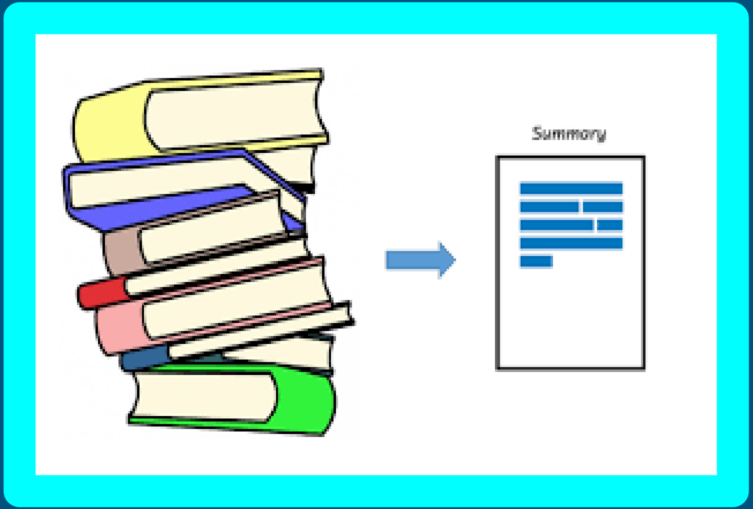
Text Summarization

- Document/ Text reading
    - Short summary
    - Long Summary



Created by:
Rasika Gulhane
Mahesh Jadhav
Chaitanya Dave

Contents:

- Introduction
- Features
- User Interface
- Usage
- Analysis
- Model Implementations
    - Pegasus
    - Hugging Face BERT
    - Hugging face BART
    - Hugging face T5_base Model
- Fine Tuning Hugging Face BART Model
- REST API work
- Findings
- Docker & AWS EC2
- Github
- Conclusion

Introduction:

Text Summarization is Generative AI for research and document summarization is an advanced solution that utilizes natural language processing (NLP) and deep learning techniques to automatically generate concise and coherent summaries of lengthy documents quickly and efficiently. The application provides a user-friendly interface where users can input their text or upload a PDF document for summarization.

The summarization process can be customized based on the user's preference, with options for short or long summaries. It utilizes the Hugging Face library and the framework to provide a user-friendly interface for text summarization.

Features:

1. Text Placement:

- User can either paste or type the article text directly into the input textarea provided.
- The word count of the input text is dynamically displayed, giving users an idea of the text length.

2. Upload Document (PDF):

- User have the option to upload a PDF document for summarization.
- The uploaded PDF is processed, and the text content is extracted for summarization.
- The extracted text is displayed in the input textarea, allowing users to make modifications if needed.

Features:

3. Summary Type:

- User can choose between a short or long summary type based on their requirements.
- The selected summary type determines the level of detail in the summarized output.

4. Summarized Text:

- The application generates a summarized version of the input text using advanced natural language processing techniques.
- The summarized text is displayed in a readonly textarea, allowing users to easily view and copy the summary.
- The word count of the summarized text is dynamically displayed, providing users with the length of the summary.

Features:

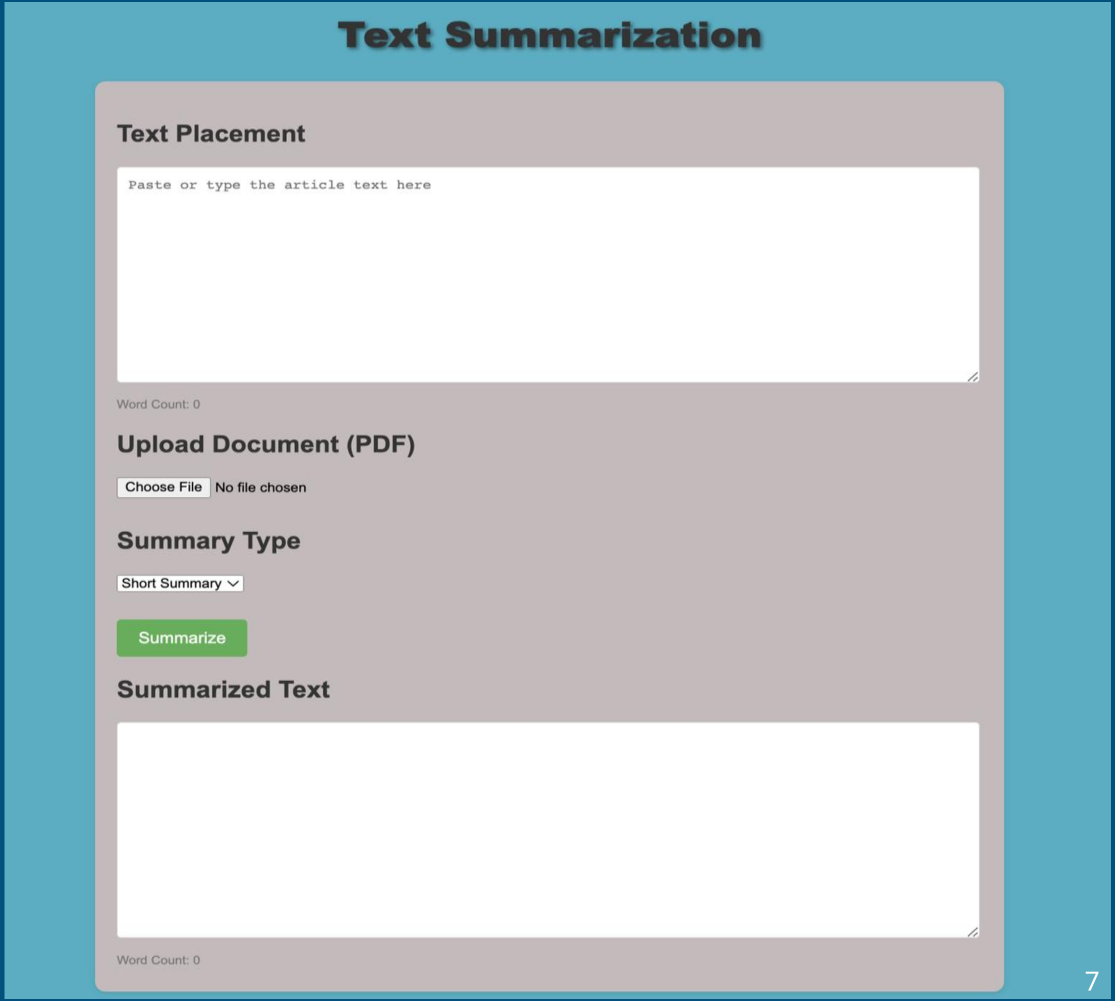
5. Summarize Button:

- Clicking the "Summarize" button initiates the summarization process.
- A loading message is displayed during the summarization process to indicate progress.
- Upon completion, the summarized text is updated, and the word count is recalculated.

6. Technical Details:

- The web application is built using HTML, CSS, and JavaScript.
- It leverages the PDF.js library to extract text content from uploaded PDF documents.
- The text summarization process is handled through a server-side component that utilizes natural language processing algorithms.
- AJAX is used to send asynchronous requests to the server for summarization and receive the summarized text as a response.
- The application dynamically updates the word count of the input text and summarized text using JavaScript functions.

User Interface:



Text Summarization

Text Placement

Paste or type the article text here

Word Count: 0

Upload Document (PDF)

No file chosen

Summary Type

Short Summary

Summarized Text

Word Count: 0

Usage:

1. Python 3.11.1
2. Anaconda 4.12
- VScode IDE

1. Required Libraries
- numpy
- pandas
- nltk
- tensorflow
- torch
- transformers
- pytorch-lightning
- rouge
- sentencepiece
- scikit-learn
- protobuf==3.20.0
- accelerate

4. Laptop Configuration:

- Windows 17 processor
- Nvidia
- GPU compatible

5. Github repository

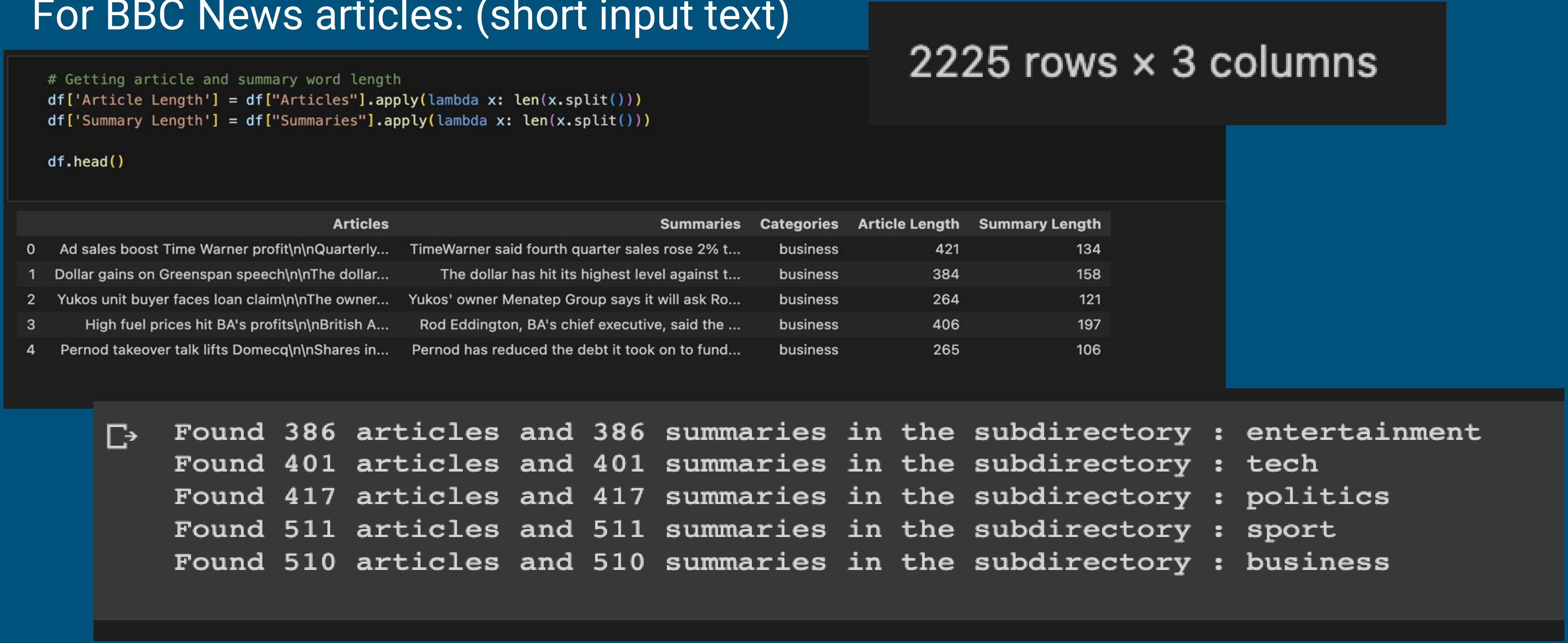
- https://github.com/Rasika-Gulhane/Document_Summarization

6. Docker Desktop

6. AWS EC2 cloud platform

Analysis:

For BBC News articles: (short input text)



2225 rows × 3 columns

```
# Getting article and summary word length
df['Article Length'] = df["Articles"].apply(lambda x: len(x.split()))
df['Summary Length'] = df["Summaries"].apply(lambda x: len(x.split()))

df.head()
```

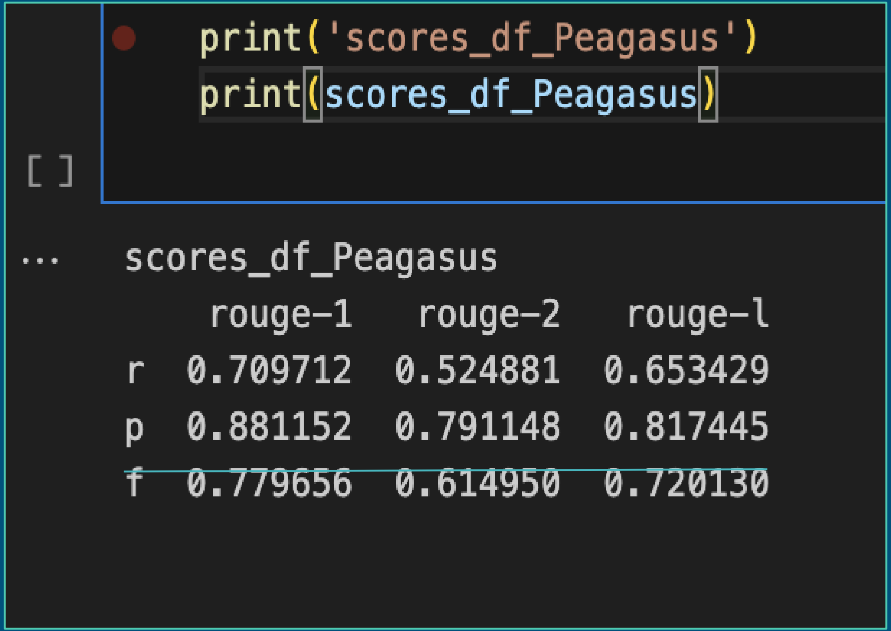
	Articles	Summaries	Categories	Article Length	Summary Length
0	Ad sales boost Time Warner profit\n\nQuarterly...	TimeWarner said fourth quarter sales rose 2% t...	business	421	134
1	Dollar gains on Greenspan speech\n\nThe dollar...	The dollar has hit its highest level against t...	business	384	158
2	Yukos unit buyer faces loan claim\n\nThe owner...	Yukos' owner Menatep Group says it will ask Ro...	business	264	121
3	High fuel prices hit BA's profits\n\nBritish A...	Rod Eddington, BA's chief executive, said the ...	business	406	197
4	Pernod takeover talk lifts Domecq\n\nShares in...	Pernod has reduced the debt it took on to fund...	business	265	106

```
Found 386 articles and 386 summaries in the subdirectory : entertainment
Found 401 articles and 401 summaries in the subdirectory : tech
Found 417 articles and 417 summaries in the subdirectory : politics
Found 511 articles and 511 summaries in the subdirectory : sport
Found 510 articles and 510 summaries in the subdirectory : business
```

Pegasus Summarization

Summarization on BBC article using Pegasus library with `model_name = 'google/pegasus-cnn_dailymail'`

Evaluation of actual summary vs generated summary using Rouge score.
(F1 score is considered)



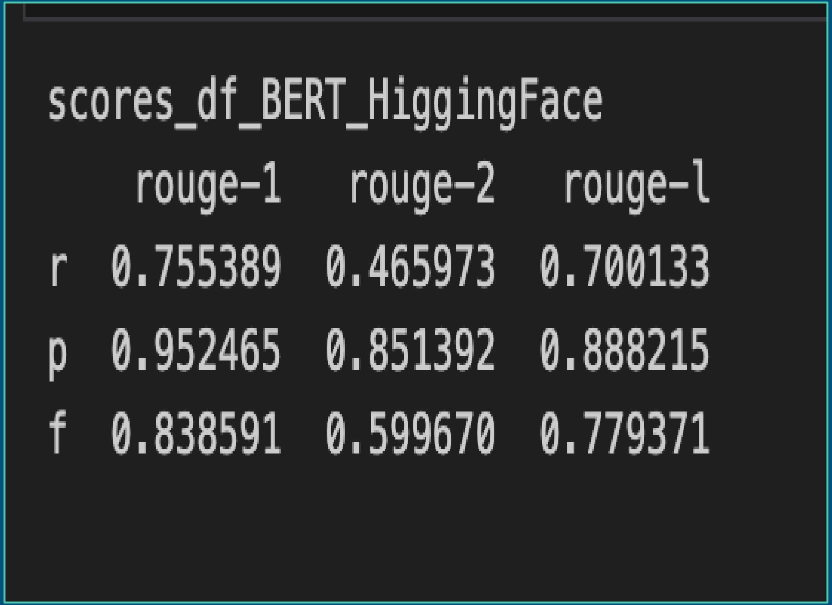
```
print('scores_df_Peagasus')
print(scores_df_Peagasus)
```

```
scores_df_Peagasus
    rouge-1   rouge-2   rouge-l
r  0.709712  0.524881  0.653429
p  0.881152  0.791148  0.817445
f  0.779656  0.614950  0.720130
```

Hugging Face BERT Summarization

Summarization on BBC article using pipeline summarization from transformers with Hugging Face

Evaluation of actual summary vs generated summary using Rouge score.
(F1 score is considered)



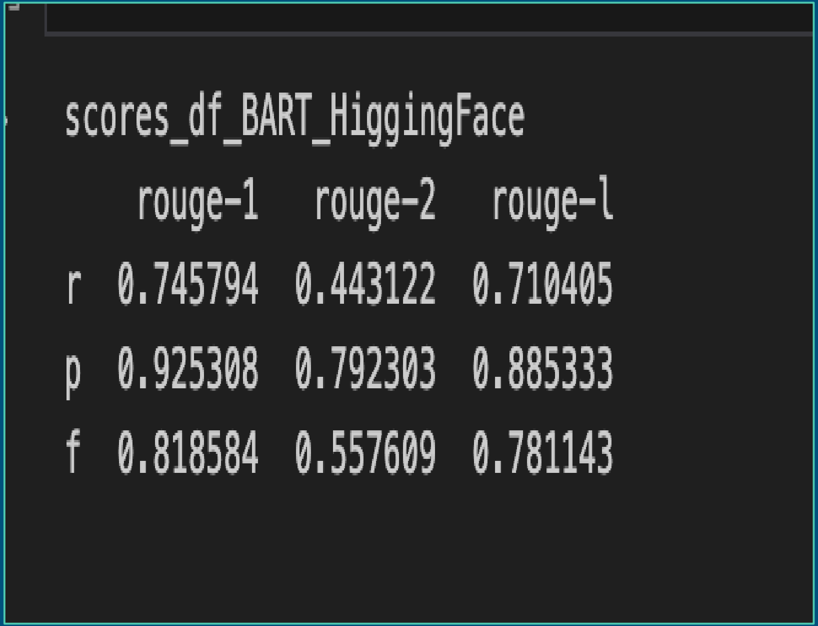
```
scores_df_BERT_HiggingFace
    rouge-1   rouge-2   rouge-l
r  0.755389  0.465973  0.700133
p  0.952465  0.851392  0.888215
f  0.838591  0.599670  0.779371
```

Hugging Face BART Summarization

Summarization on BBC article usind
model_name = "facebook/bart-large-cnn"

Evaluation of actual summary vs generated summary using Rouge score.
(F1 score is considered)

Since the average score is good for BERT and BART , Tried Fine tuning both with some parameters.



```
scores_df_BART_HiggingFace
    rouge-1   rouge-2   rouge-l
r  0.745794  0.443122  0.710405
p  0.925308  0.792303  0.885333
f  0.818584  0.557609  0.781143
```

Hugging Face T5_base Summarization

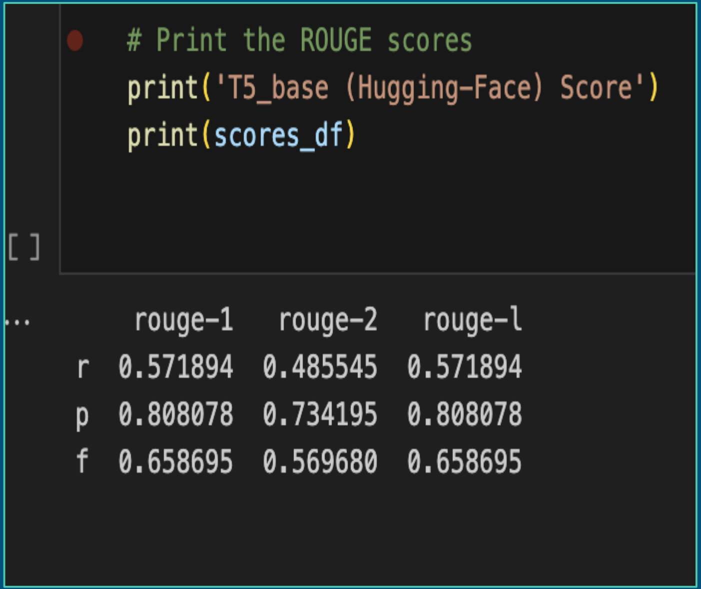
Summarization on BBC article using PyTorch lightning library

For tokenizer use of MODEL_NAME = 't5-base'

Evaluation of actual summary vs generated summary using Rouge score.
(F1 score is considered)

Tried manipulating epoches but taking longer time to execute.

Decreasing epoches affects score



```
# Print the ROUGE scores
print('T5_base (Hugging-Face) Score')
print(scores_df)
```

```
     rouge-1   rouge-2   rouge-l
r  0.571894  0.485545  0.571894
p  0.808078  0.734195  0.808078
f  0.658695  0.569680  0.658695
```

Fine Tune Hugging face BART model

For better Accuracy and fastest result using device function of torch library

```
device = torch.device("cuda" if torch.cuda.is_available()
else "cpu")
```

Model Used:

```
model_name = "sshleifer/distilbart-cnn-12-6"
```

```
summarizer = pipeline("summarization", model=model_name, revision="a4f8f3e")
```

Split the text into chunks of maximum 600 words

```
max_words = 600
```

```
chunks = [text[i:i+max_words] for i in range(0, len(text), max_words)]
```

Chunks and size varies depending on received input for Short and Long summary.

API process:

GPU Compatibility:

The application checks for the availability of a GPU for accelerated processing. If a GPU is available, it is utilized for enhanced performance.

Code Structure:

The application code is written in Python, utilizing Flask Restful API for the web framework.

Within the function, the Hugging Face summarization pipeline is loaded using the specified model name and revision.

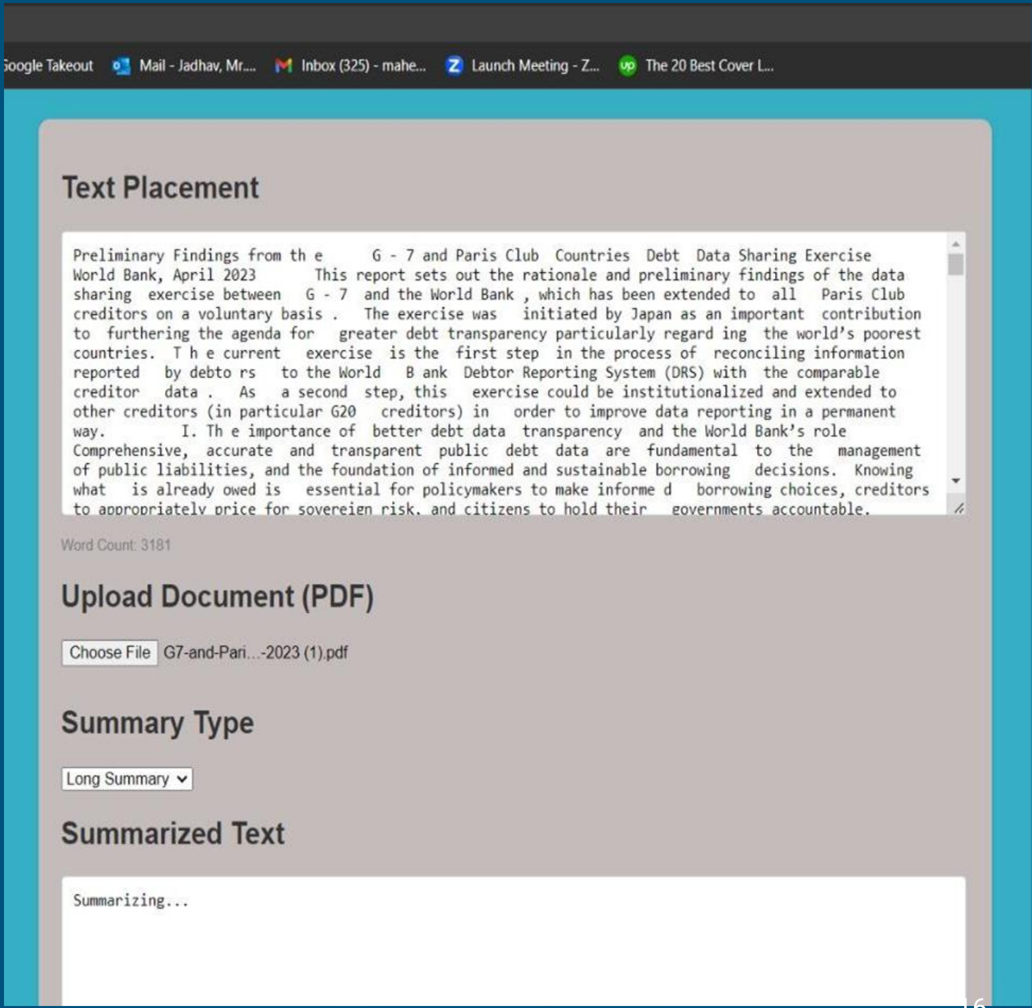
The transformers library from Hugging Face is used for text summarization.

The application code is structured into different routes, including the home route and the summary route.

To efficiently process large texts, the input text is divided into chunks of maximum 600 words. Each chunk is then passed through the summarization pipeline separately.

Document reading

Here, we can replace the text or we can upload any longer document. Input text has not limit of word count.



Text Placement

Preliminary Findings from th e G - 7 and Paris Club Countries Debt Data Sharing Exercise World Bank, April 2023 This report sets out the rationale and preliminary findings of the data sharing exercise between G - 7 and the World Bank , which has been extended to all Paris Club creditors on a voluntary basis . The exercise was initiated by Japan as an important contribution to furthering the agenda for greater debt transparency particularly regard ing the world's poorest countries. T h e current exercise is the first step in the process of reconciling information reported by debto rs to the World B ank Debtor Reporting System (DRS) with the comparable creditor data . As a second step, this exercise could be institutionalized and extended to other creditors (in particular G20 creditors) in order to improve data reporting in a permanent way. I. Th e importance of better debt data transparency and the World Bank's role Comprehensive, accurate and transparent public debt data are fundamental to the management of public liabilities, and the foundation of informed and sustainable borrowing decisions. Knowing what is already owed is essential for policymakers to make informe d borrowing choices, creditors to appropriately price for sovereign risk, and citizens to hold their governments accountable.

Word Count: 3181

Upload Document (PDF)

Choose File G7-and-Pari...-2023 (1).pdf

Summary Type

Long Summary

Summarized Text

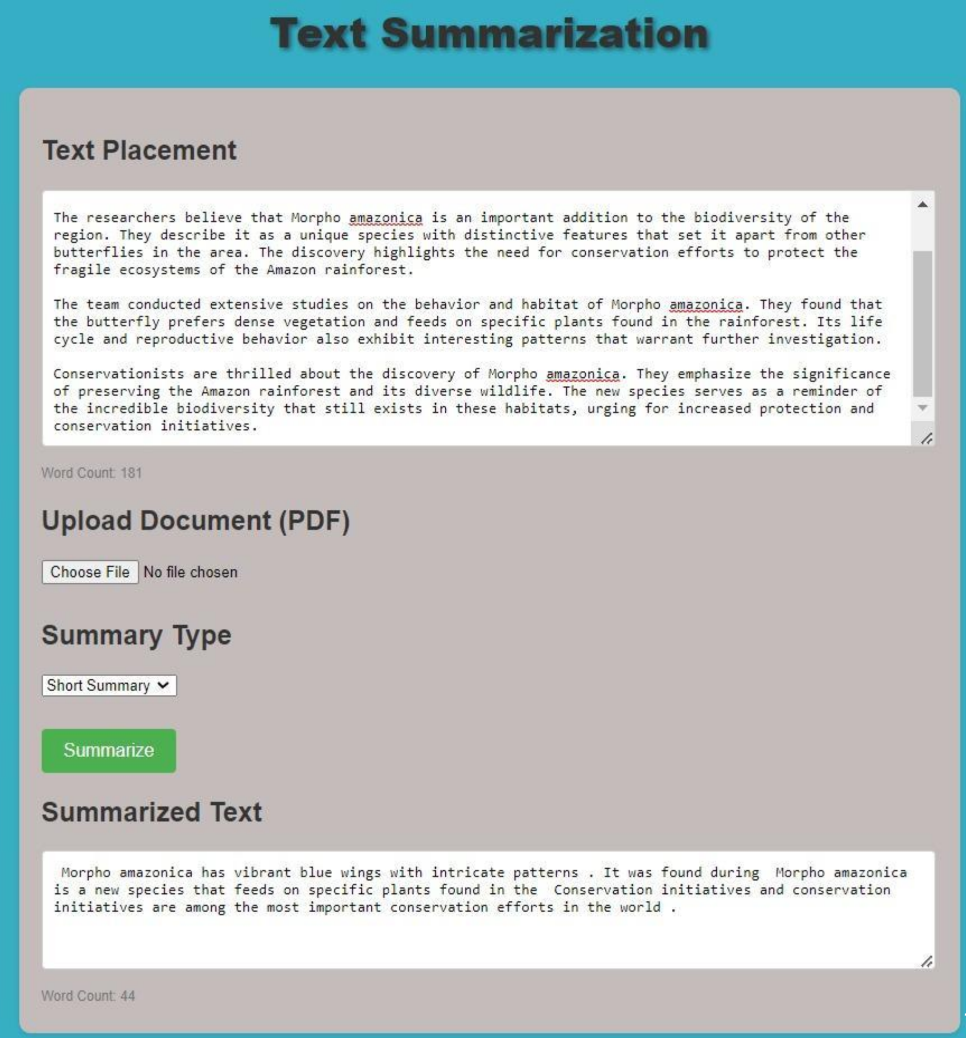
Summarizing...

Findings:

Summarization result depends on what type of summary user has selected short/long.

With shorter document of less than 1000 words :

- Short summary takes approx. 10 sec. on local system.
- Long summary takes approx. 15 sec.



Text Summarization

Text Placement

The researchers believe that Morpho amazonica is an important addition to the biodiversity of the region. They describe it as a unique species with distinctive features that set it apart from other butterflies in the area. The discovery highlights the need for conservation efforts to protect the fragile ecosystems of the Amazon rainforest.

The team conducted extensive studies on the behavior and habitat of Morpho amazonica. They found that the butterfly prefers dense vegetation and feeds on specific plants found in the rainforest. Its life cycle and reproductive behavior also exhibit interesting patterns that warrant further investigation.

Conservationists are thrilled about the discovery of Morpho amazonica. They emphasize the significance of preserving the Amazon rainforest and its diverse wildlife. The new species serves as a reminder of the incredible biodiversity that still exists in these habitats, urging for increased protection and conservation initiatives.

Word Count: 181

Upload Document (PDF)

No file chosen

Summary Type

Short Summary

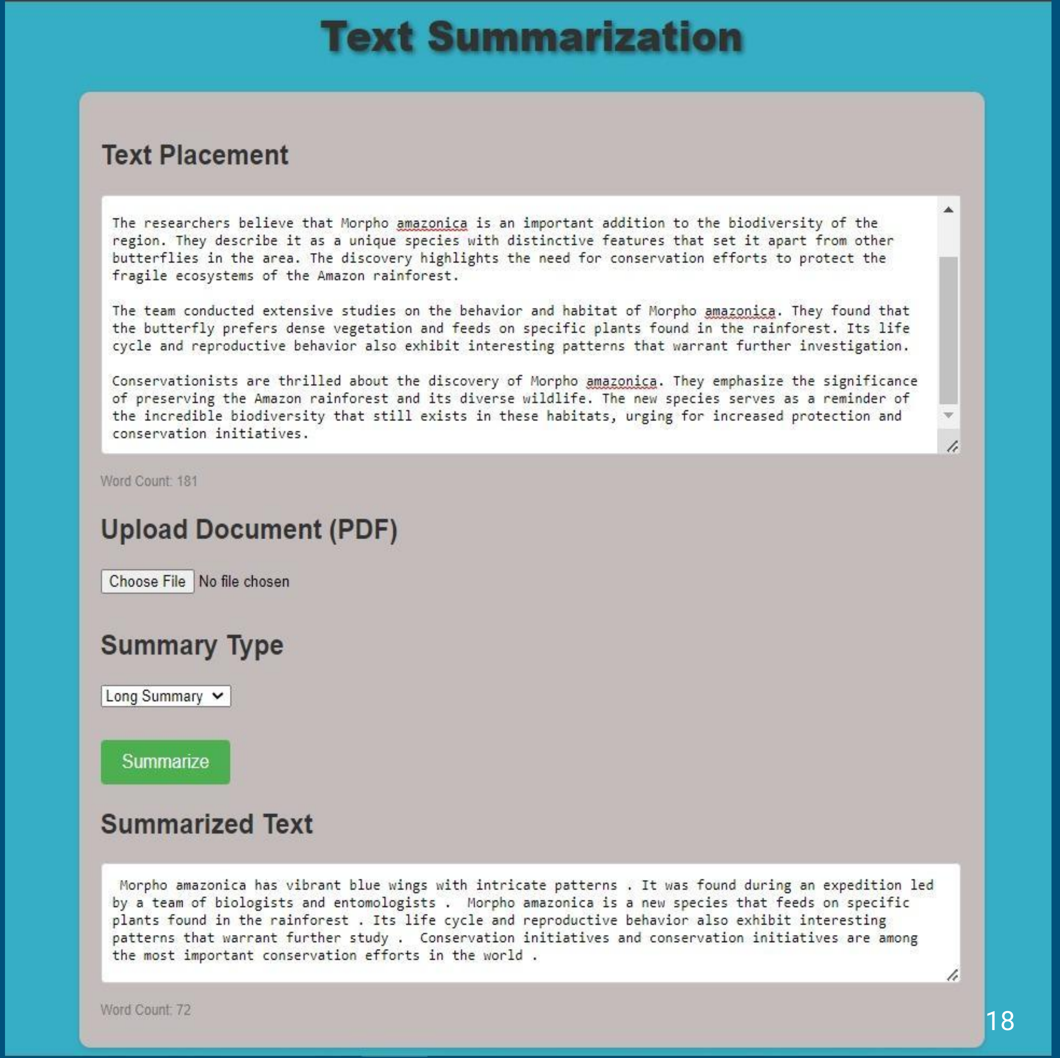
Summarized Text

Morpho amazonica has vibrant blue wings with intricate patterns . It was found during Morpho amazonica is a new species that feeds on specific plants found in the Conservation initiatives and conservation initiatives are among the most important conservation efforts in the world .

Word Count: 44

Findings:

- Greater the text greater is time taken for generating text summary.
- If text more than 600 words get process with chunking of each batches of 600 words.



Text Summarization

Text Placement

```
The researchers believe that Morpho amazonica is an important addition to the biodiversity of the
region. They describe it as a unique species with distinctive features that set it apart from other
butterflies in the area. The discovery highlights the need for conservation efforts to protect the
fragile ecosystems of the Amazon rainforest.

The team conducted extensive studies on the behavior and habitat of Morpho amazonica. They found that
the butterfly prefers dense vegetation and feeds on specific plants found in the rainforest. Its life
cycle and reproductive behavior also exhibit interesting patterns that warrant further investigation.

Conservationists are thrilled about the discovery of Morpho amazonica. They emphasize the significance
of preserving the Amazon rainforest and its diverse wildlife. The new species serves as a reminder of
the incredible biodiversity that still exists in these habitats, urging for increased protection and
conservation initiatives.
```

Word Count: 181

Upload Document (PDF)

Summary Type

Long Summary

Summarized Text

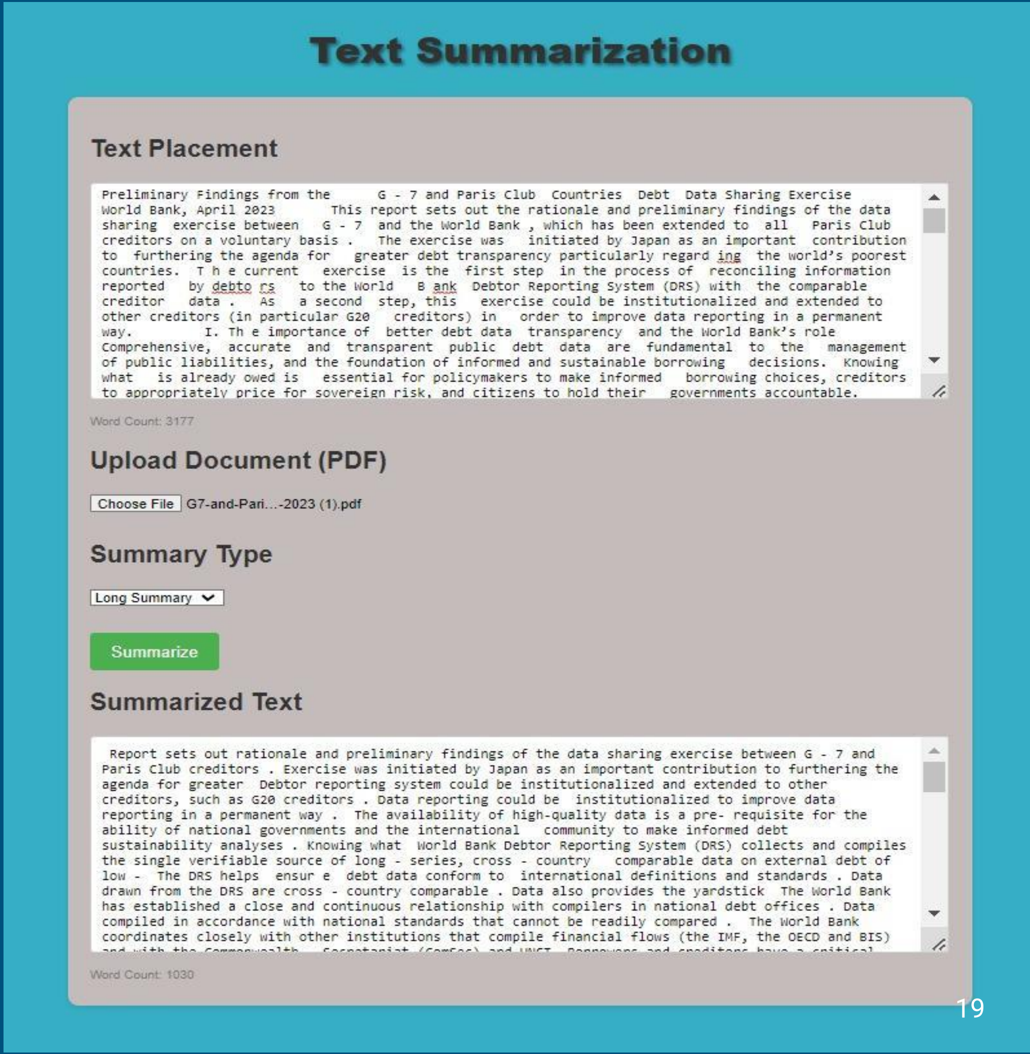
```
 Morpho amazonica has vibrant blue wings with intricate patterns . It was found during an expedition led
by a team of biologists and entomologists .  Morpho amazonica is a new species that feeds on specific
plants found in the rainforest . Its life cycle and reproductive behavior also exhibit interesting
patterns that warrant further study .  Conservation initiatives and conservation initiatives are among
the most important conservation efforts in the world .
```

Word Count: 72

Findings:

With longer document considering around 4000+ :

- Short summary takes approx. 2 min. on local system it can be reduced over cloud with fast processor
- Long summary takes approx. 3 min.



Text Summarization

Text Placement

Preliminary Findings from the   G - 7 and Paris Club Countries Debt Data Sharing Exercise
World Bank, April 2023   This report sets out the rationale and preliminary findings of the data
sharing exercise between   G - 7 and the World Bank , which has been extended to all   Paris Club
creditors on a voluntary basis .   The exercise was   initiated by Japan as an important contribution
to furthering the agenda for   greater debt transparency particularly regard ing the world's poorest
countries. T h e current   exercise is the first step in the process of reconciling information
reported   by debto rs   to the World   B ank Debtor Reporting System (DRS) with the comparable
creditor   data .   As   a second step, this   exercise could be institutionalized and extended to
other creditors (in particular G20   creditors) in   order to improve data reporting in a permanent
way.   I. Th e importance of better debt data transparency and the World Bank's role
Comprehensive, accurate and transparent public debt data are fundamental to the   management
of public liabilities, and the foundation of informed and sustainable borrowing   decisions. Knowing
what   is already owed is   essential for policymakers to make informed   borrowing choices, creditors
to appropriately price for sovereign risk, and citizens to hold their   governments accountable.

Word Count: 3177

Upload Document (PDF)

Choose File G7-and-Pari...-2023 (1).pdf

Summary Type

Long Summary

Summarize

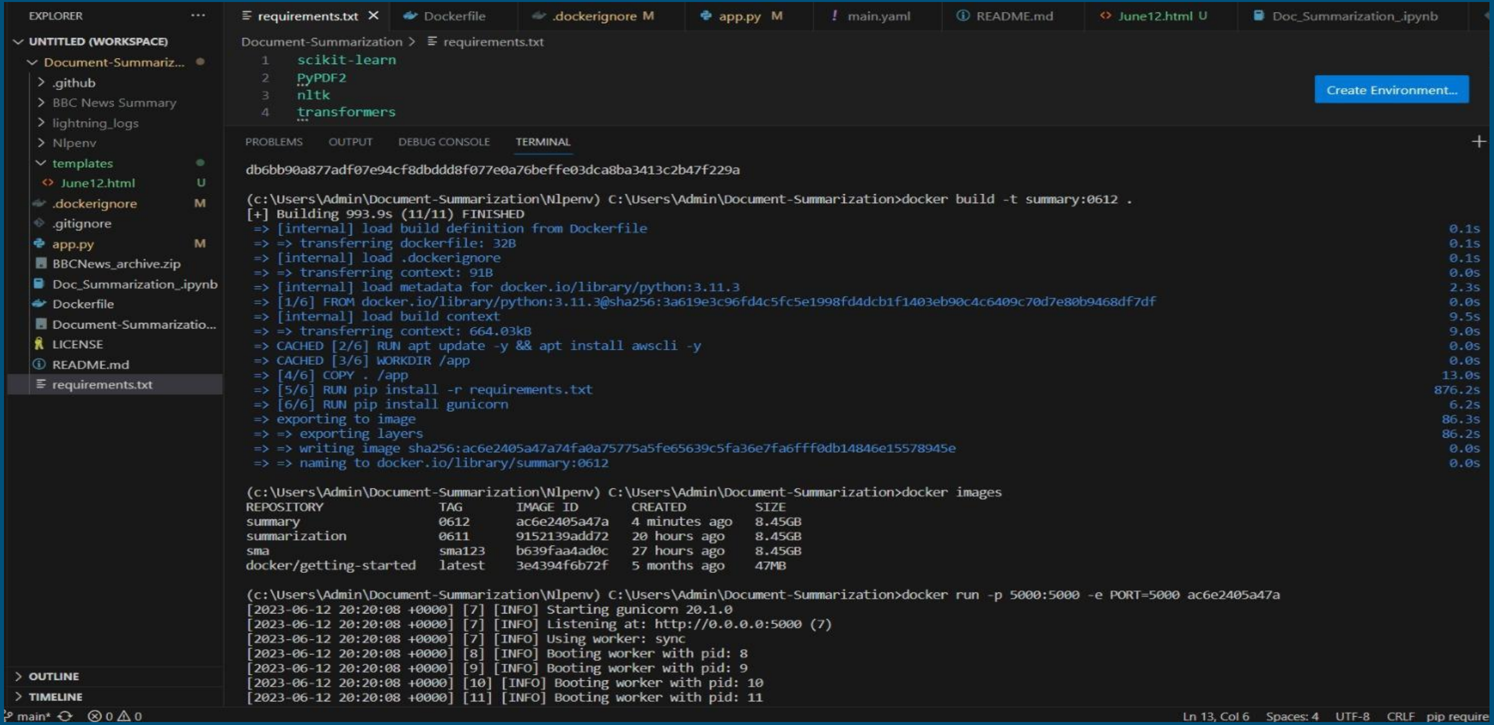
Summarized Text

Report sets out rationale and preliminary findings of the data sharing exercise between G - 7 and
Paris Club creditors . Exercise was initiated by Japan as an important contribution to furthering the
agenda for greater Debtor reporting system could be institutionalized and extended to other
creditors, such as G20 creditors . Data reporting could be institutionalized to improve data
reporting in a permanent way . The availability of high-quality data is a pre- requisite for the
ability of national governments and the international   community to make informed debt
sustainability analyses . Knowing what World Bank Debtor Reporting System (DRS) collects and compiles
the single verifiable source of long - series, cross - country   comparable data on external debt of
low - The DRS helps ensur e debt data conform to international definitions and standards . Data
drawn from the DRS are cross - country comparable . Data also provides the yardstick The World Bank
has established a close and continuous relationship with compilers in national debt offices . Data
compiled in accordance with national standards that cannot be readily compared . The World Bank
coordinates closely with other institutions that compile financial flows (the IMF, the OECD and BIS)
and with the Commonwealth Secretariat (ComSec) and UNCT. Borrowers and creditors have a critical

Word Count: 1030

Docker:

A docker image for future use and updation within project



```
scikit-learn
PyPDF2
nltk
transformers
```

```
db6bb90a877adf07e94cf8dbddd8f077e0a76beffe03dca8ba3413c2b47f229a

(c:\Users\Admin\Document-Summarization\Nlpenv) C:\Users\Admin\Document-Summarization>docker build -t summary:0612 .
[+] Building 993.9s (11/11) FINISHED
 => [internal] load build definition from Dockerfile                                                                          0.1s
 => => transferring dockerfile: 32B                                                                                           0.1s
 => [internal] load .dockerignore                                                                                             0.1s
 => => transferring context: 91B                                                                                              0.0s
 => [internal] load metadata for docker.io/library/python:3.11.3                                                              2.3s
 => [1/6] FROM docker.io/library/python:3.11.3@sha256:3a619e3c96fd4c5fc5e1998fd4dcb1f1403eb90c4c6409c70d7e80b9468df7df        0.0s
 => [internal] load build context                                                                                             9.5s
 => => transferring context: 664.03kB                                                                                         9.0s
 => CACHED [2/6] RUN apt update -y && apt install awscli -y                                                                   0.0s
 => CACHED [3/6] WORKDIR /app                                                                                                 0.0s
 => [4/6] COPY . /app                                                                                                        13.0s
 => [5/6] RUN pip install -r requirements.txt                                                                               876.2s
 => [6/6] RUN pip install gunicorn                                                                                            6.2s
 => exporting to image                                                                                                       86.3s
 => => exporting layers                                                                                                      86.2s
 => => writing image sha256:ac6e2405a47a74fa0a75775a5fe65639c5fa36e7fa6fff0db14846e15578945e                                  0.0s
 => => naming to docker.io/library/summary:0612                                                                               0.0s

(c:\Users\Admin\Document-Summarization\Nlpenv) C:\Users\Admin\Document-Summarization>docker images
REPOSITORY               TAG       IMAGE ID       CREATED          SIZE
summary                  0612      ac6e2405a47a   4 minutes ago    8.45GB
summarization            0611      9152139add72   20 hours ago     8.45GB
sma                      sma123    b639faa4ad0c   27 hours ago     8.45GB
docker/getting-started   latest    3e4394f6b72f   5 months ago     47MB

(c:\Users\Admin\Document-Summarization\Nlpenv) C:\Users\Admin\Document-Summarization>docker run -p 5000:5000 -e PORT=5000 ac6e2405a47a
[2023-06-12 20:20:08 +0000] [7] [INFO] Starting gunicorn 20.1.0
[2023-06-12 20:20:08 +0000] [7] [INFO] Listening at: http://0.0.0.0:5000 (7)
[2023-06-12 20:20:08 +0000] [7] [INFO] Using worker: sync
[2023-06-12 20:20:08 +0000] [8] [INFO] Booting worker with pid: 8
[2023-06-12 20:20:08 +0000] [9] [INFO] Booting worker with pid: 9
[2023-06-12 20:20:08 +0000] [10] [INFO] Booting worker with pid: 10
[2023-06-12 20:20:08 +0000] [11] [INFO] Booting worker with pid: 11
```

ReadMe Instruction to Run the Project:

Clone the project using link https://github.com/Rasika-Gulhane/Document_Summarization.git

- open in IDE (recommended VSCODE)
- open terminal of project

Create Python/Conda Environment

Command to run:

```
pip install -r requirements.txt
python app.py
```

click the link in terminal output

or

follow link http://127.0.0:5000

Conclusion:

The Text Summarization project provides a convenient and efficient solution for summarizing large amounts of text. Whether it's academic papers, news articles, legal documents or any other type of lengthy text, users can easily obtain concise summaries tailored to their needs. With its user-friendly interface and customizable summarization options, the Text Summarization application is a valuable tool for enhancing productivity and extracting key information from voluminous texts.

The text summarization uses ‘hugging face BART model’ which provides a convenient way to generate summaries and offers flexibility with short and long summary modification as per input received and it utilizes GPU acceleration for faster processing. The application can be further customized and enhanced to meet specific requirements.

Note: This report provides an overview of the application's features and functionality based on the provided code and HTML page. Further improvements and refinements can be made as per specific project requirements.