# Text Summarization Using Natural Language Processing

Tanveer Mahmood Hasan
BCS, Information and
Communication Technology,
International Islamic University
Gombak, Selangor, Malaysia
tanveer.iium@gmail.com

Jaki Fayek Alvi Rahman
BCS, Information and
Communication Technology,
International Islamic University
Gombak, Selangor, Malaysia
fayek.lucse44@gmail.com

MD Raisul Islam
BCS, Information and
Communication Technology,
International Islamic University
Gombak, Selangor, Malaysia
mdraisulislam48@gmail.com

Liu Yufei
BCS, Information and
Communication Technology,
International Islamic University
Gombak, Selangor, Malaysia
arthas1061476137@gmail.com

## ABSTRACT

The method of spontaneously generating and condensing the shape of a given document is text summarization. Summarizing significant text documents manually is a daunting job for humans. Extractive and abstractive are the two main methods of summing up text. Though extractive summary is primarily concerned with what summary content the frequency words, phrases and sentences from the original document should be used. This research paper proposes a tf-idf model to summarize from large articles to give the reader a clear vision about that particular article.

**KEYWORD:**

Text Summarization, Tf-Idf, Reuters, Sentence positions, F-1 Score, Rouge score, Blue score, Extractive.

## I.INTRODUCTION

The goal of summarization is to capture the important information contained in large volumes of text, and present it in a brief, representative, and consistent summary. A well written summary can significantly reduce the amount of work needed to digest large amounts of text on a given topic. It is a challenging task for humans to summarize manually substantial documents of text. Text summarization is the process of spontaneously creating and condensing the form of a given record and safeguarding its information source into a shorter adaptation with by and large importance. Nowadays text summarization is one of the most favourite research territories in natural language processing and could attract more attention from NLP researchers.

In this project we introduce an application of the Tf-Idf model to the problem of multi-document summarization. TF-IDF, short for term frequency-inverse document frequency, is a numeric measure that is used to score the importance of a word in a document based on how often it appeared in that document and a given collection of documents. The intuition behind this measure is: If a word appears frequently in a document, then it should be important, and we should give that word a high score. But if a word appears in too many other documents, it is probably not a unique identifier, therefore we should assign a lower score to that word. Our experiments will show that improve the performance of a state-of-the-art summarization framework which strongly indicate the benefits of the TF-IDF model for this task.

## II. RELATED WORK

The extractive based summarization using K-Means Clustering with TF-IDF (Term Frequency-Inverse Document Frequency) for summarization. The research paper also reflects the idea of true K and using that value of K divides the sentences of the input document to present the final summary. In the experiment, used the K-means and TF-IDF method for extractive text summarization with K value predefined following the two well known methods that are used to get widespread for true K determination. It is very clear that the statistical measures approach results in best output and researchers used two types of comparison evaluation to so. One thing that makes clear during all these experiments that before summarizing the online crawled or corpus texts should give proper look to the preprocessing step and make sure that all unnecessary characters, keywords, tags, and punctuations [1]. Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm. This research proposes a sentence based clustering algorithm (K-Means) for a single document. For feature extraction, this article used Gensim word2vec which is intended to automatically extract semantic topics from documents in the most efficient way possible. In this research model, all the sentences were clustered using the K-Means clustering algorithm. Sentence scoring algorithm rates a sentence based on the occurrence of numerical values and nouns. These techniques were implemented on BBC news article datasets. The model showed the best performance on the business articles because the business article contains more numerical values and the sentence scoring algorithm gives priority to numerical values [2]. The use of continuous vector representations for semantically aware representations of sentences as a basis for measuring similarity also evaluate different compositions for sentence representation on a standard dataset using the ROUGE evaluation measures. These research experiments show that the evaluated methods improve the performance of a state-of-the-art summarization framework and strongly indicate the benefits of continuous word vector representations for automatic summarization. The results of this paper show great potential for employing word and phrase embeddings in summarization [3].

## III. TECHNICAL BACKGROUND:

### 1. NLTK:

NLTK, which refers to "Natural Language Toolkit", is a leading platform for building Python programs to work with human language data. It is a suite of libraries and programs for symbolic and statistical NLP for English written in the Python programming language. As the applications, NLTK supports classification, tokenization, stemming, tagging, parsing, and semantic reasoning functionalities.

In our text summarization model, we have such packages like nltk corpus which are reuters and stopwords, nltk stemming which is PorterStemmer, and function on bleu scores.

### 2. Reuter Corpus:

Corpus is an important segment inside NLTK libraries, it is a single large collection of text used in linguistics research, may be in the form of written or spoken material. Furthermore, corpus provides grammarians, lexicographers, and researchers in NLP with better descriptions of a language. There are several famous corpus like British National Corpus (BNC), containing 100 million word collection of samples of written and spoken language from a wide range of sources, and Brown Corpus, containing 500 samples of English-language text compiled from works published in the United States in 1961, etc.

In our model, we choose Reuter Corpus. Reuter Corpus is a large collection of Reuters News stories for use in research and development of natural language processing, information retrieval, and machine learning systems. It has three available corpus: "Reuters Corpus, Volume 1' (RCV1), "Reuters Corpus, Volume 2' (RCV2), "Thomson Reuters Text Research Collection" (TRC2). Differently, RCV1 is written in a single English language while RCV2 is in thirteen languages (Dutch, French, German, Chinese, Japanese, Russian, Portuguese, Spanish, Latin American Spanish, Italian, Danish, Norwegian, and Swedish). The stories are not parallel but are written by local reports in each language.

### 3. Position of Sentence:

Generally, the position of sentences in a paragraph or an article have a great impact on extracting the main ideas from the paragraph or the article.

Therefore, it is frequently used by automatic summarization systems as an attribute for sentence selection. Especially, sentences close to the beginning of the document are supposed to deal with the main topic and thus are selected for the summary. This criterion has shown to be very effective when summarizing some types of documents, such as news items.

In our model, we borrow the values from a user's github account for helping us assign the values between 0 to 1 on each sentence based on their different position in an article. We assume that sentences at the very beginning and ends of the article have a higher weight.

**4.Tf-Idf:**

To measure an individual term, we have the tf-idf approach. "tf" means "term frequency", measuring how frequently a term occurs in a document while "idf" means "inverse document frequency" and measures how important a term is. The formula are below:

$$tf = \frac{Number\ of\ times\ term\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

$$idf = \frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}$$

In the given article. Sentences are scored as the sum of their TF-IDF word frequencies. Other than tf-idf, we add our own methods to assist the operation of the model. These are, we assign each sentence a score between 0 to 1 based on the percentage of words common to the headline, assign a score based on how close the sentence's length is to the ideal length, assign a score corresponding to the sentence's position in the article. In short, we have four elements that need to be considered: tf-idf, headline, length and position.

**5.Evaluation:**

The application of Bilingual Evaluation Understudy (BLEU) is essential in our testing. As known, BLEU is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Its score is for comparing a candidate translation of text to one or more reference translations. For each category length of sentence, we have 1-gram, 2-gram, 3-gram and 4-gram with an average weight which is 25% respectively, followed by the final BLEU scores.

What is more, we also have rouge packages in our model. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. Rouge scores are for measuring length-3, length-5, and length-8 testing. Also we have ROUGE-1, ROUGE-2, ROUGE-l for each approach of summary. ROUGE-1 and ROUGE-2 is under ROUGE-N classification. And ROUGE-N refers to overlap of N-grams between the system and reference summaries. ROUGE-l indicates Longest common subsequence problem taking into account sentence level structure similarity naturally and identifies.

## IV. METHODOLOGY

In this project we are trying to summarize the text by using real data articles. So, we collect some news and articles from different online sources. Moreover, this summarization can give more or less accuracy to analyse the real outcomes. To avoid less accuracy, we endeavour to collect more articles or news.

1. **Data Collection :**

   Articles are collected from malaymail and nst. Article is uploaded by Keertan Ayamany on 20th january 2020.

2. **Tools:**

   Jupyter Notebook and Google Colab.

3. **Model Creation:**

   To create our model we are using some approaches such as position of sentence, top down approaches, retaive corpus, tf-idf method etc, for text summarization. We use a top down approach, because the description of the framework is formulated in a top-down manner, defining but not describing any first-level subsystems. Tf-idf weight is often used in text mining which is a statistical measure to evaluate how important a word is to a document. We take an article and split it into sentences and words. Then we measure some impact factors and score them. Based on that score our system summarizes a whole document.

4. **Workflow:**

   We are using the tf-idf method to create a model for text summarization. We start our work retrieving articles or news and then we extract headline and body from the document. After that, we check the length of the sentences whether it is empty or not. If the article contains below than 5 sentences the model will print the existing article as it is. If the article contains more than five sentences the model will continue with two different processes. First one is stemming and tokenization. And the other one is average scoring. Inside these two processes there are also some sub processes that our system will consider such as tf-idf model, Splitting sentences, Length score, Sentence position and so on. After all of these our system summarizes the whole document and prints the result and save In a file. Then we evaluate our system using Rogue Score and Bleu Score.

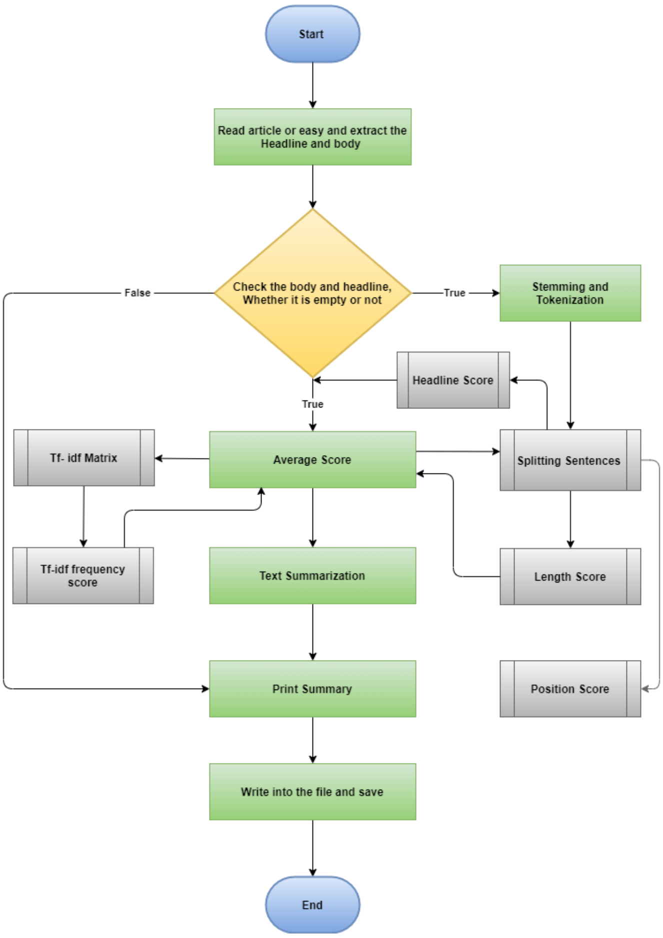
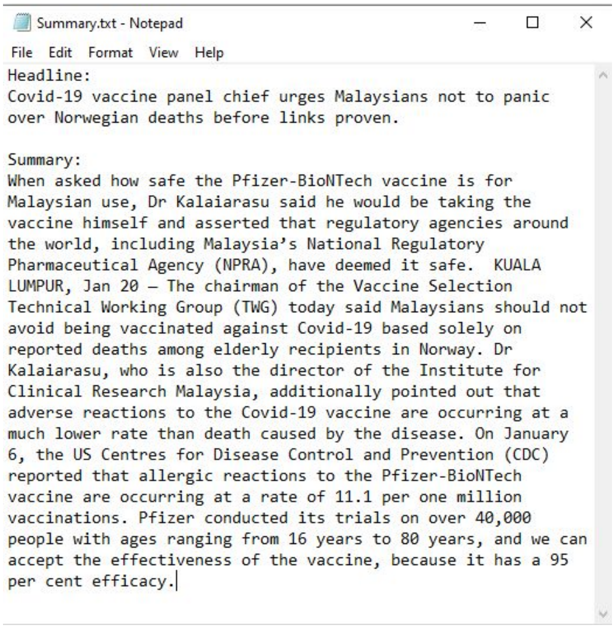

Start
Read article or easy and extract the Headline and body
Check the body and headline, Whether it is empty or not
False
True
Stemming and Tokenization
Headline Score
True
Tf- idf Matrix
Average Score
Splitting Sentences
Tf-idf frequency score
Text Summarization
Length Score
Print Summary
Position Score
Write into the file and save
End

*Figure 1: Project Flowchart*

## V. RESULT ANALYSIS

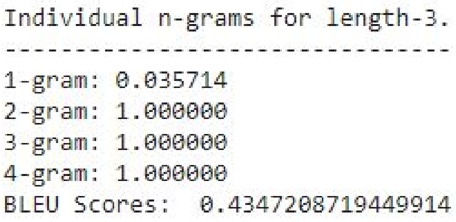
After pushing our selected article into Summarizer class to test using our tf-idf model, we got the summarized result of the given article as shown in Fig. 2 below.



```
Headline:
Covid-19 vaccine panel chief urges Malaysians not to panic
over Norwegian deaths before links proven.

Summary:
When asked how safe the Pfizer-BioNTech vaccine is for
Malaysian use, Dr Kalaiarasu said he would be taking the
vaccine himself and asserted that regulatory agencies around
the world, including Malaysia’s National Regulatory
Pharmaceutical Agency (NPRA), have deemed it safe.  KUALA
LUMPUR, Jan 20 — The chairman of the Vaccine Selection
Technical Working Group (TWG) today said Malaysians should not
avoid being vaccinated against Covid-19 based solely on
reported deaths among elderly recipients in Norway. Dr
Kalaiarasu, who is also the director of the Institute for
Clinical Research Malaysia, additionally pointed out that
adverse reactions to the Covid-19 vaccine are occurring at a
much lower rate than death caused by the disease. On January
6, the US Centres for Disease Control and Prevention (CDC)
reported that allergic reactions to the Pfizer-BioNTech
vaccine are occurring at a rate of 11.1 per one million
vaccinations. Pfizer conducted its trials on over 40,000
people with ages ranging from 16 years to 80 years, and we can
accept the effectiveness of the vaccine, because it has a 95
per cent efficacy.
```

*Figure 2: Model Result*

To evaluate our model result we used Quilbot, a summarizer web tool, to compare the summary results.

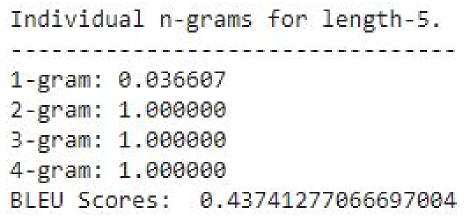


```
Individual n-grams for length-3.
--------------------------------
1-gram: 0.035714
2-gram: 1.000000
3-gram: 1.000000
4-gram: 1.000000
BLEU Scores:  0.4347208719449914
```

*Figure 3: BLEU score of Summary of length 3*

Fig. 3 shows the individual n-gram and BLEU scores of summary of length 3. Individual n-gram is the evaluation of just matching grams of specific order, such as single words (1-gram) or word pairs (2-gram or higher). And BLEU score is the function for evaluating a candidate sentence against one or more reference sentences. According to our result for length 3, for the 2-gram, 3-gram and 4-gram our summarizer gives an outstanding output which is exactly 1. That means the accuracy is the same whereas the individual 1-gram score is very low compared to others. Consequently, the final outcome of BLEU scores results as 0.4347. This happens because the website (Quilbot) uses machine learning algorithms in their text summarizer.

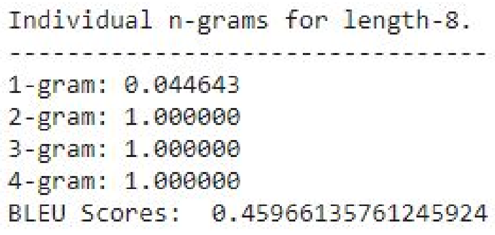
We also tried to tweak between length 5 and 8. Surprisingly length 5 shows better BLEU scores of 0.4374 shown in fig. 4.



```
Individual n-grams for length-5.
--------------------------------
1-gram: 0.036607
2-gram: 1.000000
3-gram: 1.000000
4-gram: 1.000000
BLEU Scores:  0.43741277066697004
```

***Figure 4:  BLEU score of Summary of length 5***

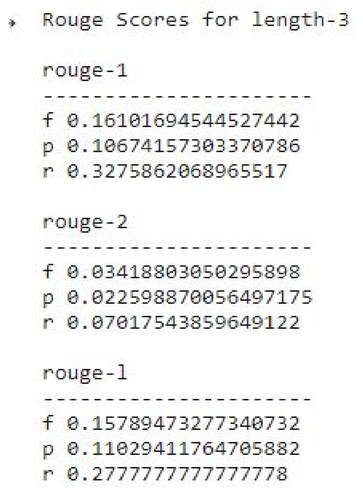
Whereas, the result for length is much better compared to others, which is 0.4596.



```
Individual n-grams for length-8.
--------------------------------
1-gram: 0.044643
2-gram: 1.000000
3-gram: 1.000000
4-gram: 1.000000
BLEU Scores:  0.45966135761245924
```

***Figure 5: BLEU score of Summary of length 8***

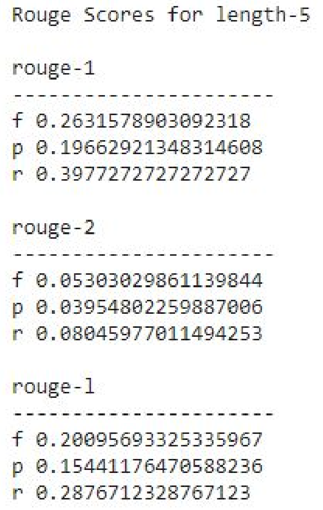
To get the F1-score, Precision and Recall we used Rogue scores. Rogue-1 is for unigram, Rogue-2 is for bigram and Rogue-l is for the longest sequence of matching words.



```
Rouge Scores for length-3

rouge-1
----------------------
f 0.16101694544527442
p 0.10674157303370786
r 0.3275862068965517

rouge-2
----------------------
f 0.03418803050295898
p 0.022598870056497175
r 0.07017543859649122

rouge-l
----------------------
f 0.15789473277340732
p 0.11029411764705882
r 0.2777777777777778
```

***Figure 6: Rogue scores for length-3***

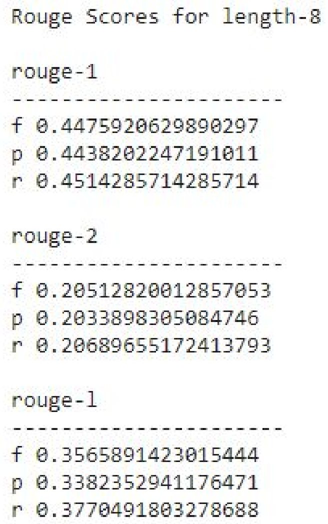
F1-score is the measure for accuracy which is way below our expectation level for length 3. Also precision which is the number of overlapping words divided by total words in the system summary and recall which is the number of overlapping words divided by total words in reference summary are very low.



```
Rouge Scores for length-5

rouge-1
----------------------
f 0.2631578903092318
p 0.19662921348314608
r 0.3977272727272727

rouge-2
----------------------
f 0.05303029861139844
p 0.03954802259887006
r 0.08045977011494253

rouge-l
----------------------
f 0.20095693325335967
p 0.15441176470588236
r 0.2876712328767123
```

***Figure 7: Rogue scores for length-5***

In fig. 7, we get the result of Rogue scores where summary length of the articles (Model result summary and Quilbot result summary) is 5.



```
Rouge Scores for length-8

rouge-1
----------------------
f 0.4475920629890297
p 0.4438202247191011
r 0.4514285714285714

rouge-2
----------------------
f 0.20512820012857053
p 0.2033898305084746
r 0.20689655172413793

rouge-l
----------------------
f 0.3565891423015444
p 0.3382352941176471
r 0.3770491803278688
```

***Figure 8: Rogue scores for length-8***

In fig. 7, we get the result of Rogue scores where summary length of the articles (Model result summary and Quilbot result summary) is 8. After comparing the results for length 3,5 and 8, we can see that length-8 has comparably better scores than others which is around 50%.

## VI. EVALUATION

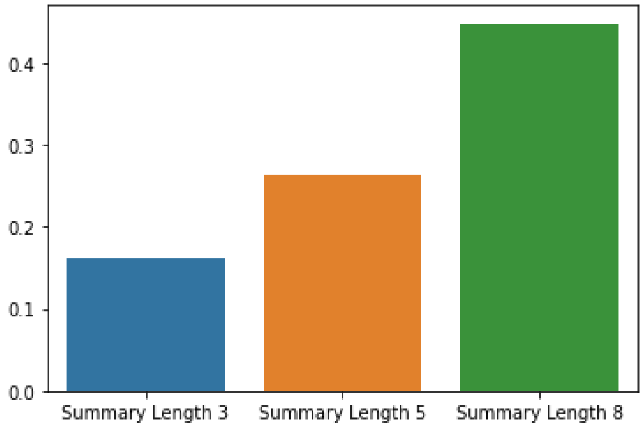


*Figure 9: F1-scores of Rogue-1*

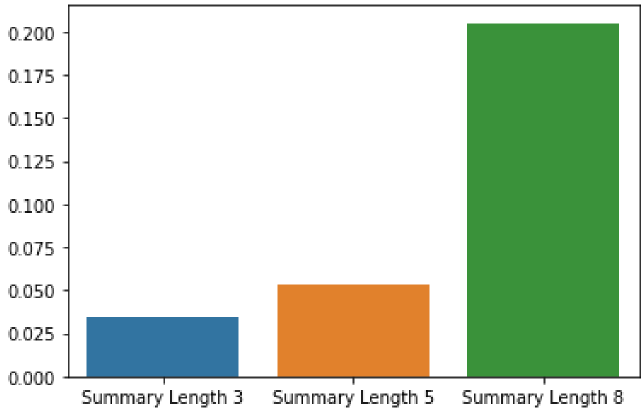


*Figure 10: F1-scores of Rogue-2*

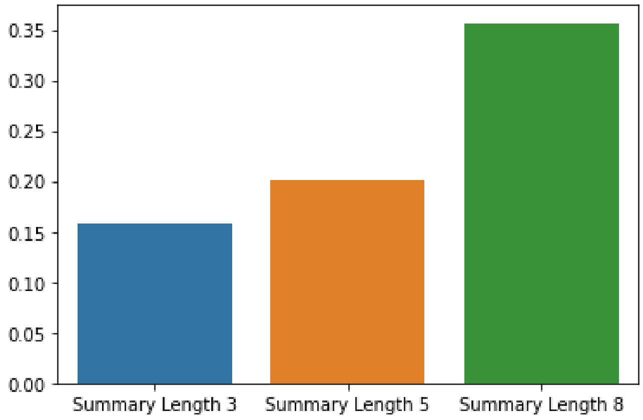


*Figure 11: F1-scores of Rogue-l*

According to the bar-plots in fig. 9,10 and 11, summaries of length-8 always get the highest scores.

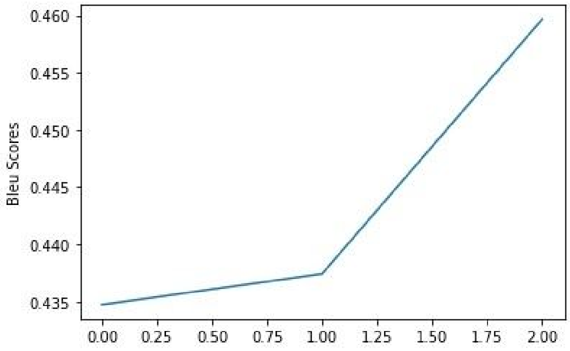


*Figure 12: Line-plot of BLEU scores*

Our model performs well when the summary length is 8 compared to length 3 or 5. We know that, higher the BLEU score, the better the model is.

## VII. CONCLUSION

According to the above analysis of BLEU Scores and Rouge Scores we can conclude that our model works best for summary length 8. In the Rouge scores summary length 8 is also ahead of summary length 3 and 5. Thus, we can consider our model is best for 8 lines summary. In the future, we will try to implement machine learning algorithms in our model to upgrade our model as well as performance and also to work on multiple articles simultaneously.

## VIII. REFERENCES

Khan, R., Qian, Y., & Naeem, S. (2019). Extractive based Text Summarization Using KMeans and TF-IDF. *International Journal of Information Engineering and Electronic Business*, *11*(3), 33–44. https://doi.org/10.5815/ijieeb.2019.03.05

Haider, M. M., Hossin, M. A., Mahi, H. R., & Arif, H. (2020). Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm. *2020 IEEE Region 10 Symposium, TENSYMP 2020*, *June*, 283–286. https://doi.org/10.1109/TENSYMP50017.2020.9230670

Kågebäck, M., Mogren, O., Tahmasebi, N., & Dubhashi, D. (2015). *Extractive Summarization using Continuous Vector Space Models*. 31–39. https://doi.org/10.3115/v1/w14-1504

https://quillbot.com/summarize

https://www.malaymail.com/news/malaysia/2021/01/20/covid-19-vaccine-panel-chief-urges-malaysians-not-to-panic-over-norwegian-d/1942343

https://github.com/xiaoxu193/PyTeaser

https://www.malaymail.com/news/malaysia/2021/01/20/covid-19-vaccine-panel-chief-urges-malaysians-not-to-panic-over-norwegian-d/1942343https://www.nst.com.my/news/nation/2021/01/657965/covid-19-vaccine-timeline-allows-malaysia-study-efficacy-and-safety

https://machinelearningmastery.com/calculate-bleu-score-for-text-python/

https://trec.nist.gov/data/reuters/reuters.html