbreak conducted by the joint commission of china and who. The patients that haven't started recovering by k7 get most serious call send breathing difficulty fever can get high up to 38 degrees celsius patients may develop a headache and body pain and diarrhea if there's any many patients are admitted to hospital at this day. The symptoms are likely to be worse in for the patient who has co-existing medical conditions. Breath becomes more frequent temperature reading goes well above 38 degrees in one of the study they know anything average time when sepsis does to affect 40% of patients. 211. Gold ring imaging test like chest x-ray to capture the severity of respiratory distress in patients patients adding loss of appetite and may be facing abdominal pain. Ulcer needs immediate treatment in icu. 214. The symptoms can be well-managed at this point steve attends to get better and breathing difficulties may start to see someday 13. May still be affected by a mild cough even after hospital discharge. Into 16 day 15 is the opposite condition for the rest of the minority patients the franchise group must prepare for the possibility of acute cardiac injury or kidney injury. Seems anointing. Teen fatality cases helping us around day 18. Vulnerable patients may develop a secondary infection caused by a new pathogen in the lower respiratory tract. Addition made in lead to blood coagulation and ischemia. C22 this involving patients are recovered completely from the disease i don't discharge from the hospital.	The patient's throat painful body temperature bring it around 36.5 degrees c levittown, other symptoms like diarrhea. Seems like feeling weak and joint pain start to manifest the patient may show a temperature reading between 36.5 degrees to 37° celsius. 6. The patient show a temperature reading above 37.2 degrees c the second most common symptom dry cough also of his. Breathing difficulty mayekar occasionally most patients in this stage are easily feeling tired of the symptoms remain about the same. The patients that haven't started recovering by k7 get most serious call send breathing difficulty fever can get high up to 38 degrees celsius patients may develop a headache and body pain and diarrhea if there's any many patients are admitted to hospital at this day. The symptoms are likely to be worse in for the patient who has co-existing medical conditions.

Figure 4.1.1 Transcribed Text Summarization with Ratio 0.3

Figure 4.1.1 portrays the transcribed text (audio-to-text) and summarized text. The audio is entitled *Recognizing Day to Day Signs and Symptoms of Coronavirus* in WAV format. It was then converted into text and summarized. Some of the words might be a bit off from the original audio but the contents are still understandable. But it will affect the summarized text where the actual meaningful terms might become a different word when the audio is converted to text. So, the result of the summarized text may not consist of all the main points. If this issue happens, users have the freedom to add or edit the text. In addition, the results of the summarized text is supported by the evaluation from the testing where the article objectives, summary relevancy, main idea,

and words rearrangement got the same percentage where 85.7 % answered 'Yes' and 14.3% answered 'No'.

Keywords: patients, symptoms, degrees, hospital, feeling, temperature, reading, pain, breathing, day, proceeding, proper, note, general, overview, compiled, self-assessment, vary, individual, immediately

Figure 4.1.2 Keywords of the Transcribed Text

The keywords of the transcribed text can be seen clearly in Figure 4.1.2 where the keywords provided are relevant with the transcribed text. Thus, the results and percentage from the evaluation satisfies the objectives of this project.

CHAPTER FIVE
CONCLUSION AND FUTURE WORKS

Based on the results, it is proven that the system can meet the project objectives successfully which are to provide related note taking applications as well as using them to ease the process by providing speech to text transcription function, summarization function and note organization.

There are lots of things that can be improved in the system as the system itself has not been developed before. First of all, the system could give a better performance by having a more powerful server and hosting site. To improve more, a cloud storage space would be able to afford heavy media for processing. Next, the speech recognition system can be more sophisticated that can produce a more accurate transcription and cancel surrounding noise. Lastly, the summarization can be enhanced by providing a more robust model to do the summarization process.

Note Assist is the first note taking web application that can do summarization for the notes taken. It can help the students and individuals in their lessons and also can be used as a daily driver tool for important points. For future works, the user interface and user experience can be improved, and the summarization algorithm can be further studied to produce a more accurate system.

REFERENCES

[1] Bijl, D., & Hyde-Thomson, H. (2001). *U.S. Patent No. 6,173,259*. Washington, DC: U.S. Patent and Trademark Office.

[2] Furui, S., Kikuchi, T., Shinnaka, Y., & Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing*, *12*(4), 401-408.

[3] Jing, Hongyan. "Sentence Reduction for Automatic Text Summarization." Proceedings of the Sixth Conference on Applied Natural Language Processing -, 2000, doi:10.3115/974147.974190.

[4] Stinson, M. S., Elliot, L. B., Kelly, R. R., & Liu, Y. (2009). Deaf and hard-of-hearing students' memory of lectures with speech-to-text and interpreting/note taking services. *The Journal of Special Education*, *43*(1), 52-64.

[5] Balcerzak, B., Jaworski, W., Wierzbicki, A.: Application of textrank algorithm for credibility assessment. In: Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01. pp. 451{454. WI-IAT '14, IEEE Computer Society,Washington, DC, USA (2014), http://dx.doi.org/10.1109/WI-IAT.2014.70

[6] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

[7] Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. Information processing & management, 33(2), 193-207.

[8] Adorf, J. (2013). Web speech API. *KTH Royal Institute of Technology*.

[9] Mihalcea, R., Tarau, P.: "Textrank: Bringing order into texts". In: Lin, D., Wu, D. (eds.) Proceedings of EMNLP 2004. pp. 404–411. Association for Computational Linguistics, Barcelona, Spain. July 2004.

[10] Barrios, F., López, F., Argerich, L., Wachenchauzer, R.: "Variations of the Similarity Function of TextRank for Automated Summarization". Anales de las 44JAIIO. Jornadas Argentinas de Informática, Argentine Symposium on Artificial Intelligence, 2015.

[11] Campos, R., Mangaravite, V., Pasquali, A., Jatowt, A., Jorge, A., Nunes, C. and Jatowt, A. (2020). YAKE! Keyword Extraction from Single Documents using Multiple Local Features. In Information Sciences Journal. Elsevier, Vol 509, pp 257-289

[12] Campos, R., & Mangaravite, V., & Pasquali, A., & Jorge, A., & Nunes, C., & Jatowt, A. (2018). A Text Feature Based Automatic Keyword Extraction Method for Single Documents. In Gabriella Pasi et al. (Eds.), Lecture Notes in Computer Science - Advances in Information Retrieval - 40th European Conference on Information Retrieval (ECIR'18). Grenoble, France. March 26 – 29. (Vol. 10772(2018), pp. 684 - 691). *

APPENDICES

A. Gantt Chart

NO	TASK	DURATION	SEM 1 19/20														SEM 2 19/20													
			W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14
1	Idea & Proposal	2W																													
2	Abstract	1W																													
3	Gantt Chart	1W																													
4	Literature Reviews	4W																													
5	Previous Work Reviews	4W																													
6	Projects Synopsis	2W																													
7	1st Implementation	1W																													
8	Methodology Planning	3W																													
9	Existing Design Reviews	2W																													
10	Flowchart	1W																													
11	Interface Design Mockup	2W																													
12	Algorithm Development	4W																													
13	1st Prototype	3W																													
14	1st Prototype Testing	2W																													
15	Quality Check	1W																													
16	Review Prototype Design	1W																													
17	FYP1 Poster	1W																													
18	FYP1 Showcase	2W																													
19	Interim Report	7W																													
20	Development	4W																													
21	Final Prototype	1W																													
22	2nd Implementation	3W																													
23	Final Prototype Testing	2W																													
24	Result Documentation	2W																													
25	Technical Report	2W																													
26	Preliminary Report	2W																													
27	FYP 2 Showcase	2W																													
28	Final Report & Poster	2W																													

B. Testing Form Responses

Note Assist Evaluation

Assalamualaikum and greetings,

We are Undergraduate Computer Science Students of International Islamic University Malaysia (IIUM). Therefore, we are currently conducting a study to assess our Final Year Project on our accuracy of summarization website.

Website: https://noteassist.pythonanywhere.com/

Google Drive: https://drive.google.com/drive/folders/1WqkuXrGkJaZWsfpepbcGjMjeg3TEtmiI?usp=sharing

Instructions:
1. Click the Google Drive link, go to 'Video' folder to watch the video // listen to the audio in the 'Audio' folder for the same content.

2. Go to Audio-To-Text folder -> Download the file based on the title video you have watched (Recommended to read the contents first).

3. Please go to the website.

4. Register / Login using id:test pwd:123

5. Go to summarization page.

6. On "Option 1", upload the file that you have downloaded and click the Summarize button.

7. Evaluate the summary in the form below.

8. Feel free to browse around the website!

Notes:
**In instruction 2, you can actually get the Audio-To-Text from our Speech Note page but for this evaluation purpose, we have done it for you and save your time. But if you want to test the audio-to-text function;
1. Download any audio from the 'Audio' folder in GDrive.
2. Go to Speech Note Page
3. Upload the audio and click convert (The audio only supports .wav format at the moment)
4. May take a while to convert based on the connection.

**In Google Drive link, there are also sample articles provided that you can test in our Summarization page (not compulsory)

The summary give an objective outline of the whole piece of the article/video.   *

○ Yes

○ No

The summary include any examples, details, or information from which are relevant from the article/video. *

Questions   Responses **7**

The keywords match the content from the original text or video. *

◯ Yes

◯ No

The summary include the main idea of each paragraph, and the main evidence supporting that idea *

◯ Yes

◯ No

The words rearrangement in the summary are keeping the same structure of the context *

◯ Yes

◯ No

Do you find it easy to use our summarization? *

◯ Yes

◯ No

Do you have any feedback on the website? *

Long answer text