**ABSTRACT**

**TEXT SUMMARIZATION**

Summarization is of two types:

# **EXTRACTIVE SUMMARIZATION:**

As the word suggests, we find essential sentences from a text and put them together to give us the summary.

# 

**ALGORITHM**

1. Import all the required libraries(nltk, nltk.corpus, nltk.tokenize, pandas, matplotlib and markovify).
2. Provide the Input (any text).
3. Store Stop\_Words in an array (Stop Words are is, in, off, from, we, you,etc.).
4. Tokenize the sentences.
5. Create a table to keep track of the words and how many times a word is used (use a Dictionary).
6. We use another dictionary to store all sentences, and it's token.
7. We find the average by the (Total value of the sentences / Total number of sentences).
8. We then print the summary of the given text.

**APPLICATIONS**

This program can be used to generate the summary of books, reports, documents etc. It can be used in the field of literature where you want to understand a large text, in the field of law for a summary, in the field of medical sciences to get the summary of a report.

**LIMITATIONS**

It can not generate a summary for huge paragraphs (Count < 2000). A standard threshold value does not hold good for all cases.

**HOW CAN IT BE IMPROVED?**

The program can be improved by making it capable of handling text of any size, By adding some minor tweaks like providing the meaning of a complex word.

**HOW TO MONETIZE?**

This method of summarization has a small market. Most summarizers out there depend mostly on ads for revenue.

**THINGS THAT I WASN’T ABLE TO DO?**

There aren’t many things to do in this program. The Extractive side has limited functionality. I wasn’t able to control the threshold value(If I change it manually, I get bits of a summary).

**WHY CREATING THIS?**

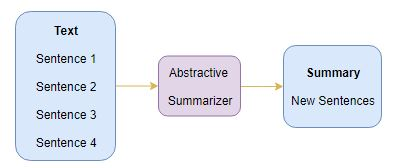
To solve the problem of reading and understanding huge amounts of text in a short period of time.

**WHO ARE THE TARGET AUDIENCE?**

Doctors, Writers, Readers, Students, Researchers, Lawyers, Reporters. We need to manage time these days, and this is a good way to manage some time.

# **ABSTRACTIVE SUMMARIZATION:**

Here, we form new sentences for our summary.



**ALGORITHM**

1. Import Attention Layer and required libraries(Numpy, pandas, urllib, keras,nltk, TensorFlow, bs4.etc).
2. Provide input(Give a file that has text sorted).
3. Drop duplicates and null values from the dataset.
4. Processing the data.
5. Convert text into lowercase.
6. Remove (‘s).
7. Remove stopwords.
8. Remove special characters.

5. Remove empty rows.

6. Understand sequence distribution.

7. Text and Summary tokenization.

8. Build a model.

9. Inference for Encoder and Decoder.

**APPLICATIONS**

1. This program can be used to generate summaries for customer reviews in a short amount of time(Depends on the power of your computer).
2. Can be used to get summaries for any text.
3. We can have a summary that we can understand easily for any text (like reports, books, documentation, articles etc.)

**WHY CREATING THIS?**

This summarizer understands the text and puts it in a way that we understand it more clearly. This model will save a lot of time and energy. It will give you exactly what you need from the text.

**LIMITATIONS**

1. Doesn’t have the compatibility with file extensions other than .csv file.
2. Need to train the model for more cases if not done correctly, we may not get the desired results.

**HOW CAN IT BE IMPROVED?**

* By implementing bi-directional LSTM (Long Short Term Memory Model).
* By increasing the training dataset size. (I have used only 15000 data samples).
* Evaluate the performance of the model using the BLEU (Bilingual Evaluation Understudy Score). If the result is perfect, it gives a score of 1.0 else 0.

**HOW TO MONETIZE?**

* We can host a website with the summarizer is a free tool and get revenue from ads.
* We can make this as an application on various platforms and put a paywall. Abstractive summarization has potential in the market.

**THINGS THAT I WASN’T ABLE TO DO?**

1. I wasn’t able to give more training data to the model due to the lack of computational power.
2. I didn’t have enough resources and knowledge to take the program further.

**WHO ARE THE TARGET AUDIENCE?**

Doctors, Writers, Readers, Students, Researchers, Lawyers, Reporters. We need to manage time these days, and this is a good way to manage some time.