CAN THO UNIVERSITY
COLLEGE OF INFORMATION AND COMMUNICATION TECHNOLOGY

THESIS IN
INFORMATION TECHNOLOGY
(HIGH-QUALITY PROGRAM)

Topic

TEXT SUMMARIZATION WITH T5 MODEL

Student: Hồ Minh Nhựt
ID: B2005889
Course: K46

Can Tho, 02/2023

CAN THO UNIVERSITY
COLLEGE OF INFORMATION AND COMMUNICATION TECHNOLOGY

THESIS IN
INFORMATION TECHNOLOGY
(HIGH-QUALITY PROGRAM)

Topic

TEXT SUMMARIZATION WITH T5 MODEL

Advisor: Dr. Lam Nhut Khang

Student: Ho Minh Nhut
ID: B2005889
Course: K46

Can Tho, 02/2023

ACKNOWLEDGEMENTS

First and foremost I am extremely grateful to my advisor Dr. Lam Nhut Khang for her assistance at every stage of this thesis. Her immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life.

I also appreciate all the support I received from the rest of my family. Thank their financial support and encouragement, so I can complete this research.

Lastly, I would like to extend my sincere thanks to friends and classmates for their insightful comments and suggestions.

Table of Contents

CHAPTER 1: INTRODUCTION

1. Problems

In this new era, where tremendous information is available on the internet, it is most important to provide the improved mechanism to extract the information quickly and most efficiently. It is very difficult for human beings to manually extract the summary of a large document of text. It is also a relevant application in today’s information society given the exponential growth of textual information online and the need to promptly assess the contents of text collections. In other words, humans prefer to read the summary of news or articles more than reading the entire it in their busy life. Instead of reading all the articles, reading a summary will give us a brief about the story and save our time to select suitable articles.

Text Summarization topic has been researched to offer an efficient solution for these problems. Text Summarization is a subtask of Natural Language Processing (NLP) through advanced algorithms and techniques that analyze and extract the most essential information from a given text. The benefits of text summarization extend to a wide range of fields and professions, making it a highly useful tool for anyone looking to stay informed and stay ahead in today's information-saturated world.

2. Object

Text-To-Text Transfer Transformer (T5), a new pre-trained model on a data-rich task before being fine-tuned on a downstream task, has emerged as a powerful technique in natural language processing (NLP). By combining the insights from Google’s exploration with scale and our new “Colossal Clean Crawled Corpus”, they achieve state-of-the-art results on many benchmarks covering summarization, question answering, text classification, and more. Thus, I chose this model for my project topic.

3. Research scope

CNN/DailyMail Dataset

CHAPTER 2: LITERATURE REVIEW / THEORY

1. Python

Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed. Python is a popular language for machine learning and AI-based projects due to its simplicity, consistency, flexibility, platform independence, and access to great libraries and frameworks for AI and machine learning (ML)

https://www.python.org/doc/essays/blurb

2. Transformer

Transformer is a Novel Neural Network Architecture for Language Understanding based on a self-attention mechanism. This is The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best-performing models also connect the encoder and decoder through an attention mechanism

NLP’s Transformer is a new architecture that aims to solve tasks sequence-to-sequence while easily handling long-distance dependencies. Computing the input and output representations without using sequence-aligned RNNs or convolutions and relies entirely on self-attention.

In general, the Transformer model is based on the encoder-decoder architecture. The encoder and decoder consist of two and three sublayers, respectively. Multi-head self-awareness, fully connected feedforward network, and encoder-decoder self-awareness in the case of decoders (called multi-head attention) with the following visualizations. The below image is the basic architecture of Transformer

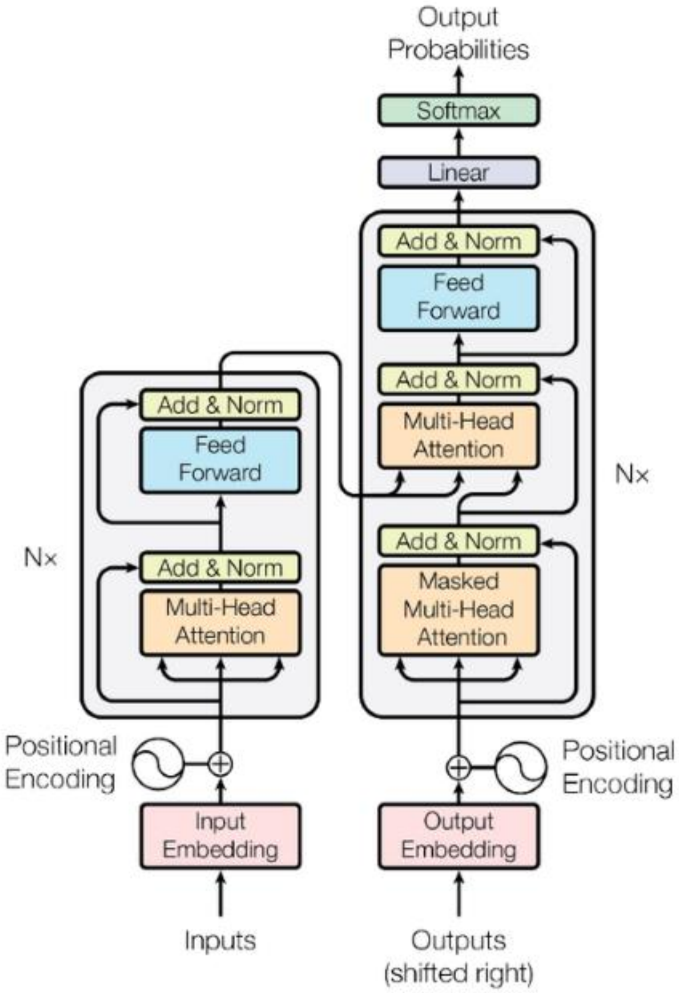


Figure 1. Basic architecture of Transformer

Inputs/outputs embedding is the step that text is parsed into tokens by a pair of encoding tokenizer. Each token is converted into a vector (array of numbers) via word embedding. After that, the positional information is added to the word embedding.

Encoder: each encoder consists of two major components: a self-attention mechanism and a feed-forward neural network. .The encoder is responsible for stepping through the input time steps and encoding the entire sequence into a fixed-length vector called a context vector that contains the information which parts of the inputs are relevant to each other. These output encodings are then passed to the next encoder as its input, as well as to the decoders.

Decoder: each decoder consists of three major components: a self-attention mechanism, an attention mechanism over the encodings, and a feed-forward neural network. An additional attention mechanism (encoder-decoder attention) is inserted which instead draws relevant information from the encodings generated by the encoders.

The word embeddings of the input sequence after being added positional information are passed to the first encoder. Next, these are then transformed and propagated to the next encoder. And then, the output from the last encoder in the encoder-stack is passed to all the decoders in the decoder-stack as shown in the figure above

https://blog.knoldus.com/what-are-transformers-in-nlp-and-its-advantages/

3. Transfer learning with T5 model

T5, or *Text-to-Text Transfer Transformer*, is a Transformer based architecture that uses a text-to-text approach. Every task – including translation, question answering, and classification – is cast as feeding the model text as input and training it to generate some target text. This allows for the use of the same model, loss function, hyperparameters, etc. across our diverse set of tasks.

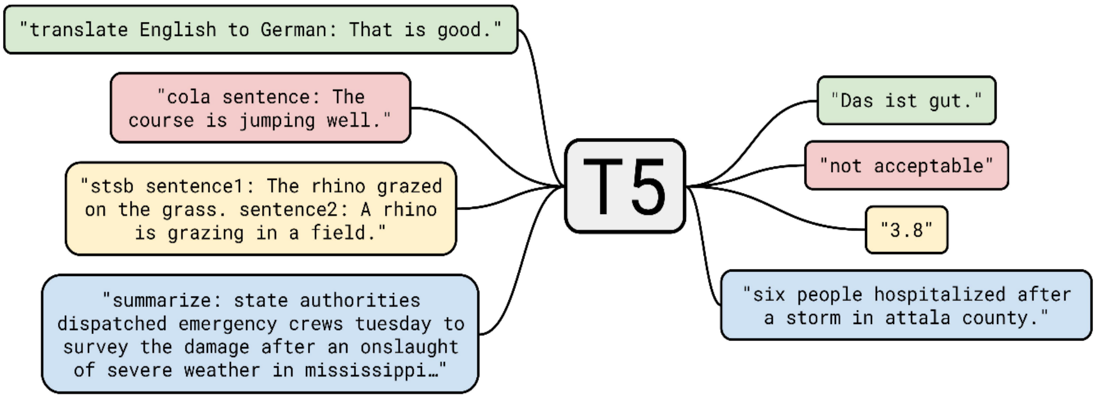

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Figure 2. Text-to-Text Tranfer Tranformer (T5)

The T5 model does not work with raw text. Instead, it requires the text to be transformed into numerical form in order to perform training and inference. The following transformations are required for the T5 model: tokenize text, convert tokens to IDs(integer), truncate the sequences to a specified maximum length, add end-of-sequence (EOS) and padding token IDs.

C4 is a new open-source pre-training dataset *Colossal Clean Crawled Corpus*. The T5 model, pre-trained on C4, achieves state-of-the-art results on many *Natural Language Processing* (NLP) benchmarks while being flexible enough to be fine-tuned to a variety of important downstream tasks.

With T5, Google proposes reframing all NLP tasks into a unified text-to-text format where the input and output are always text strings, in contrast to BERT-style models that can only output either a class label or a span of the input. The text-to-text framework allows us to use the same model, loss function, and hyperparameters on any NLP task, including machine translation, document summarization, question answering, and classification tasks (e.g., sentiment analysis). We can apply T5 to regression tasks by training it to predict the string representation of a number instead of the number itself.

https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.ht

4. Simple T5

SimpleT5 is built on top of PyTorch-lightning and Transformers which lets you quickly train your T5 models. This package supports processing data efficiently and trains the T5 model promptly with some simple lines of code.

5. Google Colaboratory

Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary Python code through the browser and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs. Colab notebooks execute code on Google's cloud servers, meaning you can leverage the power of Google hardware, including GPUs and TPUs, regardless of the power of your machine.

https://colab.research.google.com/#scrollTo=OwuxHmxllTwN

6. Datasets

Datasets is a library for easily accessing and sharing datasets for Audio, Computer Vision, and Natural Language Processing (NLP) tasks. Load a dataset in a single line of code, and use our powerful data processing methods to quickly get your dataset ready for training in a deep learning model. Backed by the Apache Arrow format, process large datasets with zero-copy reads without any memory constraints for optimal speed and efficiency.

6.1 CNN/DailyMail

The dataset of my project is the CNN/DailyMail Dataset which is an English-language dataset containing just over 300,000 unique news articles written by journalists at CNN and the Daily Mail. For each instance of this dataset, there is a string for the article, a string for the highlights, and a string for the id. The current version of CNN/DailyMail supports both extractive and abstractive summarization, so this dataset is very suitable for this project.

https://huggingface.co/datasets/cnn_dailymail

CHAPTER 3: TRAINNING MODEL

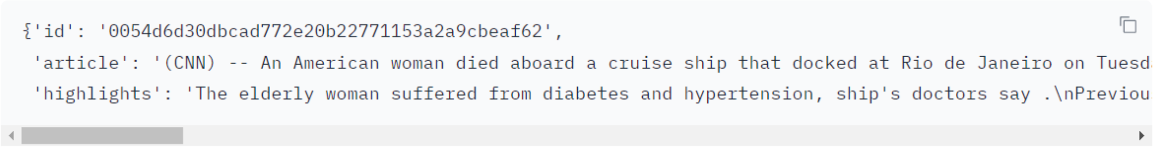
1. Hardware requirements

- Python 3.6 and above
- All stock of Python Machine Learning libraries
- Pytorch and Transformers

2. Dataset Structure

2.1 Dataset instance

In CNN_DailyMail dataset for each row, there is a string for the article, a string for the highlights, and a string for the id. However, in this project I just mainly focus on the article a and the highlights for training model.



```
{'id': '0054d6d30dbcad772e20b22771153a2a9cbeaf62',
 'article': '(CNN) -- An American woman died aboard a cruise ship that docked at Rio de Janeiro on Tuesd
 'highlights': 'The elderly woman suffered from diabetes and hypertension, ship's doctors say .\nPreviou
```

Figure 3. CNN_DailyMail dataset instance

2.2 Data fields

- Id: a string containing the heximal formated SHA1 hash of the url where the story was retrieved from
- Article: a string containing the body of the news article
- Highlights: a string containing the highlight of the article as written by the article author

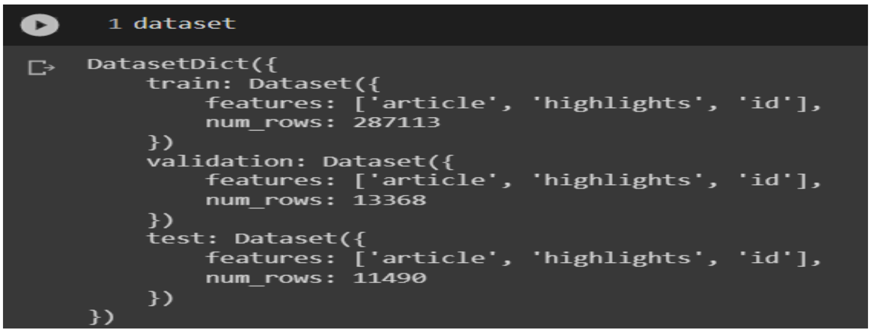
2.3 Data splits

The CNN/DailyMail dataset has three splits: train, validation, and test. Below are the statistics for Version 3.0.0 that I use for my for my project.

3. Prepare data for training

3.1 Load dataset

I use function *load_datase*t function of package *datasets* to load CNN_DailyMail from Hugging Face. After downloading, the dataset has the structure like below figure



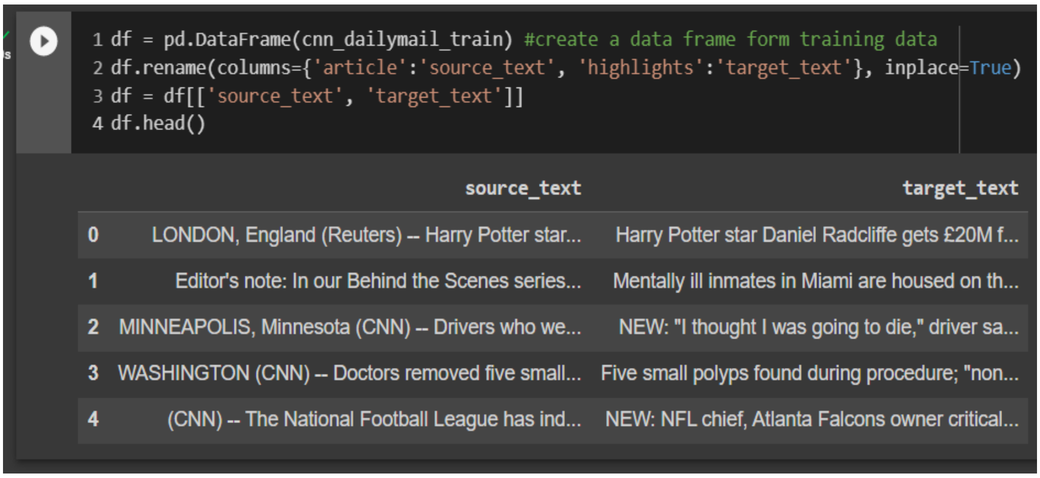
```
dataset
```

```
DatasetDict({
    train: Dataset({
        features: ['article', 'highlights', 'id'],
        num_rows: 287113
    })
    validation: Dataset({
        features: ['article', 'highlights', 'id'],
        num_rows: 13368
    })
    test: Dataset({
        features: ['article', 'highlights', 'id'],
        num_rows: 11490
    })
})
```

Figure 4. The dataset structure

3.2 Pre-process data

Converting the article and highlights because SimpleT5 expects input data frame have two columns "source_text" and "target_text".



```
df = pd.DataFrame(cnn_dailymail_train) #create a data frame form training data
df.rename(columns={'article':'source_text', 'highlights':'target_text'}, inplace=True)
df = df[['source_text', 'target_text']]
df.head()
```

	source_text	target_text
0	LONDON, England (Reuters) -- Harry Potter star...	Harry Potter star Daniel Radcliffe gets £20M f...
1	Editor's note: In our Behind the Scenes series...	Mentally ill inmates in Miami are housed on th...
2	MINNEAPOLIS, Minnesota (CNN) -- Drivers who we...	NEW: "I thought I was going to die," driver sa...
3	WASHINGTON (CNN) -- Doctors removed five small...	Five small polyps found during procedure; "non...
4	(CNN) -- The National Football League has ind...	NEW: NFL chief, Atlanta Falcons owner critical...

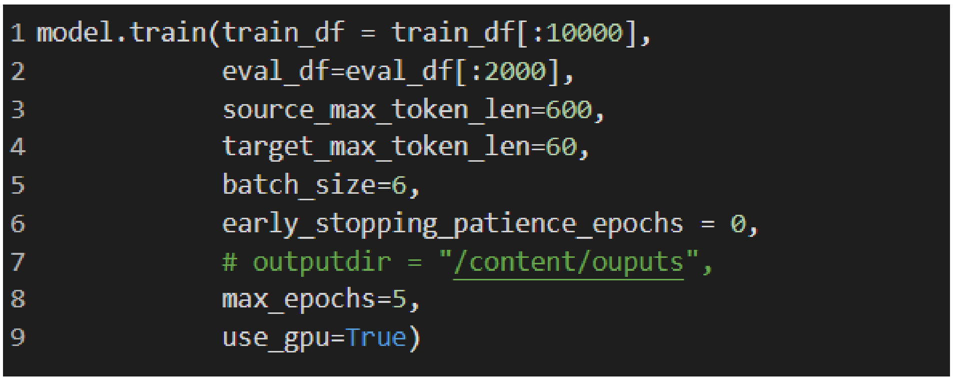
Figure 5. The prepared data for training

To prepare for train step, I split the training data into two parts: train(80%) and eval(20%).

4. Train model

Training T5 model by importing SimpleT5 class, download the pre-trained T5 model and then train it on our dataset in train_df and test_df. I also specify some optional arguments to make it appropriate with my input data.

Because of Colab's free GPU compute session, I must reduce the size of entire dataset when training model. I taking 10000 rows from *train_df* for training input and 2000 rows for evaluating from the *evalt_df* , the max length of input is 512 and summary text length is 60.



```
model.train(train_df = train_df[:10000],
            eval_df=eval_df[:2000],
            source_max_token_len=600,
            target_max_token_len=60,
            batch_size=6,
            early_stopping_patience_epochs = 0,
            # outputdir = "/content/ouputs",
            max_epochs=5,
            use_gpu=True)
```

Figure 6. Training model

5. Test and evaluate model

In this testing step, I divide into thee stage with the length of input: 500-1500 words, 1500-2500 words and above 2500.

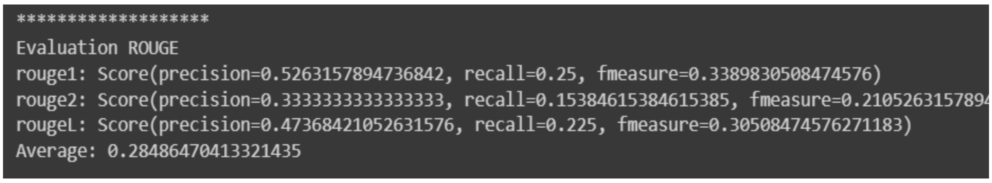
To evaluate model, ROUGE value is a set of metrics used to evaluate the quality of text summaries by comparing them to reference summaries created by humans. The ROUGE score is a scalar value in the range [0,1]. A ROUGE score close to zero indicates poor similarity between the candidate summary and the reference summaries, while a ROUGE score close to one indicates strong similarity.

5.1 Test with the article from 500-1500 words

- In this test the input article has length 624

+ Highlights: Bob Barker returned to host "The Price Is Right" on Wednesday . Barker, 91, had retired as host in 2007

+ Generated summary: Bob Barker hosted "The Price Is Right" for 35 years before stepping down in 2007. Barker handled the first price-guessing game of the TV show. He's been away from the show for most of the past eight years



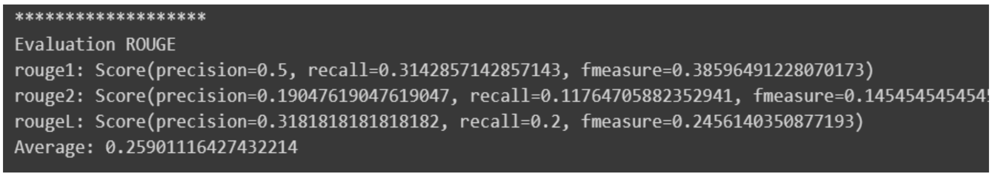
```
*******************
Evaluation ROUGE
rouge1: Score(precision=0.5263157894736842, recall=0.25, fmeasure=0.3389830508474576)
rouge2: Score(precision=0.3333333333333333, recall=0.15384615384615385, fmeasure=0.210526315789
rougeL: Score(precision=0.47368421052631576, recall=0.225, fmeasure=0.3050847457627118З)
Average: 0.28486470413321435
```

Figure 7. Result of testing with input 624 length

- In this test the input article has a length of 1071

+ Highlights: Deion Sanders calls out son for "hood doughnuts" comments. "You're a Huxtable with a million $ trust fund. Stop the hood stuff!

+ Generated summary: Deion Sanders Jr. reminded his son he has a trust fund, condo and clothing line. Junior is a wide receiver at Southern Methodist University. He has gone on record with his love for "hood doughnuts



```
*******************
Evaluation ROUGE
rouge1: Score(precision=0.5, recall=0.3142857142857143, fmeasure=0.38596491228070173)
rouge2: Score(precision=0.19047619047619047, recall=0.11764705882352941, fmeasure=0.145454545454
rougeL: Score(precision=0.3181818181818182, recall=0.2, fmeasure=0.2456140350877193)
Average: 0.25901116427432214
```

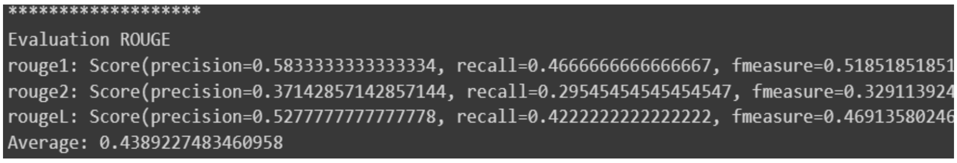
Figure 8. Result of testing with input 1071 length

5.2 Test with the article from 1500-2500 words

- In this test the input article has a length of 1913

+ Highlights: NASA chief scientist Ellen Stofan believes we're close to finding alien life . Indications within a decade; definitive evidence within "20 to 30 years," she said . Finding water on other celestial bodies is key to determination

+ Generated summary: NASA chief scientist Ellen Stofan: "We're going to have definitive evidence within 20 to 30 years" Scientists found evidence of water on a number of celestial bodies, including Jupiter's moon Europa. NASA isn't talking about intelligent alien civilizations from the Alpha Quadrant

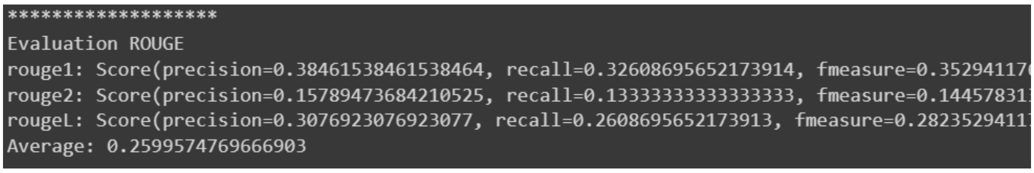


```
*******************
Evaluation ROUGE
rouge1: Score(precision=0.5833333333333334, recall=0.4666666666666667, fmeasure=0.51851851851
rouge2: Score(precision=0.37142857142857144, recall=0.29545454545454547, fmeasure=0.329113924
rougeL: Score(precision=0.5277777777777778, recall=0.4222222222222222, fmeasure=0.46913580246
Average: 0.4389227483460958
```

Figure 9. Result of testing with input 1913 length

- In this test the input article has a length of 2408

+ Highlights: Robert Bates said he meant to subdue a suspect with a Taser but accidentally shot him . The preliminary hearing is scheduled for July 2 . The judge said Bates was free to travel to the Bahamas for a family vacation . + Generated summary: Robert Bates said he meant to subdue a suspect with a Taser but accidentally shot him . The preliminary hearing is scheduled for July 2 . The judge said Bates was free to travel to the Bahamas for a family vacation .



```
*******************
Evaluation ROUGE
rouge1: Score(precision=0.38461538461538464, recall=0.32608695652173914, fmeasure=0.3529411
rouge2: Score(precision=0.15789473684210525, recall=0.13333333333333333, fmeasure=0.14457831
rougeL: Score(precision=0.3076923076923077, recall=0.2608695652173913, fmeasure=0.2823529411
Average: 0.2599574769666903
```

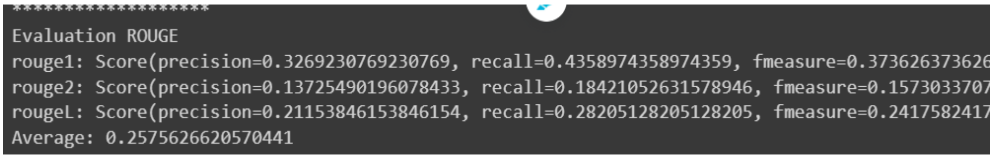
Figure 10. Result of testing with input 2408 length

5.3 Test with the article from 2500 words

- In this test the input article has a length of 4514

+ Highlights: Many girls in Nima,one of Accra's poorest slums, receive little or no education . Achievers Ghana is a school funded by the community to give the next generation a better chance of success . Girls are being taught to code by tech entrepreneur Regina Agyare, who believes her students will go far .

+ Generated summary: Tech entrepreneur Regina Agyare is helping 250 girls in Nima, Ghana, study computer science. Achievers Ghana provides school funding to help the girls shape their own future. Agyare: "I definitely feel [technology] has given them more of a voice"



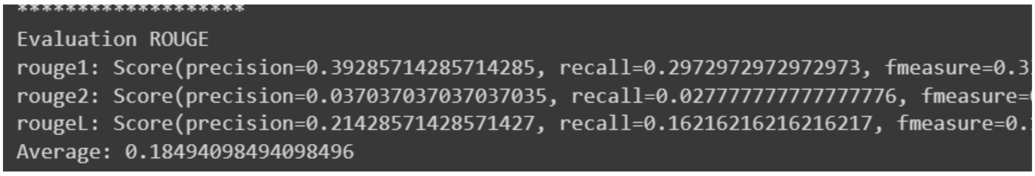
```
********************
Evaluation ROUGE
rouge1: Score(precision=0.3269230769230769, recall=0.4358974358974359, fmeasure=0.373626373626
rouge2: Score(precision=0.13725490196078433, recall=0.18421052631578946, fmeasure=0.1573033707
rougeL: Score(precision=0.21153846153846154, recall=0.28205128205128205, fmeasure=0.2417582417
Average: 0.2575626620570441
```

Figure 11. Result of testing with input 4514 length

- In this test the input article has a length of 6190

+ Highlights: Kenyans gather in Nairobi to remember victims of a terrorist attack that stunned a nation . The attack at a Garissa university last week killed 147 people, mostly students

+ Generated summary: Kenya launches airstrikes targeting Al-Shabaab training camps in Somalia, a military source says. Kenyan authorities have not released the names of the victims. A vigil was held at a Kenyan university in Garissa on Tuesday.



```
*********************
Evaluation ROUGE
rouge1: Score(precision=0.39285714285714285, recall=0.2972972972972973, fmeasure=0.3
rouge2: Score(precision=0.037037037037037035, recall=0.027777777777777776, fmeasure=
rougeL: Score(precision=0.21428571428571427, recall=0.16216216216216217, fmeasure=0.
Average: 0.18494098494098496
```

Figure 12. Result of testing with input 6190 length

5.4 Evaluate model

After some stages of testing my model. I recognized that the model has better performance with the input text has a length around 1500 – 3000 words. Below is the graph of rouge value after 1000 testing with random input articles from test split of dataset.

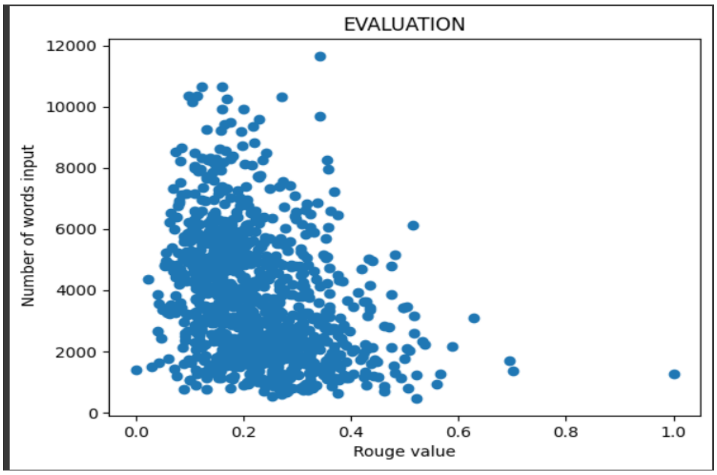


Figure 13 Graph of rouge value with different length of words input

CHAPTER 4: CONCLUSION

1 Result achieved

Find a dataset is suitable for my project that is available on the fly. Expand my horizon of knowledge about machine learning. Train and test successfully a model for task of text summarization.

2 Advantages

The best advantage of the success of my project is support of my advisor regarding sources of information, experiences, and knowledge. The free GPU of Google Colab also provides me with a free powerful environment to work with machine learning.

3 Disadvantages

This is the first time I approach machine learning, so my experience with this scope is insufficient. I need to get deep knowledge about training and evaluating a model of machine learning. The limitation of free GPU computing memory also hinders my training step.

4 Development direction

Training model with more appropriate arguments to fit the model with requirements and enhancing the rouge value of this model. Improve and apply the model to a graphical application or web extension for users easier to use.

REFERENCES

"*Attention is all you need*" *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin., 2017.*

"Exploring the Limits of Transfer Learning with a Unified Text-to-*Text Transformer" by Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu*

"Studying the Usage of Text-To-Text Transfer Transformer to Support Code-Related Tasks" Antonio Mastropaolo, Simone Scalabrino, Nathan Cooper, David Nader Palacio, Denys Poshyvanyk, Rocco Oliveto, Gabriele Bavota

"Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer" Posted by Adam Roberts, Staff Software Engineer and Colin Raffel, Senior Research Scientist, Google Research on Monday, February 24, 2020