Kulliyyah of Information and Communications Technology
International Islamic University Malaysia
Semester II 2019/2020

CSC 4309 Natural Language Processing
Section 1

Capstone Project

Newspaper Articles Summarization

Lecturer: Dr. Suriani Sulaiman

Name	**Matric No.**
Haziq Iskandar bin Suriani	1628259
Izzat Asyraf bin Haniz	1624865
Ahmad Zulfahmi bin Harum	1626867
Hifdzul Hadi bin Daud	1620509

Introduction

Natural language processing (NLP) is a wide study field covering the combination of linguistics, artificial intelligence, and computer science. NLP is one of the famous ways for machines to understand and analyze human language. In fact, Google is also using NLP for the use of words translation, their keyword search ability and others. On the internet, there are billions of data and information that can be searched and read. There are many types of data that have been published on the internet including texts, videos, images and many more. Anyhow, text is the most basic thing that we can see on any website. However, not all parts in the text are useful and not important for us. Thus, the most effective way to save our time from reading unimportant parts in a text is by summarizing it.

Text summarization is one of the branches in NLP, it is a process of extracting the information from a text and summarizing it into a small chunk of text. There are two types of text summarization which are abstractive summarization and extractive summarization. Abstractive summarization is concerned about the semantic understanding of every word in a text. While extractive summarization weighs the important part of a sentence by ranking it with scores. In this project, the process of text summarization is done by using NLTK and it is categorized as extractive summarization. NLTK is an open source library in Python programming language, that provides many language processing modules. There are several important steps in text summarization by using NLTK which create a frequency table for the words, split the text into words word by word, finding the score and lastly generate a summary for the text.

The capability of technologies nowadays is so incredible that they can do almost all the tasks that humans do. The evolution of NLP and machine learning has served a lot of good implications to us as its analyzing ability is sometimes more consistent, critical and done in an unbiased way.

Problem Statement

Newspaper is a document that contains news that is happening around the world. The news might consist of information that is not important and useful to us. Thus, it is taking more time to digest the important information by reading the whole news.

Objectives of Project

- To determine the important information from the newspaper articles.
- To generate a summary that contains information relevant with the newspaper articles.

Methodology

Summarization with NLP has two approaches, Abstractive summarization and Extractive summarization. Extractive summarization identifies the significant and meaningful information from the text while Abstractive summarization is a deep learning where it tries to understand the semantic of the articles and the summary might change entirely but still relevant.

Thus, this project is implementing the Extraction summarization approach where the definition meets the objective of the project in which identifies the important points from the newspaper articles. The summary will be generated from the top rank of sentences that have the high sentence scores.

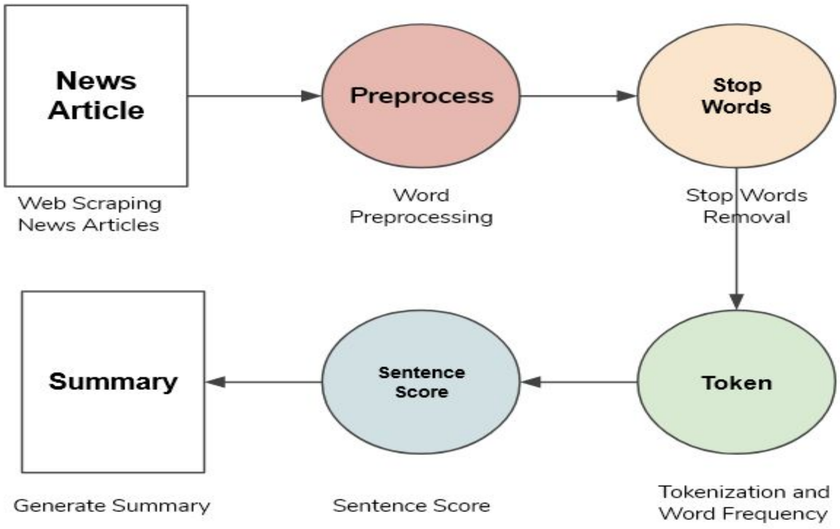


Figure 1 Text Summarization Pipeline

Text Summarization Pipeline

- **Word Preprocessing**

Initial stage would require a bit of data cleaning and preprocessing. This stage is important in order to run a smooth process and high accuracy system. The raw data text that is imputed may be dirty and contains noise elements. In this project, the data are processed and cleaned by removing square brackets and extra spaces in the text that are considered as unnecessary elements. Later it would remove the special characters and digits that will cause uncertainties in the sentences and words scores. By using this, data can be processed in a smooth way without analyzing unintended characters and elements.

- **Stop Words Removal**

Stop Words removal is part of the data cleaning process. The process will remove any characters that are indicated as a stop word in the existing corpus: ‘,’,, ,. , ! and more. In this case, the stopwords corpus used here is the English stop words from NLTK library. All the stop words will be removed and it will increase the efficiency to process the text imputed.

- **Tokenization**

The cleaned data text will be split into words and undergo the process of stemming where every word that contains affixes to suffixes and prefixes will be removed. The next step is to create frequency for every term. By creating the frequency of every occurrence of the terms, the system can determine the significance of the term. The higher the frequency, the more meaningful the term to be in the summary.

- **Sentence Score**

This process needs to determine the score for every sentence in the newspaper or articles. In achieving the score sentence, the sentences from articles will be tokenized first and the value of the sentences is by its own words frequency which these tokenized words frequency are obtained from the previous step. To get the sentence value, if the words are in the sentences, the frequency value of the tokenized words will be summed up to create the sentence value for every tokenized sentence. Then, the score of the sentence is calculated with the formula of sentence value divided by the total words of the sentence.

- **Generate Summary**

After the sentence scores are created, a summary will be generated by combining all the top rank of the sentence scores. To determine the rank of sentence scores, there is a need to compute an average sentence score where the condition to get the ranking is the sentence scores have to be greater than the average sentence score. Together with these, a summary from articles or newspapers can be produced.

- **Web Scraping News Article from Newspaper website**

There are several ways in scrapping news from various sources which is to write scraping code for each website or use a package. For this project *newspaper3k* package is used to automatically extract structured information from various news websites. This package can extract several information from websites such as the authors, publish date, text and top image. For this project, text of the news article is extracted and downloaded. Then, the text is passed to the text summarization pipeline.

- **System GUI Interface**

The system interface is made using *tkinter* package ("Tk interface"). *tkinter* is the standard Python interface to the GUI toolkit. This project consists of two frames which is the upper frame which includes the text input box for users to paste the news article link and a button to summarize the news. The lower frame will display the summarized news article. Figure below shows the system interface for this project.

Tkinter provides various controls, such as buttons, labels and text boxes used in GUI application. These controls are commonly called widgets. For this project, some of the widget is used to create the GUI interface which are:-

- Canvas:-
    - The Canvas widget is used to draw shapes, such as lines and other shapes in the application. Canvas is used as a background for this project.
- Frame:-
    - The Frame widget is used as a container widget to organize other widgets. For this project frame is used to organize the layout of the button, user input field and also textbox for the summarized news article.
- Label:-
    - The Label widget is used to provide a single-line caption for other widgets. It can also contain images. For this project, Label is used to display the title of the system which is located at the top and also is used for displaying summarized news articles.
- Button:-
    - The Button widget is used to display "Summarize News" buttons in this project .

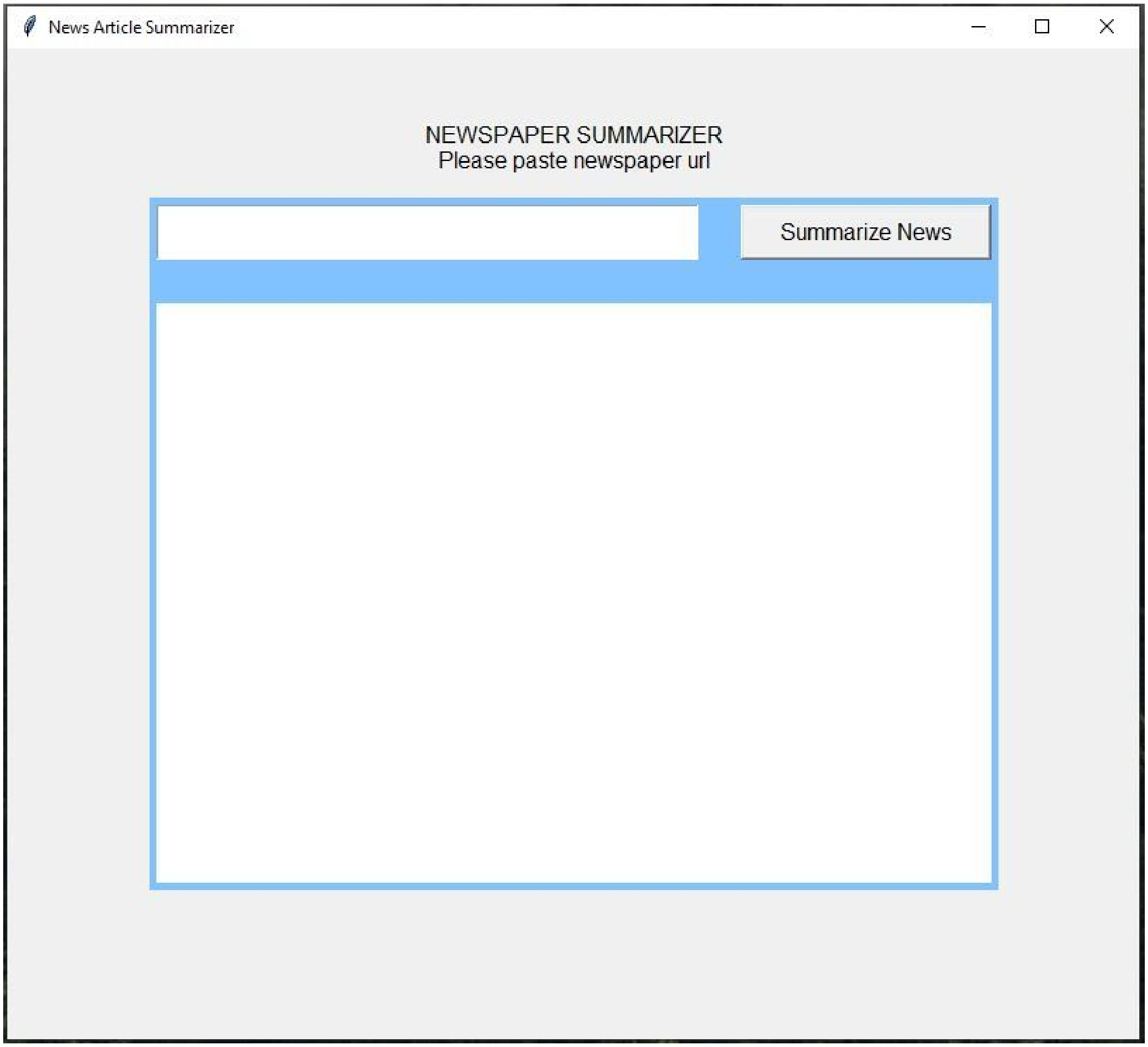
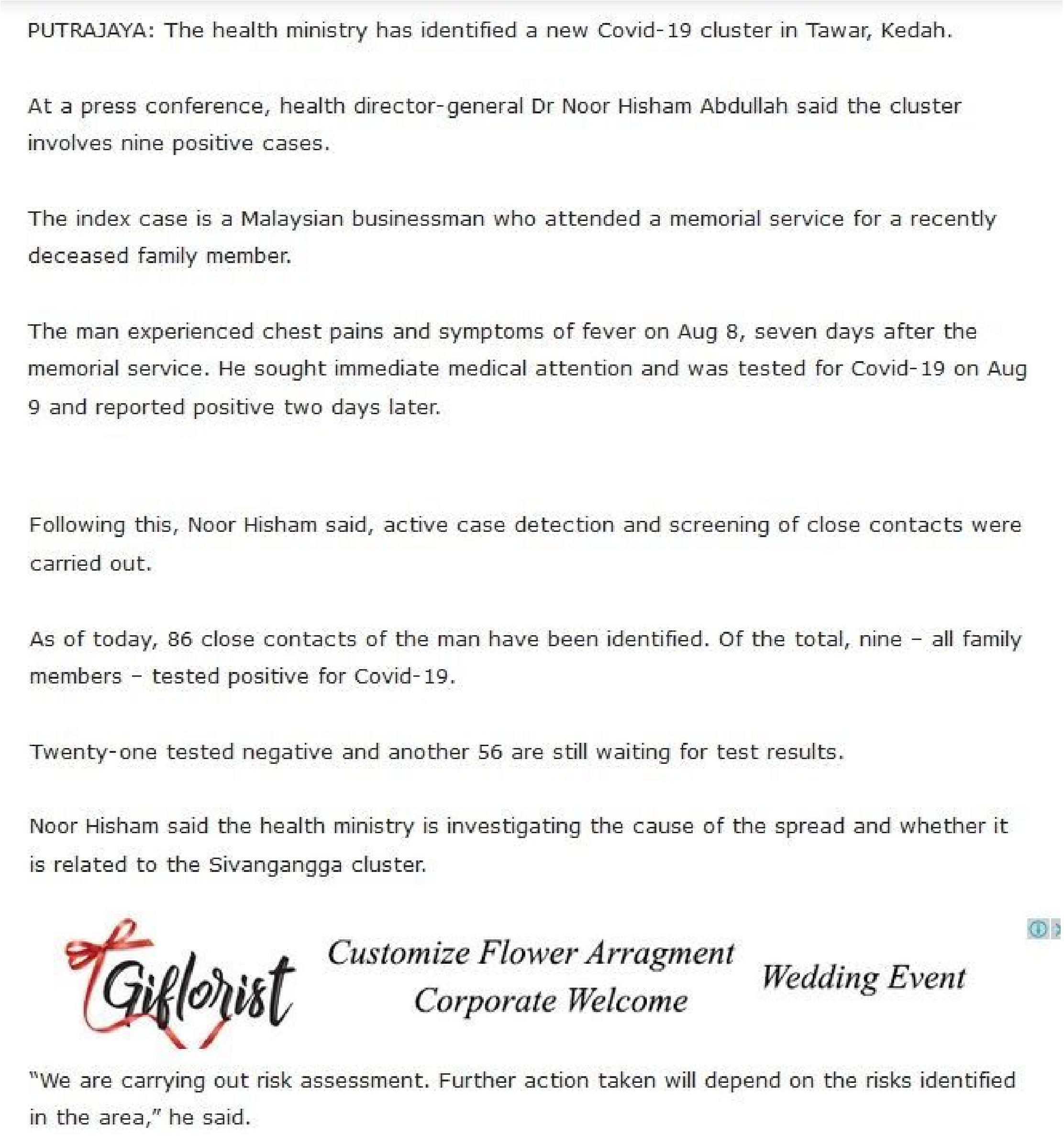


Figure 2 News Article Summarizer interface

Results

Original News Article link:
https://www.freemalaysiatoday.com/category/nation/2020/08/13/new-cluster-detected-in-kedah-with-9-positive-cases/

Before summarization, the original news article from the website contains 1184 characters.



PUTRAJAYA: The health ministry has identified a new Covid-19 cluster in Tawar, Kedah.

At a press conference, health director-general Dr Noor Hisham Abdullah said the cluster involves nine positive cases.

The index case is a Malaysian businessman who attended a memorial service for a recently deceased family member.

The man experienced chest pains and symptoms of fever on Aug 8, seven days after the memorial service. He sought immediate medical attention and was tested for Covid-19 on Aug 9 and reported positive two days later.

Following this, Noor Hisham said, active case detection and screening of close contacts were carried out.

As of today, 86 close contacts of the man have been identified. Of the total, nine – all family members – tested positive for Covid-19.

Twenty-one tested negative and another 56 are still waiting for test results.

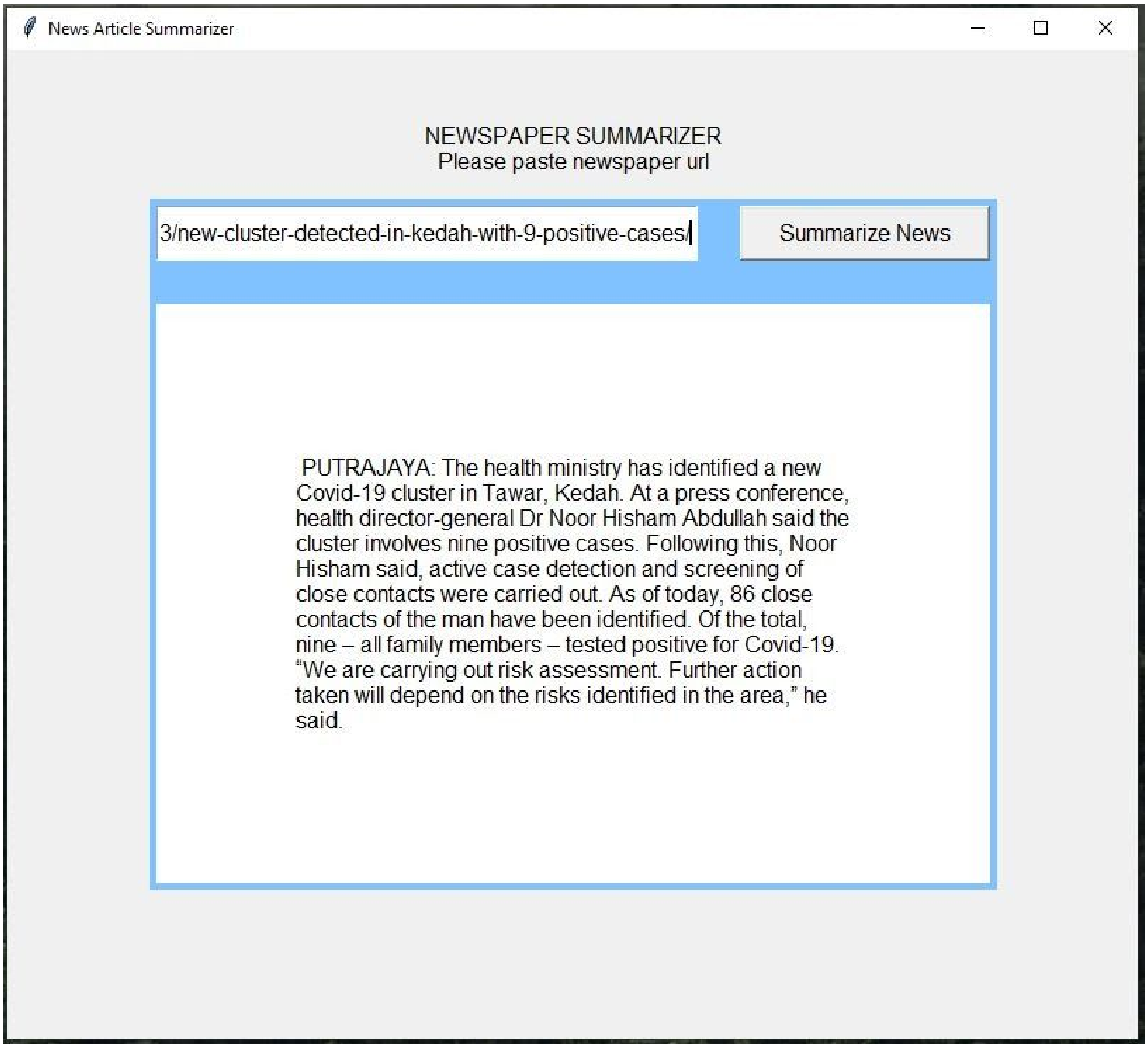
Noor Hisham said the health ministry is investigating the cause of the spread and whether it is related to the Sivangangga cluster.

"We are carrying out risk assessment. Further action taken will depend on the risks identified in the area," he said.

Figure 3 Original news article

Summarized News Article:

After summarization using NLTK, the summarized news article contains 564 characters.

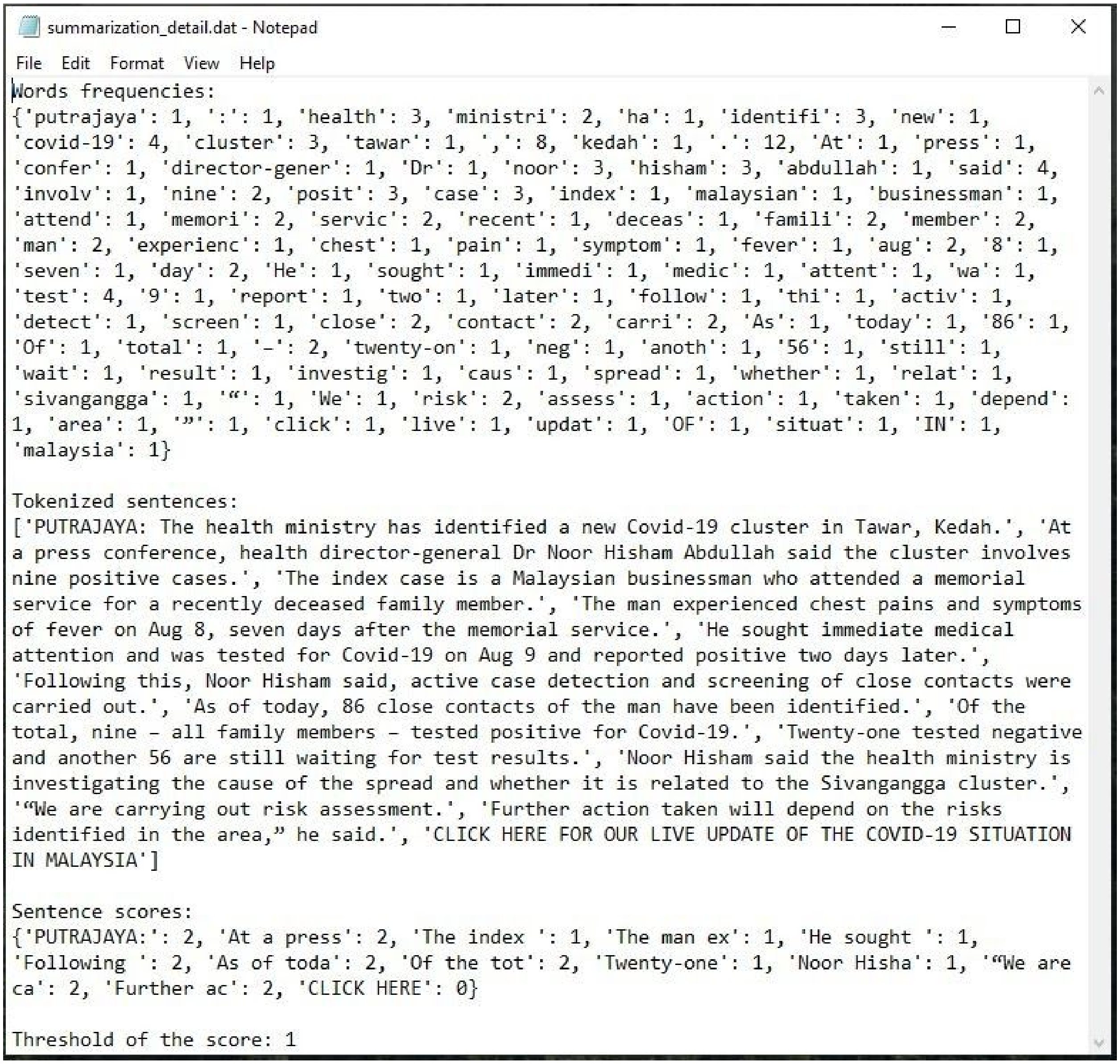


3/new-cluster-detected-in-kedah-with-9-positive-cases/

PUTRAJAYA: The health ministry has identified a new Covid-19 cluster in Tawar, Kedah. At a press conference, health director-general Dr Noor Hisham Abdullah said the cluster involves nine positive cases. Following this, Noor Hisham said, active case detection and screening of close contacts were carried out. As of today, 86 close contacts of the man have been identified. Of the total, nine – all family members – tested positive for Covid-19. "We are carrying out risk assessment. Further action taken will depend on the risks identified in the area," he said.

Figure 4 Summarized news article

Figure 5 below shows the output of the summarization process which are word frequencies, tokenized sentences, sentence score and also Threshold of the score.



summarization_detail.dat - Notepad
File Edit Format View Help

```
Words frequencies:
{'putrajaya': 1, ':': 1, 'health': 3, 'ministri': 2, 'ha': 1, 'identifi': 3, 'new': 1,
'covid-19': 4, 'cluster': 3, 'tawar': 1, ',': 8, 'kedah': 1, '.': 12, 'At': 1, 'press': 1,
'confer': 1, 'director-gener': 1, 'Dr': 1, 'noor': 3, 'hisham': 3, 'abdullah': 1, 'said': 4,
'involv': 1, 'nine': 2, 'posit': 3, 'case': 3, 'index': 1, 'malaysian': 1, 'businessman': 1,
'attend': 1, 'memori': 2, 'servic': 2, 'recent': 1, 'deceas': 1, 'famili': 2, 'member': 2,
'man': 2, 'experienc': 1, 'chest': 1, 'pain': 1, 'symptom': 1, 'fever': 1, 'aug': 2, '8': 1,
'seven': 1, 'day': 2, 'He': 1, 'sought': 1, 'immedi': 1, 'medic': 1, 'attent': 1, 'wa': 1,
'test': 4, '9': 1, 'report': 1, 'two': 1, 'later': 1, 'follow': 1, 'thi': 1, 'activ': 1,
'detect': 1, 'screen': 1, 'close': 2, 'contact': 2, 'carri': 2, 'As': 1, 'today': 1, '86': 1,
'Of': 1, 'total': 1, '–': 2, 'twenty-on': 1, 'neg': 1, 'anoth': 1, '56': 1, 'still': 1,
'wait': 1, 'result': 1, 'investig': 1, 'caus': 1, 'spread': 1, 'whether': 1, 'relat': 1,
'sivangangga': 1, '“': 1, 'We': 1, 'risk': 2, 'assess': 1, 'action': 1, 'taken': 1, 'depend':
1, 'area': 1, '”': 1, 'click': 1, 'live': 1, 'updat': 1, 'OF': 1, 'situat': 1, 'IN': 1,
'malaysia': 1}

Tokenized sentences:
['PUTRAJAYA: The health ministry has identified a new Covid-19 cluster in Tawar, Kedah.', 'At
a press conference, health director-general Dr Noor Hisham Abdullah said the cluster involves
nine positive cases.', 'The index case is a Malaysian businessman who attended a memorial
service for a recently deceased family member.', 'The man experienced chest pains and symptoms
of fever on Aug 8, seven days after the memorial service.', 'He sought immediate medical
attention and was tested for Covid-19 on Aug 9 and reported positive two days later.',
'Following this, Noor Hisham said, active case detection and screening of close contacts were
carried out.', 'As of today, 86 close contacts of the man have been identified.', 'Of the
total, nine – all family members – tested positive for Covid-19.', 'Twenty-one tested negative
and another 56 are still waiting for test results.', 'Noor Hisham said the health ministry is
investigating the cause of the spread and whether it is related to the Sivangangga cluster.',
'“We are carrying out risk assessment.', 'Further action taken will depend on the risks
identified in the area,” he said.', 'CLICK HERE FOR OUR LIVE UPDATE OF THE COVID-19 SITUATION
IN MALAYSIA']

Sentence scores:
{'PUTRAJAYA:': 2, 'At a press': 2, 'The index ': 1, 'The man ex': 1, 'He sought ': 1,
'Following ': 2, 'As of toda': 2, 'Of the tot': 2, 'Twenty-one': 1, 'Noor Hisha': 1, '“We are
ca': 2, 'Further ac': 2, 'CLICK HERE': 0}

Threshold of the score: 1
```

Figure 5 Summarization process result

Conclusion

In conclusion, this project successfully generates summarization of a text by using NLTK module in Python. As shown in the result, the summary generated is only filled with relevant and important information. NLTK seems to have a very effective way in summarizing text. Any website's link that has text in it can be used in this text summarization program. The scores of every sentence calculated is the core value that is used to generate this type of summarization technique. The summary then is the result of arranging the sentences according to its score. Apart from that, text summarization actually promotes time efficiency especially for readers and researchers. This is because this technology reduces one's reading time and speeds up the process of analyzing an article or text.

Reference

Nurfikri, F. (2020, May 15). Web Scraping News with 4 lines using Python. Retrieved August 14, 2020, from https://towardsdatascience.com/scraping-a-website-with-4-lines-using-python-200d5c858bb1

Tkinter - Python interface to Tcl/Tk¶. (n.d.). Retrieved August 14, 2020, from https://docs.python.org/3/library/tkinter.html