Summarizer using LED

TEAM NAME : AttentionIsAllWeNeed

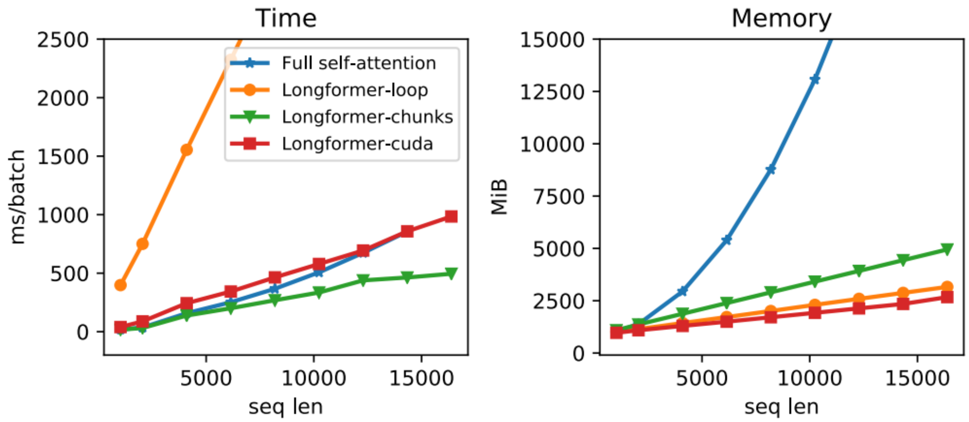
Team Members : Anwesh Saha (210178) , Arindom Bora (210183) , Ajay Sankar Makenna (210077), Khush Khandelwal (210511), Shreyash Nallawar (211010), Vineet Kumar (201121)

The heart of the model - Longformers!
1. The Summarizer Model : This app uses the ***Longformer Encoder-Decoder (LED)*** *for Narrative-Esque Long Text Summarization* architecture to summarize texts retrieved from PDFs, Documents, or web pages.
We had a couple of options in hand before we started our build:-

1. Google's T5
2. Facebook's BART
3. **Longformer Encoder-Decoder (LED) for Narrative-Esque Long Text Summarization**

We chose the third one because of the following reasons:

- **Performance and Quality:**
  Transformer-based models cannot process long sequences due to their self-attention operation, which scales quadratically with the sequence length. Longformers, on the other hand, can summarize text with a more significant number of words.
  The original Transformer consisted of an encoder-decoder architecture for sequence-to-sequence tasks, such as summarization and translation. While encoder-only Transformers are effective on various NLP tasks, pre-trained encoder-decoder Transformer models (e.g., BART and T5 have achieved strong results on tasks like summarization.) Yet, such models can't efficiently scale to seq2seq tasks with longer inputs.



. Longformer's attention mechanism is a drop-in replacement for the standard self-attention and combines local windowed attention with task-motivated global attention. It's performance and memory usage is more efficient compared to traditional self-attention.

- **Training**
  The pretraining dataset used for this model was one of the primary factors influencing our decision. LED-base-book-summary is trained on diverse sources, including books and research articles, providing a more extensive and comprehensive understanding of language. This diverse training data helps the model capture a broader range of contexts and improves its ability to generate meaningful summaries.

Architecture:- The model uses an encoder-decoder architecture for *seq2seq* modeling. What differentiates it from a traditional transformer is local and global attention.

- Self-Attention in Traditional Transformers: In traditional Transformers, self-attention simultaneously operates on the entire input sequence. It computes the attention scores between each token in the sequence and all other tokens, capturing their dependencies and relationships. This allows the model to establish both local and global connections. However, the quadratic complexity of self-attention concerning the sequence length could be more efficient for long sequences.
- Local Windowed Attention: Local windowed attention allows the model to attend to a subset of the input sequence, or a "local window," rather than the entire sequence simultaneously. This approach reduces the computational complexity and memory requirements compared to full self-attention. Limiting the attention scope allows the model to capture local dependencies within a fixed window size.
- Global Attention: As the name suggests, the model can attend to the entire input sequence. It is often used with local windowed attention to capturing local and global dependencies. While local attention focuses on nearby tokens, global attention provides a broader context by attending to all tokens in the sequence.

Training Data:- This model is pre-trained on the ***Booksum*** dataset by Salesforce. BookSum is a new collection of datasets specifically designed to address the limitations of existing text summarization datasets. Unlike other datasets mainly consisting of short-form source documents with limited dependencies, BookSum focuses on long-form narrative sources such as novels, plays, and stories. This dataset includes human-written summaries at three levels of granularity: paragraph, chapter, and book.

Number of parameters:- 162M params

Trained on	Booksum dataset (salesforce/booksum (github.com))
Model Size	~1.6G
Number of parameters	162M params

2. BERT Embeddings: BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art language representation model that has revolutionized various natural language processing (NLP) tasks. BERT embeddings capture contextual information by considering the entire sentence rather than individual words. This model can be fine-tuned on specific downstream tasks, such as text summarization, to generate meaningful representations of input text.

- BERT Pretraining: BERT is initially pre-trained on large corpora using masked language modeling and next sentence prediction tasks. This pretraining helps BERT to learn general language representations.
- Fine-tuning BERT: BERT can be fine-tuned on specific tasks, such as text summarization after pretraining. During fine-tuning, BERT is trained on labeled data specific to the task at hand, which helps it learn task-specific features and nuances.
- BERT Embeddings: The output of BERT's hidden layers can be used as embeddings for downstream tasks. These embeddings capture the contextual understanding of the input text, enabling better representation of words and sentences.

3. HNSWlib for KNN based on Cosine similarity: HNSWlib (Hierarchical Navigable Small World) is a library that provides an efficient data structure for approximate nearest neighbor search. In the context of text summarization, HNSWlib can perform a cosine similarity-based k-nearest neighbor (KNN) search on BERT embeddings.

- Building the HNSW Index: BERT embeddings of the documents in the corpus are indexed using the HNSWlib library. The HNSW index structure organizes the embeddings to allow for efficient nearest neighbor search.
- Cosine Similarity: Cosine similarity is a widely used metric to measure the similarity between two vectors. In the case of BERT embeddings, cosine similarity helps determine the semantic similarity between documents.
- K-Nearest Neighbor Search: Given a query document, the HNSW index efficiently finds the k nearest neighbors in the embedding space based on cosine similarity. These nearest neighbors can then generate a summary by extracting meaningful information from similar documents.

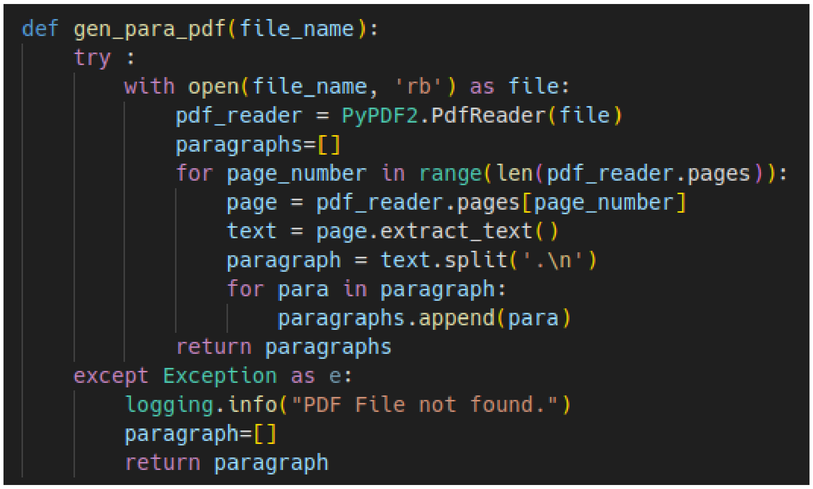
4. Generating Text Summaries: By combining BERT embeddings and HNSWlib for cosine KNN, a generative artificial intelligence system for text summarization can be developed. The following steps outline the process:

- Preprocessing: The input text, such as research papers, news articles, or documents, undergoes preprocessing to perform sentence segmentation, and tokenization.

- BERT Embeddings: The preprocessed text is passed through a fine-tuned BERT model to generate embeddings that capture the contextual understanding of the input.
- HNSW Indexing: The BERT embeddings of the documents are indexed using HNSWlib to create an efficient search index.
- Cosine KNN Search: Given a query document, the BERT embeddings are used to perform cosine similarity-based KNN search on the HNSW index, retrieving the most similar documents.
- Retrieval: The top 20 sentences similar to the query are retrieved from the HNSWlib index and post-processed to form a document.
- Summarization: Extractive or abstractive summarization techniques are applied to the retrieved documents using LED to generate concise summaries.

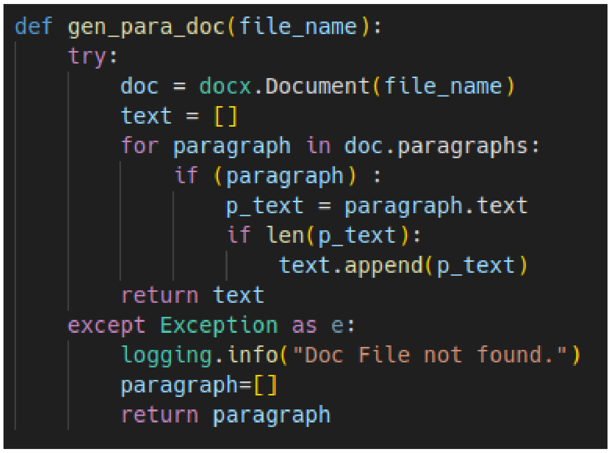
5. Text Extraction :

5.1 PDFs : PyPDF2 is a Python library for working with PDF files. We used it to extract text from PDF documents.



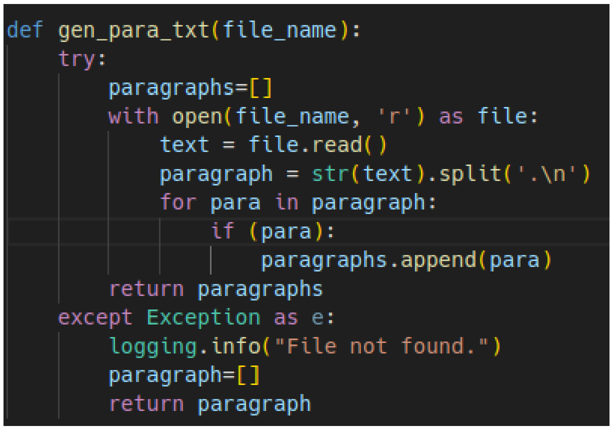
```
def gen_para_pdf(file_name):
    try :
        with open(file_name, 'rb') as file:
            pdf_reader = PyPDF2.PdfReader(file)
            paragraphs=[]
            for page_number in range(len(pdf_reader.pages)):
                page = pdf_reader.pages[page_number]
                text = page.extract_text()
                paragraph = text.split('.\n')
                for para in paragraph:
                    paragraphs.append(para)
            return paragraphs
    except Exception as e:
        logging.info("PDF File not found.")
        paragraph=[]
        return paragraph
```

5.2 Word Files : We used **python-docx,** another Python library for extracting text from Microsoft Word (.docx or .doc) files.

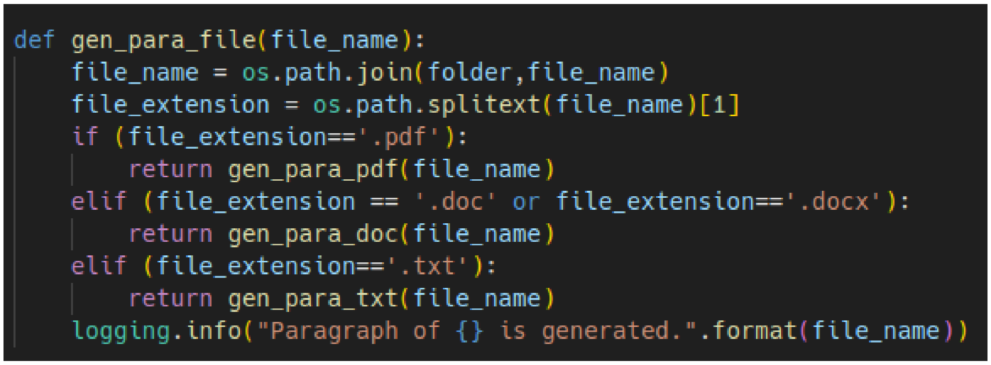


```
def gen_para_doc(file_name):
    try:
        doc = docx.Document(file_name)
        text = []
        for paragraph in doc.paragraphs:
            if (paragraph) :
                p_text = paragraph.text
                if len(p_text):
                    text.append(p_text)
        return text
    except Exception as e:
        logging.info("Doc File not found.")
        paragraph=[]
        return paragraph
```

5.3 Text Files : Python has inbuilt features to deal with text files using the **open** command.



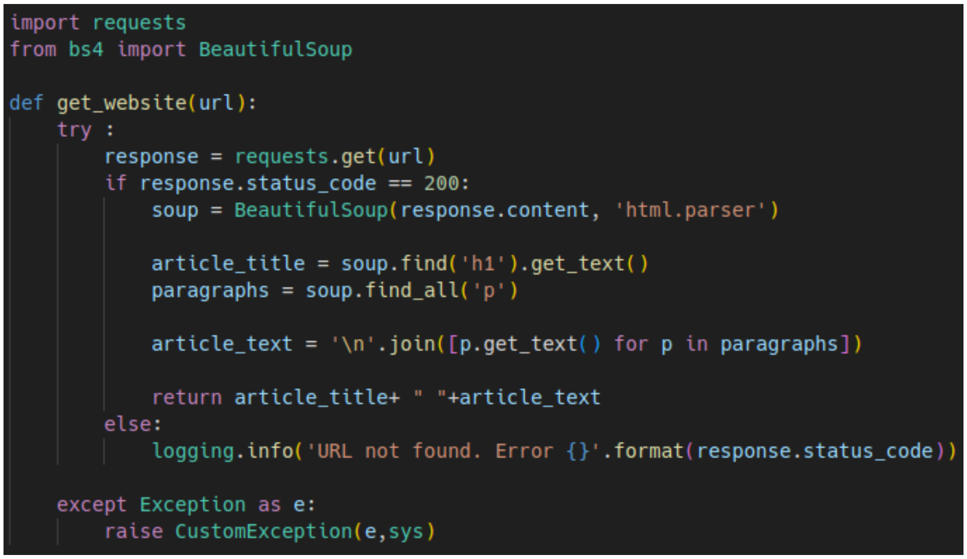
```
def gen_para_txt(file_name):
    try:
        paragraphs=[]
        with open(file_name, 'r') as file:
            text = file.read()
            paragraph = str(text).split('.\n')
            for para in paragraph:
                if (para):
                    paragraphs.append(para)
        return paragraphs
    except Exception as e:
        logging.info("File not found.")
        paragraph=[]
        return paragraph
```



```
def gen_para_file(file_name):
    file_name = os.path.join(folder,file_name)
    file_extension = os.path.splitext(file_name)[1]
    if (file_extension=='.pdf'):
        return gen_para_pdf(file_name)
    elif (file_extension == '.doc' or file_extension=='.docx'):
        return gen_para_doc(file_name)
    elif (file_extension=='.txt'):
        return gen_para_txt(file_name)
    logging.info("Paragraph of {} is generated.".format(file_name))
```

A general function that checks the filetype by the extension of the file and redirects the file to the specific function.

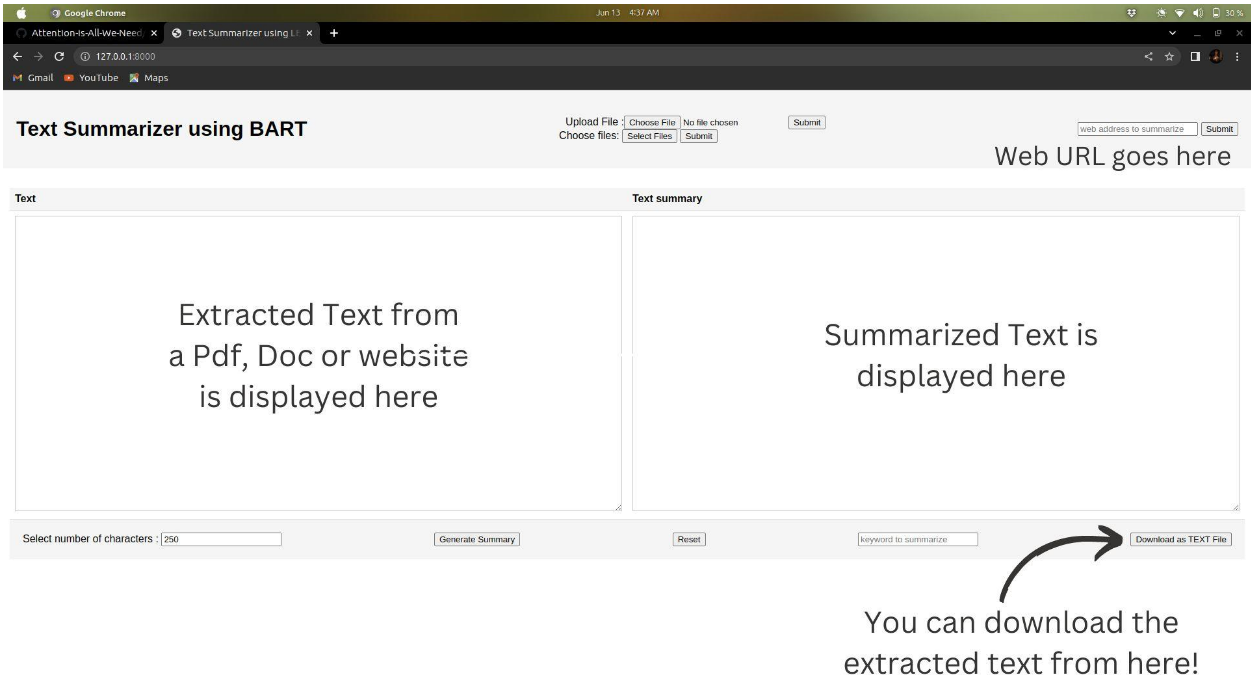
5.4 Web Pages : We used the Requests and BeautifulSoup libraries of Python to scrape web pages and extract the headings and paragraph elements from them.



```
import requests
from bs4 import BeautifulSoup

def get_website(url):
    try :
        response = requests.get(url)
        if response.status_code == 200:
            soup = BeautifulSoup(response.content, 'html.parser')

            article_title = soup.find('h1').get_text()
            paragraphs = soup.find_all('p')

            article_text = '\n'.join([p.get_text() for p in paragraphs])

            return article_title+ " "+article_text
        else:
            logging.info('URL not found. Error {}'.format(response.status_code))

    except Exception as e:
        raise CustomException(e,sys)
```

6. User Interface :

How to run: Look up the Readme file in the GIthub repo for instructions to run it.

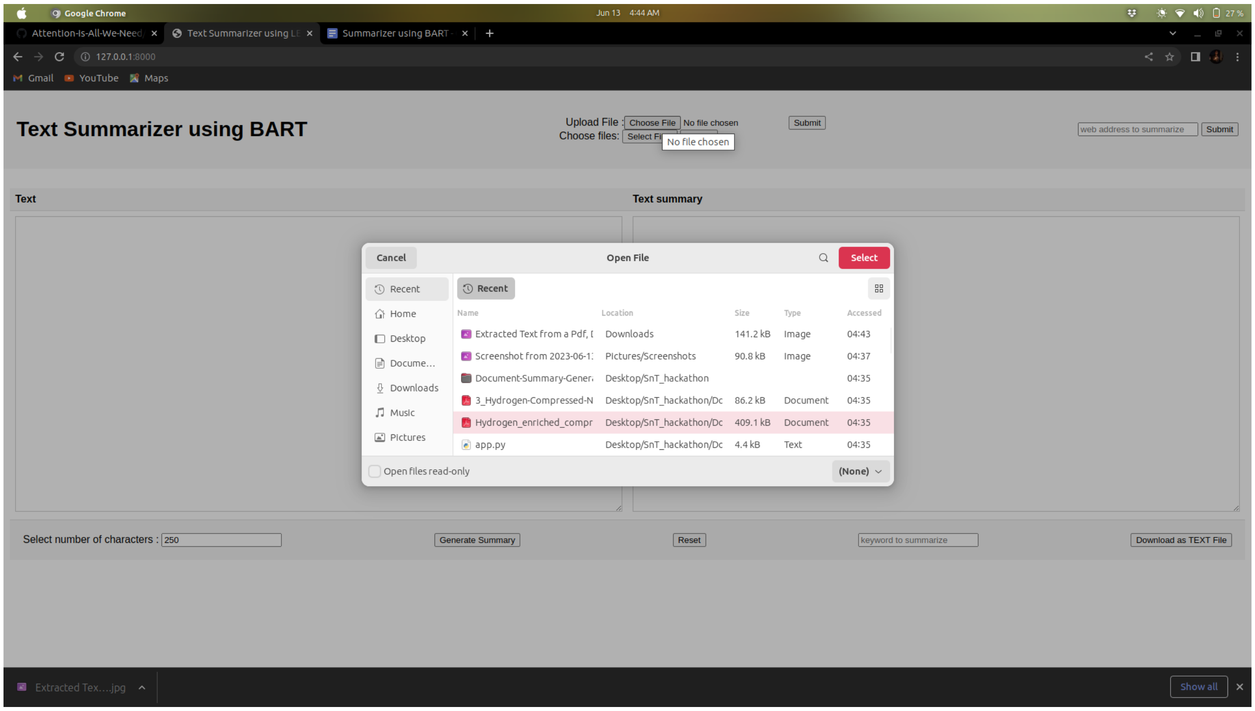


Web URL goes here

Extracted Text from a Pdf, Doc or website is displayed here

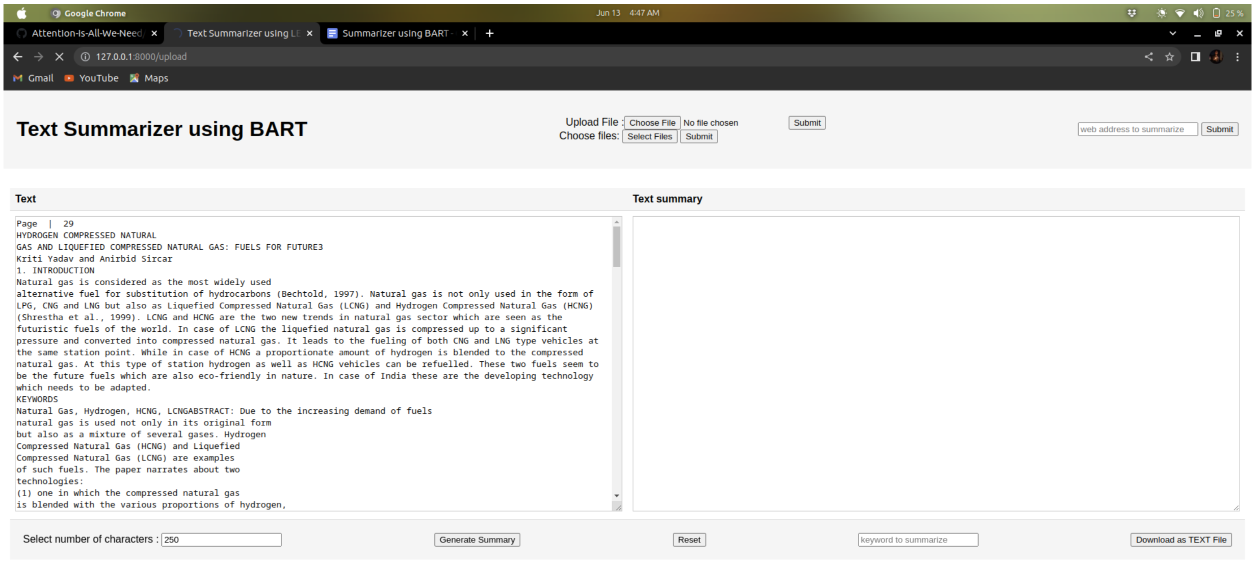
Summarized Text is displayed here

You can download the extracted text from here!



Name	Location	Size	Type	Accessed
Extracted Text from a Pdf, I	Downloads	141.2 kB	Image	04:43
Screenshot from 2023-06-1:	Pictures/Screenshots	90.8 kB	Image	04:37
Document-Summary-Genera	Desktop/SnT_hackathon			04:35
3_Hydrogen-Compressed-N	Desktop/SnT_hackathon/Dc	86.2 kB	Document	04:35
Hydrogen_enriched_compr	Desktop/SnT_hackathon/Dc	409.1 kB	Document	04:35
app.py	Desktop/SnT_hackathon/Dc	4.4 kB	Text	04:35

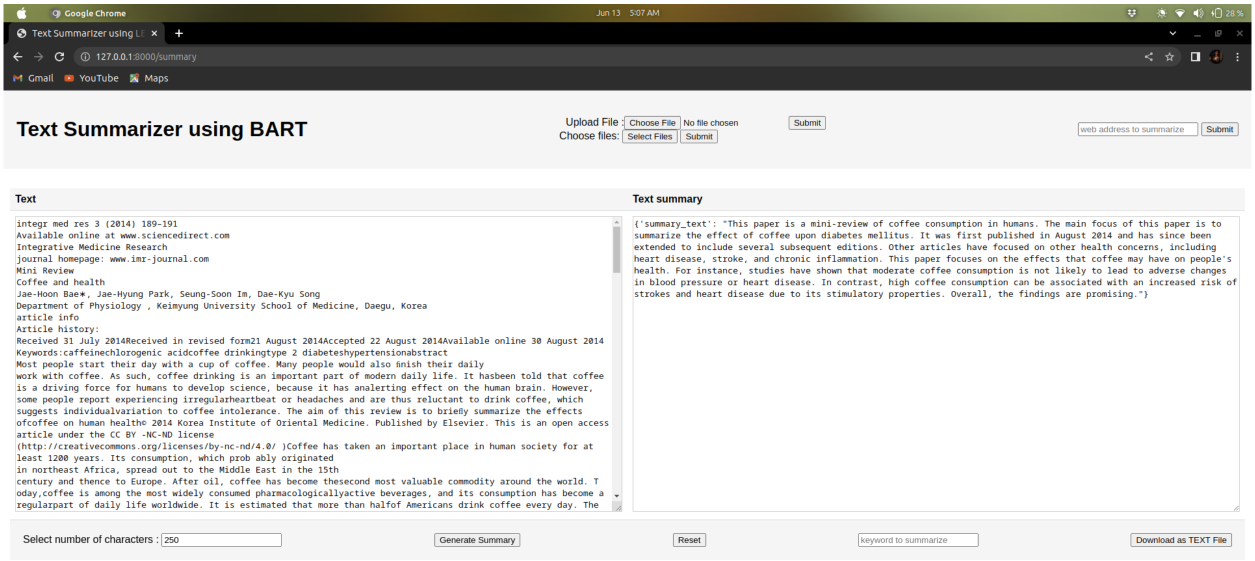
Use the choose file button to choose any PDF/DOC file from a directory. Alternatively, you may simply paste text in the text field. You can also upload multiple files from the Choose File button.



```
Page  |  29
HYDROGEN COMPRESSED NATURAL
GAS AND LIQUEFIED COMPRESSED NATURAL GAS: FUELS FOR FUTURE3
Kriti Yadav and Anirbid Sircar
1. INTRODUCTION
Natural gas is considered as the most widely used
alternative fuel for substitution of hydrocarbons (Bechtold, 1997). Natural gas is not only used in the form of
LPG, CNG and LNG but also as Liquefied Compressed Natural Gas (LCNG) and Hydrogen Compressed Natural Gas (HCNG)
(Shrestha et al., 1999). LCNG and HCNG are the two new trends in natural gas sector which are seen as the
futuristic fuels of the world. In case of LCNG the liquefied natural gas is compressed up to a significant
pressure and converted into compressed natural gas. It leads to the fueling of both CNG and LNG type vehicles at
the same station point. While in case of HCNG a proportionate amount of hydrogen is blended to the compressed
natural gas. At this type of station hydrogen as well as HCNG vehicles can be refuelled. These two fuels seem to
be the future fuels which are also eco-friendly in nature. In case of India these are the developing technology
which needs to be adapted.
KEYWORDS
Natural Gas, Hydrogen, HCNG, LCNGABSTRACT: Due to the increasing demand of fuels
natural gas is used not only in its original form
but also as a mixture of several gases. Hydrogen
Compressed Natural Gas (HCNG) and Liquefied
Compressed Natural Gas (LCNG) are examples
of such fuels. The paper narrates about two
technologies:
(1) one in which the compressed natural gas
is blended with the various proportions of hydrogen,
```

Select number of characters : 250

After uploading a file, press Submit and the text will be extracted from it and be displayed in the text field. You may then specify the number of words you desire in the summary. It is set to 250 by default.



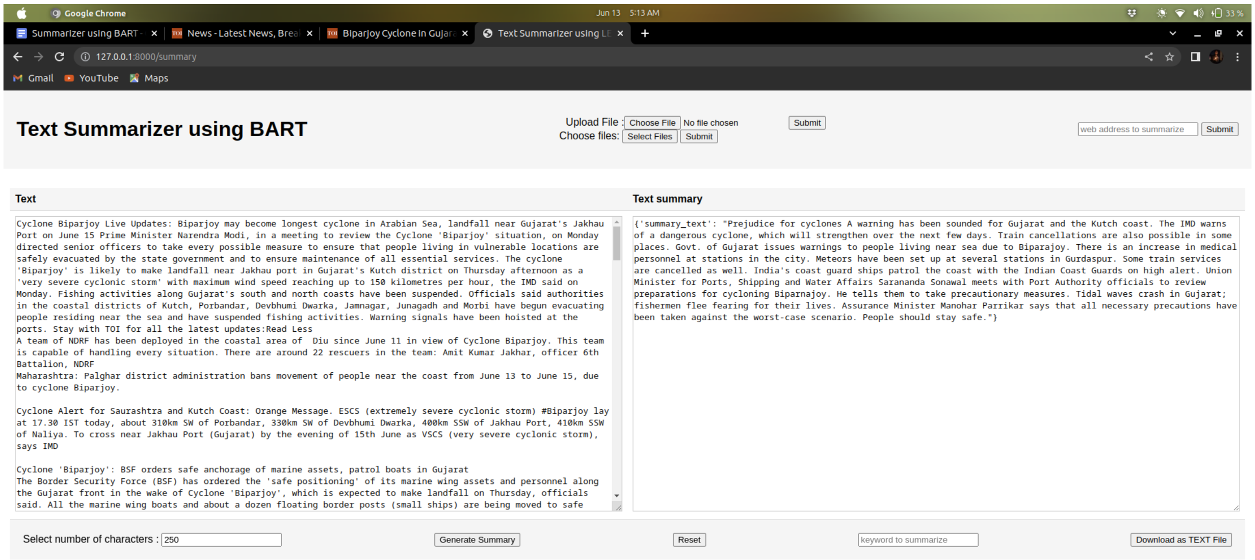
```
integr med res 3 (2014) 189–191
Available online at www.sciencedirect.com
Integrative Medicine Research
journal homepage: www.imr-journal.com
Mini Review
Coffee and health
Jae-Hoon Bae∗, Jae-Hyung Park, Seung-Soon Im, Dae-Kyu Song
Department of Physiology , Keimyung University School of Medicine, Daegu, Korea
article info
Article history:
Received 31 July 2014Received in revised form21 August 2014Accepted 22 August 2014Available online 30 August 2014
Keywords:caffeinechlorogenic acidcoffee drinkingtype 2 diabeteshypertensionabstract
Most people start their day with a cup of coffee. Many people would also finish their daily
work with coffee. As such, coffee drinking is an important part of modern daily life. It hasbeen told that coffee
is a driving force for humans to develop science, because it has analerting effect on the human brain. However,
some people report experiencing irregularheartbeat or headaches and are thus reluctant to drink coffee, which
suggests individualvariation to coffee intolerance. The aim of this review is to briefly summarize the effects
ofcoffee on human health© 2014 Korea Institute of Oriental Medicine. Published by Elsevier. This is an open access
article under the CC BY -NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/ )Coffee has taken an important place in human society for at
least 1200 years. Its consumption, which prob ably originated
in northeast Africa, spread out to the Middle East in the 15th
century and thence to Europe. After oil, coffee has become thesecond most valuable commodity around the world. T
oday,coffee is among the most widely consumed pharmacologicallyactive beverages, and its consumption has become a
regularpart of daily life worldwide. It is estimated that more than halfof Americans drink coffee every day. The
```

```
{'summary_text': "This paper is a mini-review of coffee consumption in humans. The main focus of this paper is to
summarize the effect of coffee upon diabetes mellitus. It was first published in August 2014 and has since been
extended to include several subsequent editions. Other articles have focused on other health concerns, including
heart disease, stroke, and chronic inflammation. This paper focuses on the effects that coffee may have on people's
health. For instance, studies have shown that moderate coffee consumption is not likely to lead to adverse changes
in blood pressure or heart disease. In contrast, high coffee consumption can be associated with an increased risk of
strokes and heart disease due to its stimulatory properties. Overall, the findings are promising."}
```

Select number of characters : 250

You can also summarize Blogs/news from different websites like Medium or TOI. Paste the website url in the text box in the top right side of the page.

Press “Generate Text” button and wait for a couple of minutes (Performance depends on the local system’s hardware). You will get the summarized text in the Text Summary field and press



Text Summarizer using BART

Upload File : No file chosen
Choose files:

Text

```
Cyclone Biparjoy Live Updates: Biparjoy may become longest cyclone in Arabian Sea, landfall near Gujarat's Jakhau
Port on June 15 Prime Minister Narendra Modi, in a meeting to review the Cyclone 'Biparjoy' situation, on Monday
directed senior officers to take every possible measure to ensure that people living in vulnerable locations are
safely evacuated by the state government and to ensure maintenance of all essential services. The cyclone
'Biparjoy' is likely to make landfall near Jakhau port in Gujarat's Kutch district on Thursday afternoon as a
'very severe cyclonic storm' with maximum wind speed reaching up to 150 kilometres per hour, the IMD said on
Monday. Fishing activities along Gujarat's south and north coasts have been suspended. Officials said authorities
in the coastal districts of Kutch, Porbandar, Devbhumi Dwarka, Jamnagar, Junagadh and Morbi have begun evacuating
people residing near the sea and have suspended fishing activities. Warning signals have been hoisted at the
ports. Stay with TOI for all the latest updates:Read Less
A team of NDRF has been deployed in the coastal area of  Diu since June 11 in view of Cyclone Biparjoy. This team
is capable of handling every situation. There are around 22 rescuers in the team: Amit Kumar Jakhar, officer 6th
Battalion, NDRF
Maharashtra: Palghar district administration bans movement of people near the coast from June 13 to June 15, due
to cyclone Biparjoy.

Cyclone Alert for Saurashtra and Kutch Coast: Orange Message. ESCS (extremely severe cyclonic storm) #Biparjoy lay
at 17.30 IST today, about 310km SW of Porbandar, 330km SW of Devbhumi Dwarka, 400km SSW of Jakhau Port, 410km SSW
of Naliya. To cross near Jakhau Port (Gujarat) by the evening of 15th June as VSCS (very severe cyclonic storm),
says IMD

Cyclone 'Biparjoy': BSF orders safe anchorage of marine assets, patrol boats in Gujarat
The Border Security Force (BSF) has ordered the 'safe positioning' of its marine wing assets and personnel along
the Gujarat front in the wake of Cyclone 'Biparjoy', which is expected to make landfall on Thursday, officials
said. All the marine wing boats and about a dozen floating border posts (small ships) are being moved to safe
```

Text summary

```
{'summary_text': "Prejudice for cyclones A warning has been sounded for Gujarat and the Kutch coast. The IMD warns
of a dangerous cyclone, which will strengthen over the next few days. Train cancellations are also possible in some
places. Govt. of Gujarat issues warnings to people living near sea due to Biparajoy. There is an increase in medical
personnel at stations in the city. Meteors have been set up at several stations in Gurdaspur. Some train services
are cancelled as well. India's coast guard ships patrol the coast with the Indian Coast Guards on high alert. Union
Minister for Ports, Shipping and Water Affairs Sarananda Sonawal meets with Port Authority officials to review
preparations for cycloning Biparnajoy. He tells them to take precautionary measures. Tidal waves crash in Gujarat;
fishermen flee fearing for their lives. Assurance Minister Manohar Parrikar says that all necessary precautions have
been taken against the worst-case scenario. People should stay safe."}
```

Select number of characters : 250

Also, if you want to store the summarized text in a txt file, press the Download as TEXT button in the bottom left corner.