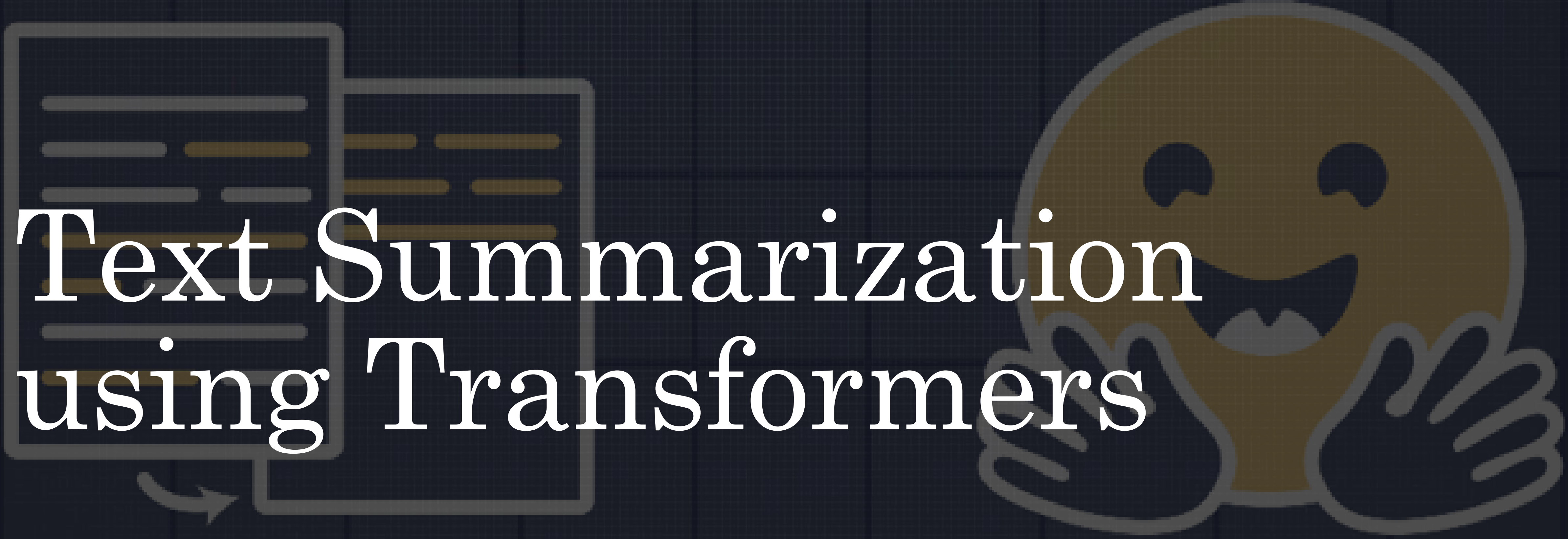
Text Summarization using Transformers



Presented By:
Shabnam Barbhuiya

OUTLINE:

Introduction on Text Summarization & Transformers

Why Use Transformers for Summarization?

How do transformers summarize text?

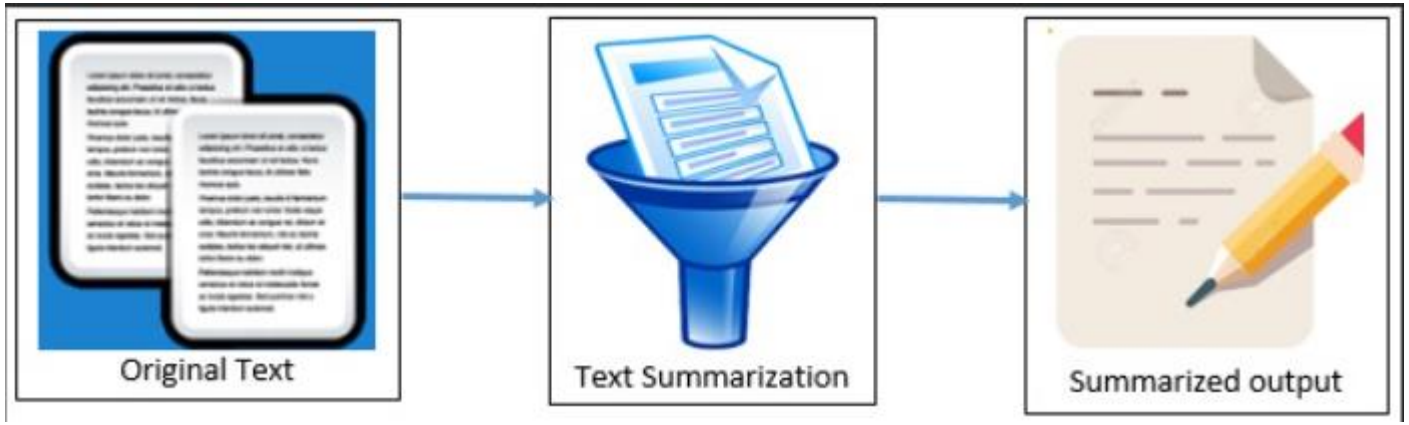
Process Overview

Implementation Steps

The Importance of Text Summarization in Today's World!

- Current Landscape:

Text summarization plays a pivotal role by distilling large volumes of text into concise, meaningful summaries



Source: https://www.analyticsvidhya.com

Real-World Applications:

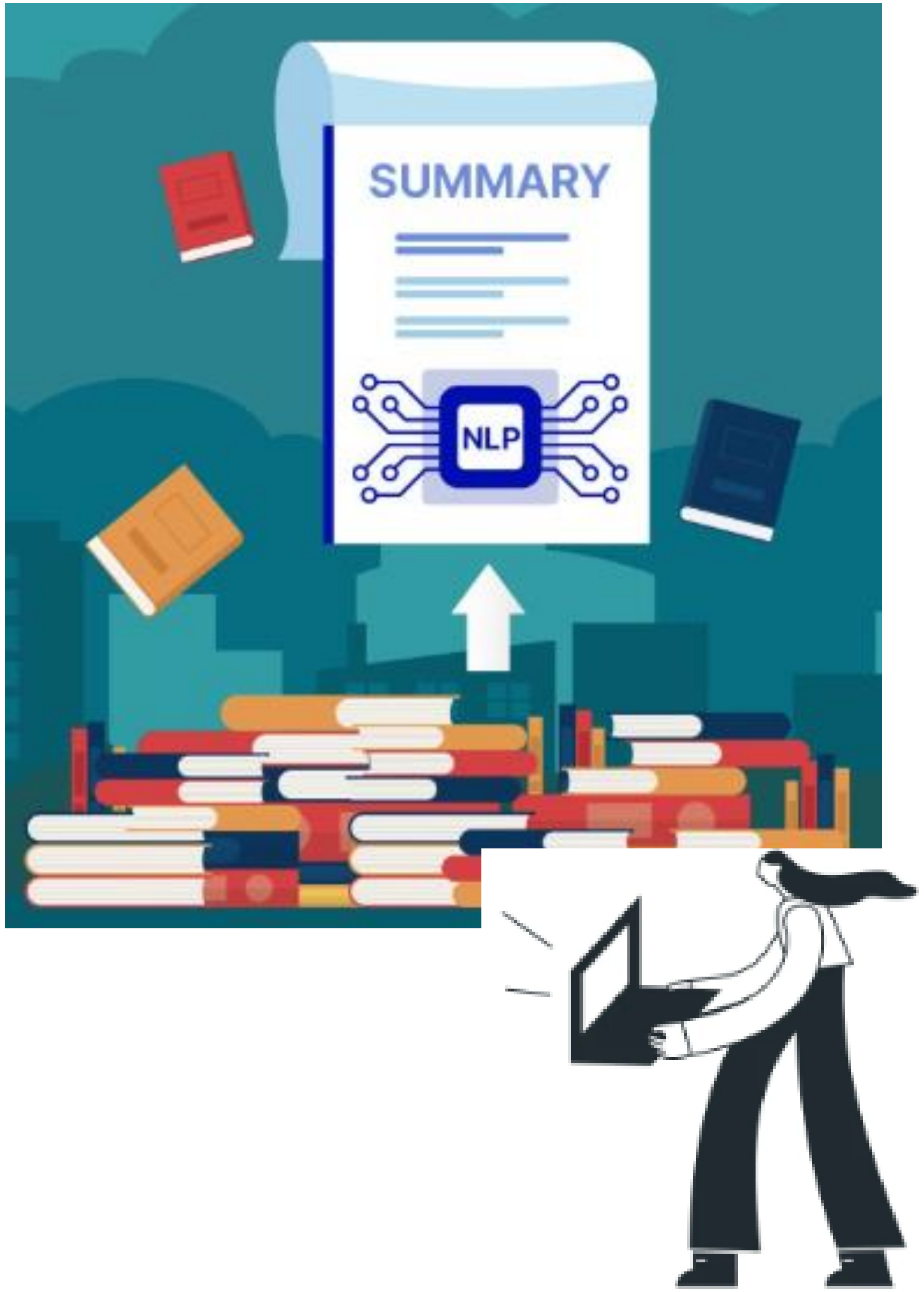
News Agencies: Automatically summarizing articles to deliver key headlines and insights swiftly.

Legal Firms: Summarizing lengthy legal documents and case studies to save time and enhance productivity.

Business Intelligence: Condensing reports and market analyses to inform strategic decisions.

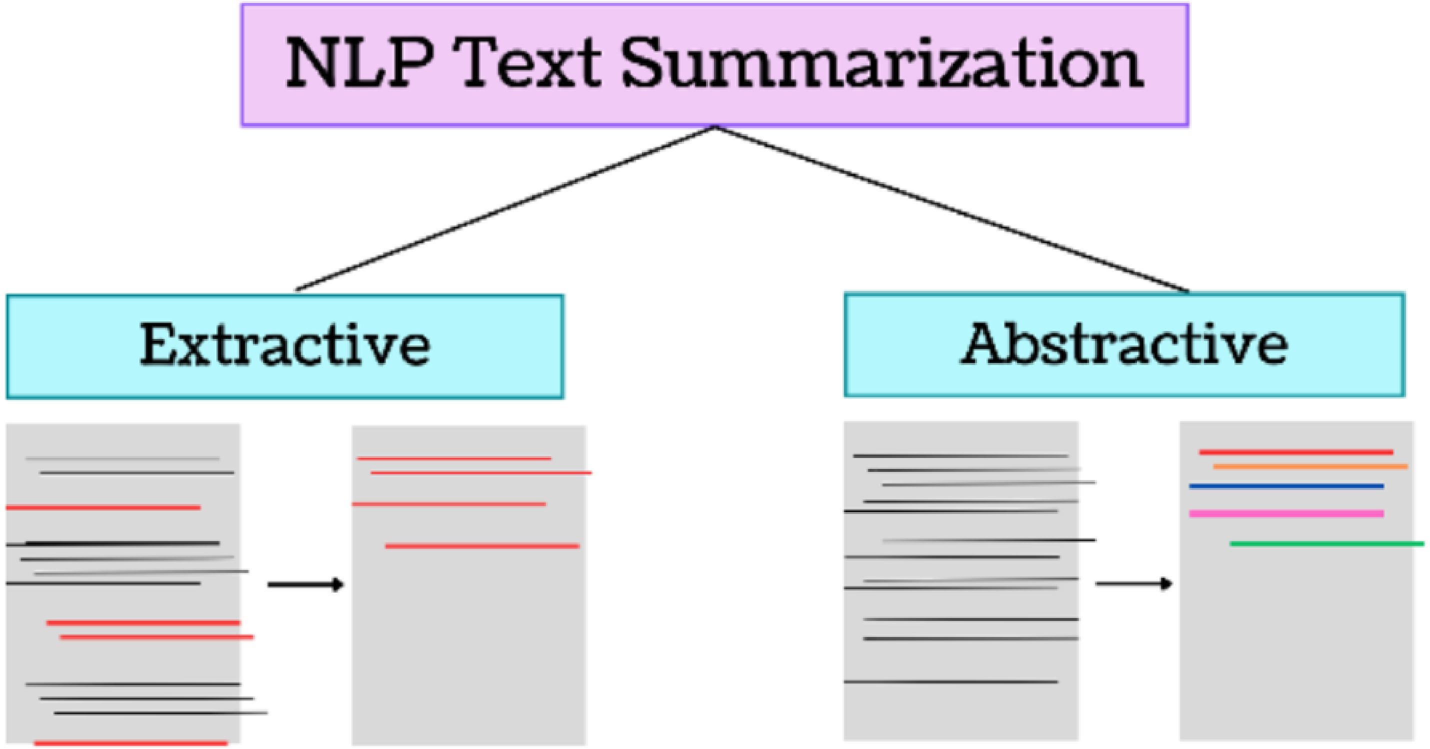
Healthcare: Summarizing patient records for quick review and efficient patient care.

Education: Creating concise study notes from extensive academic content, helping students grasp key concepts faster.



Types of Summarization

- **Abstractive Summarization:** Generating new sentences that capture the essence of the text.
- **Extractive Summarization:** Selecting key sentences from the text



Source: https://texta.ai/

What are transformers?

Transformers are a type of neural network architecture introduced in 2017 by Vaswani et al. in the paper "Attention Is All You Need".

They revolutionized the field of Natural Language Processing (NLP). Transformers have are used for various NLP tasks, including text summarization, translation, and sentiment analysis.

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Source https://doi.org/10.48550/arXiv.1706.03762

Transformer Architecture

Encoder:

- **Input Embedding**: Converts input tokens into vectors.
- **Positional Encoding**: Adds position information to the embeddings.
- **Multi-Head Attention**: Allows the model to focus on different parts of the input.
- **Feed Forward**: Applies a fully connected network to each token.

Decoder:

- **Output Embedding**: Converts output tokens into vectors.
- **Masked Multi-Head Attention**: Prevents the model from looking ahead at future tokens.
- **Multi-Head Attention**: Attends to encoder outputs for context.

Output:

- **Linear**: Converts decoder output to vocabulary-sized vector.
- **Softmax**: Produces a probability distribution for the next token prediction.

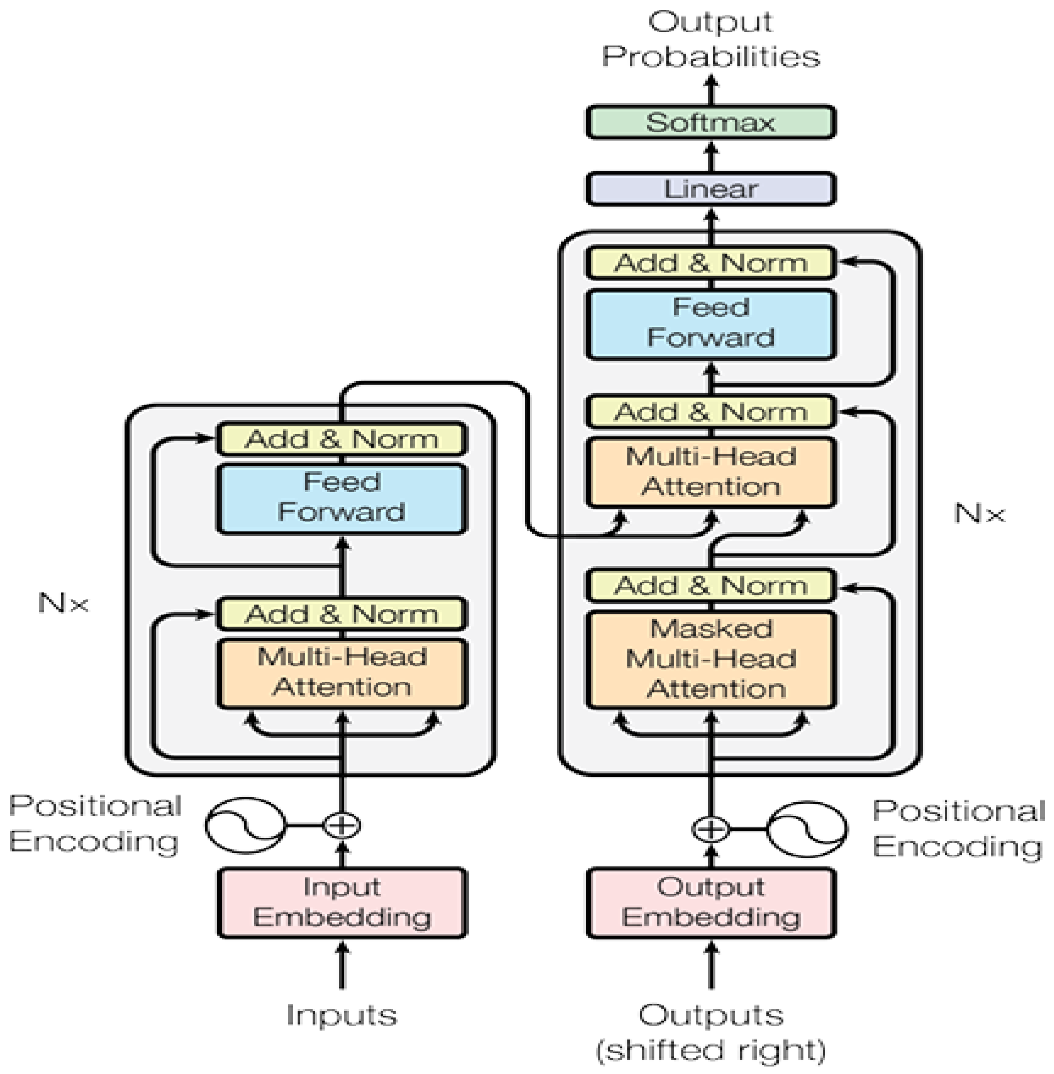


Figure 1: The Transformer - model architecture.

Source https://doi.org/10.48550/arXiv.1706.03762

Why Use Transformers for Summarization?

Task-Specific Optimization:trained on summarization datasets (e.g., CNN/DailyMail or PubMed), making them highly specialized.

perplexity

ChatGPT

Cost Considerations

Domain-Specific Pretrained Models

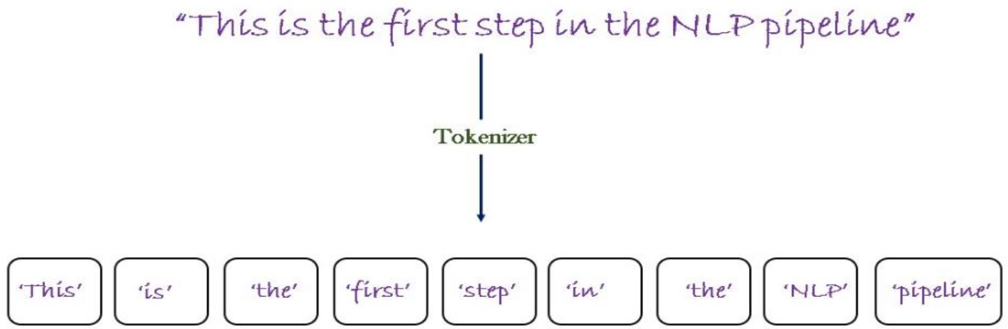
- **BioBERT** for biomedical texts.
- **LegalBERT** for legal documents.
- **FinBERT** for financial texts.

How do transformers summarize text?

Input Encoding: The input text is tokenized and converted into a format suitable for the transformer model.

Self-Attention Mechanism: The model uses self-attention to focus on different parts of the text, learning which words or phrases are important for the summary.

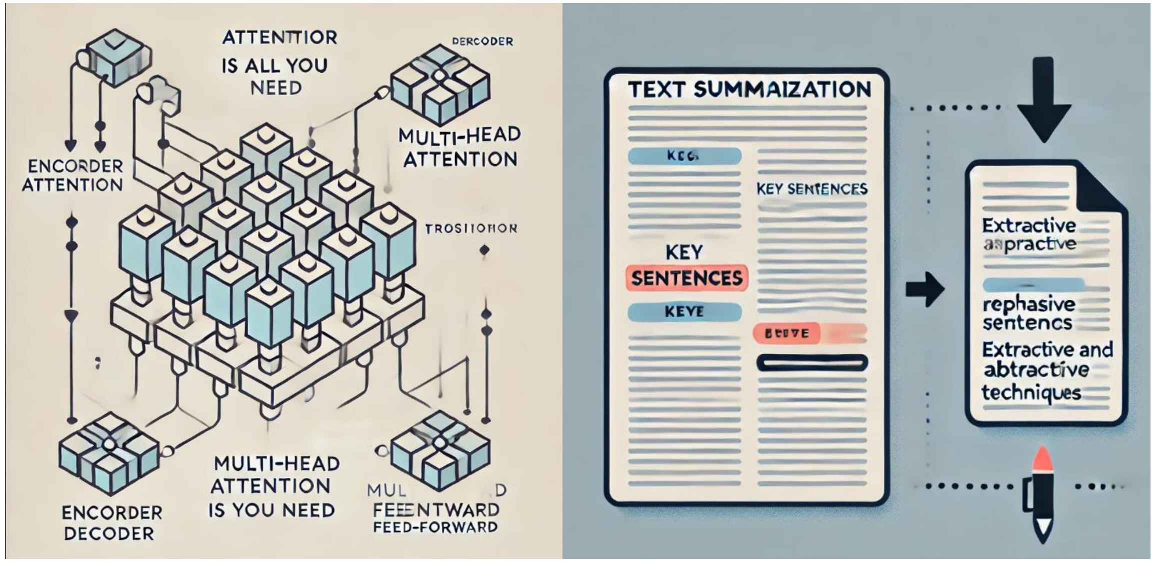
Output Decoding: The model generates a summarized version of the text, either by selecting key sentences (extractive) or by rephrasing and condensing information (abstractive).



"This is the first step in the NLP pipeline"

Tokenizer

'This' 'is' 'the' 'first' 'step' 'in' 'the' 'NLP' 'pipeline'



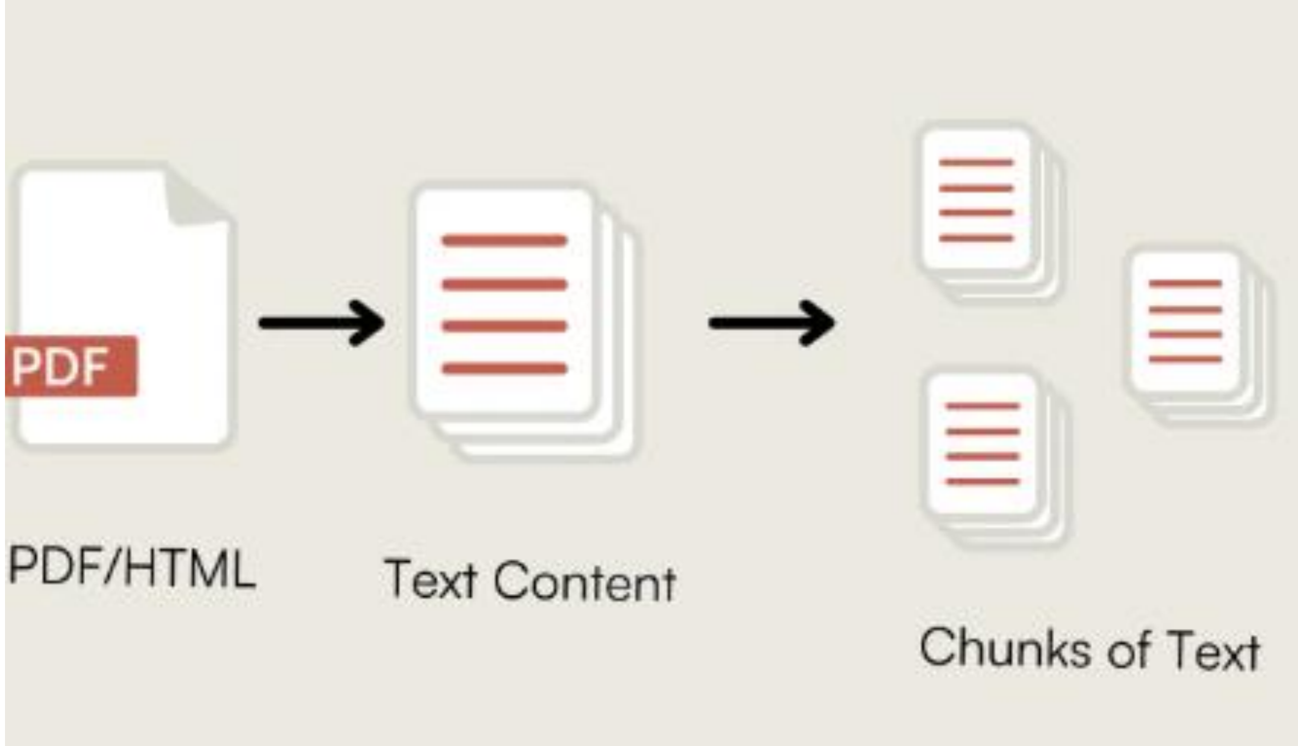
Ref:Meta AI

Process Overview

Text Extraction: using libraries like PyMuPDF

Preprocessing: Preprocess the extracted text by tokenizing it.

Text Chunking: split the preprocessed text into smaller.

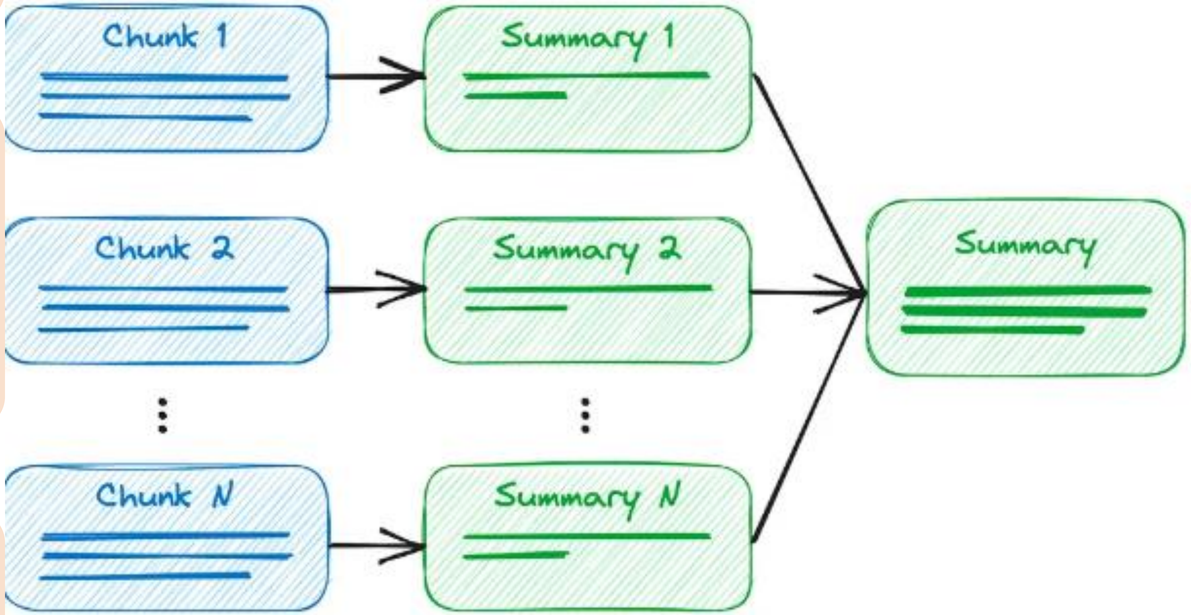


Source: https://www.analyticsvidhya.com

Model Selection: Select a pre-trained transformer model such as BERT, RoBERTa, or XLNet, or sshleifer/distilbart-cnn-12-6.

Summarization with the Model: Use the fine-tuned model to summarize each chunk of text.

Output Display



Source: https://www.analyticsvidhya.com

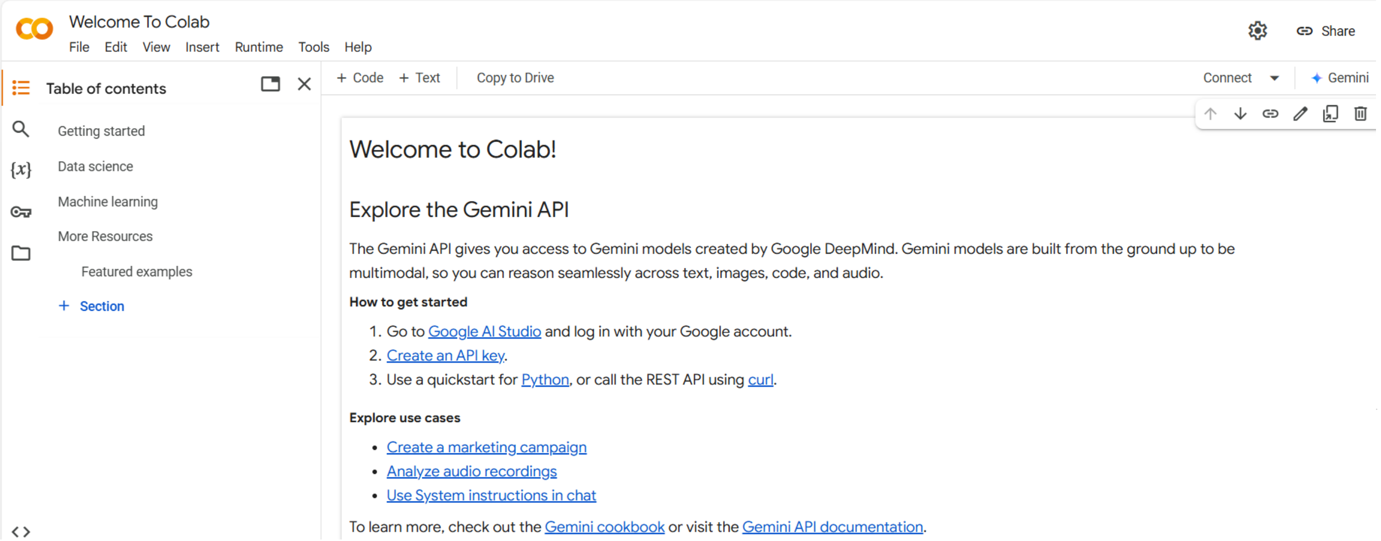
HOW ABOUT YOU DO IT YOUR SELF?

URL: https://colab.research.google.com/drive/1v-cuk4gpKMDlBDgbNFyxuElFpwKTE1ok?usp=sharing

Implementation Steps(DIY)

1. Setting Up the Environment

- Virtual Environment (Optional but Recommended)



Welcome to Colab!

Explore the Gemini API

The Gemini API gives you access to Gemini models created by Google DeepMind. Gemini models are built from the ground up to be multimodal, so you can reason seamlessly across text, images, code, and audio.

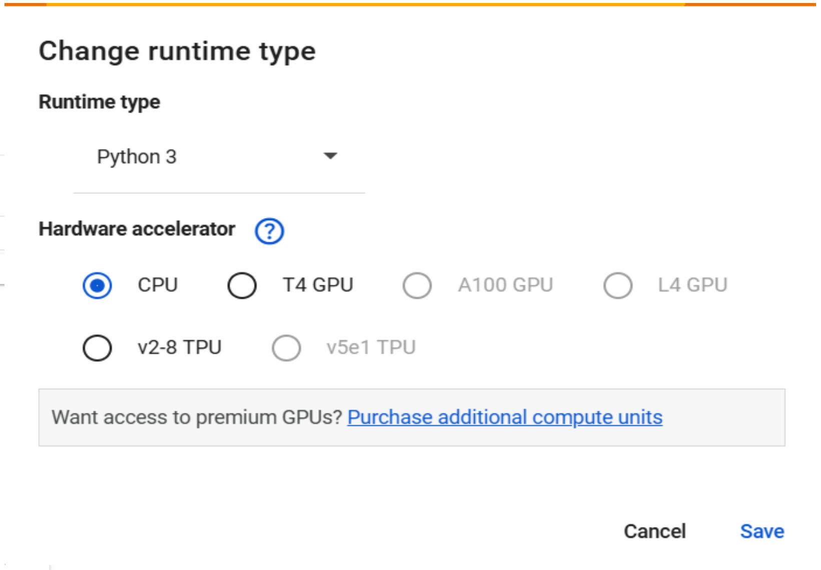
How to get started

1. Go to Google AI Studio and log in with your Google account.
2. Create an API key.
3. Use a quickstart for Python, or call the REST API using curl.

Explore use cases

- Create a marketing campaign
- Analyze audio recordings
- Use System instructions in chat

To learn more, check out the Gemini cookbook or visit the Gemini API documentation.



Change runtime type

Runtime type: Python 3

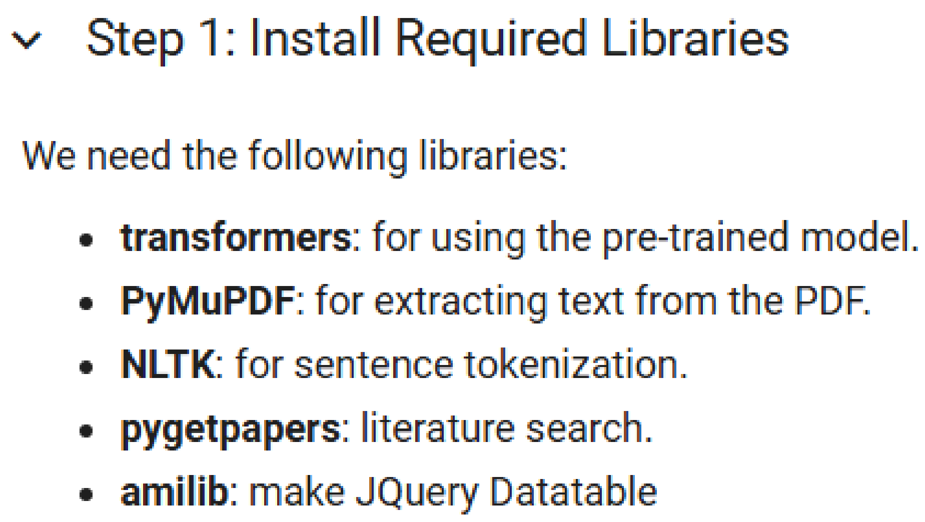
Hardware accelerator

● CPU ○ T4 GPU ○ A100 GPU ○ L4 GPU

○ v2-8 TPU ○ v5e1 TPU

Want access to premium GPUs? Purchase additional compute units

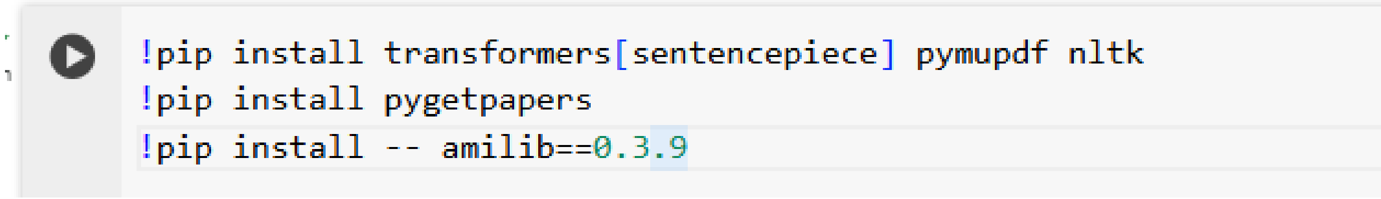
2. Installing Required Libraries



Step 1: Install Required Libraries

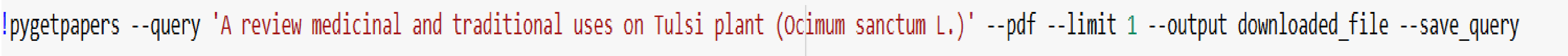
We need the following libraries:

- **transformers**: for using the pre-trained model.
- **PyMuPDF**: for extracting text from the PDF.
- **NLTK**: for sentence tokenization.
- **pygetpapers**: literature search.
- **amilib**: make JQuery Datatable

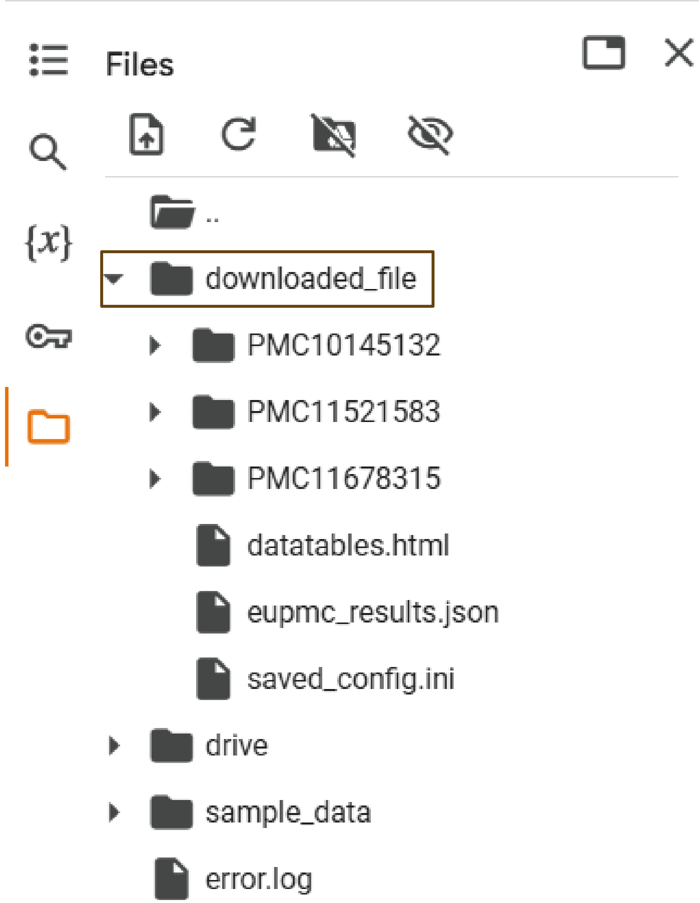


```
!pip install transformers[sentencepiece] pymupdf nltk
!pip install pygetpapers
!pip install -- amilib==0.3.9
```

3.Download research paper using pygetpapers

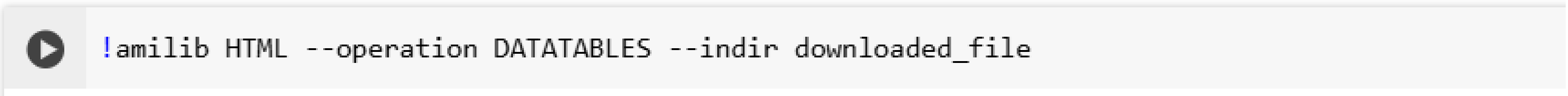


```
!pygetpapers --query 'A review medicinal and traditional uses on Tulsi plant (Ocimum sanctum L.)' --pdf --limit 1 --output downloaded_file --save_query
```

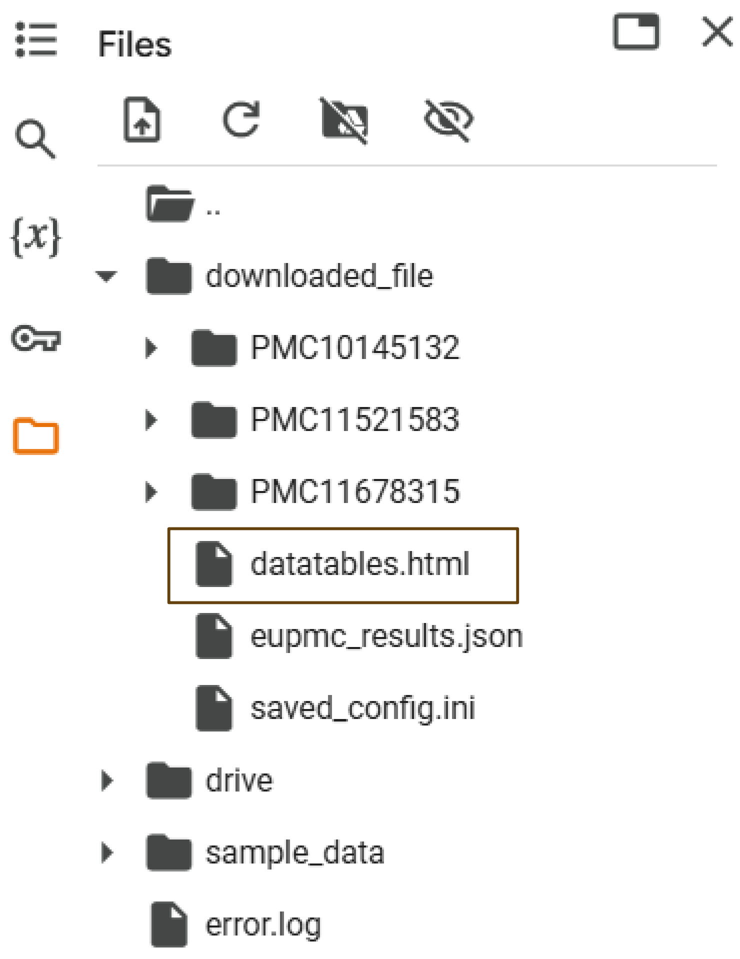


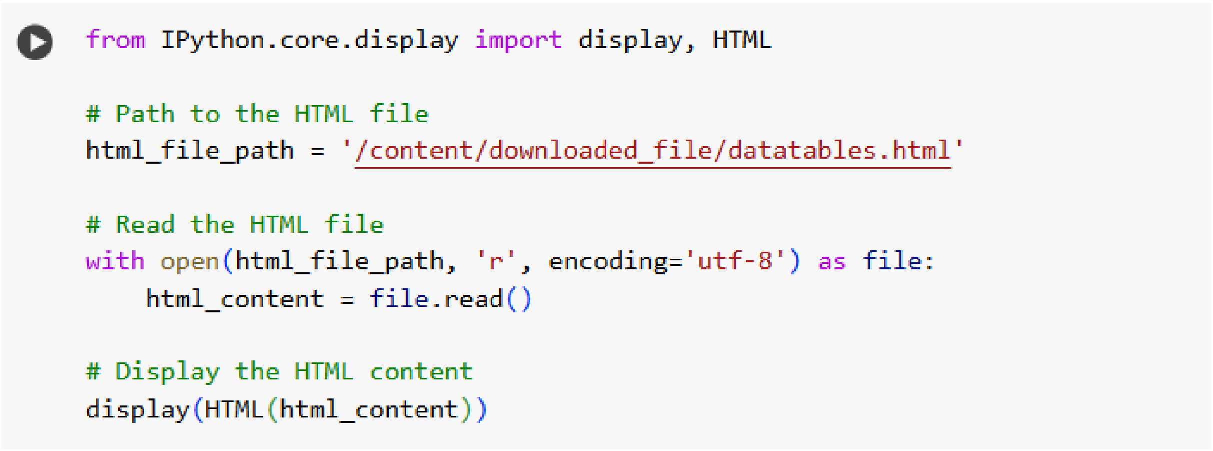
```
..
downloaded_file
    PMC10145132
    PMC11521583
    PMC11678315
    datatables.html
    eupmc_results.json
    saved_config.ini
drive
sample_data
error.log
```

4. Create Datatables for the retrieved articles using amilib

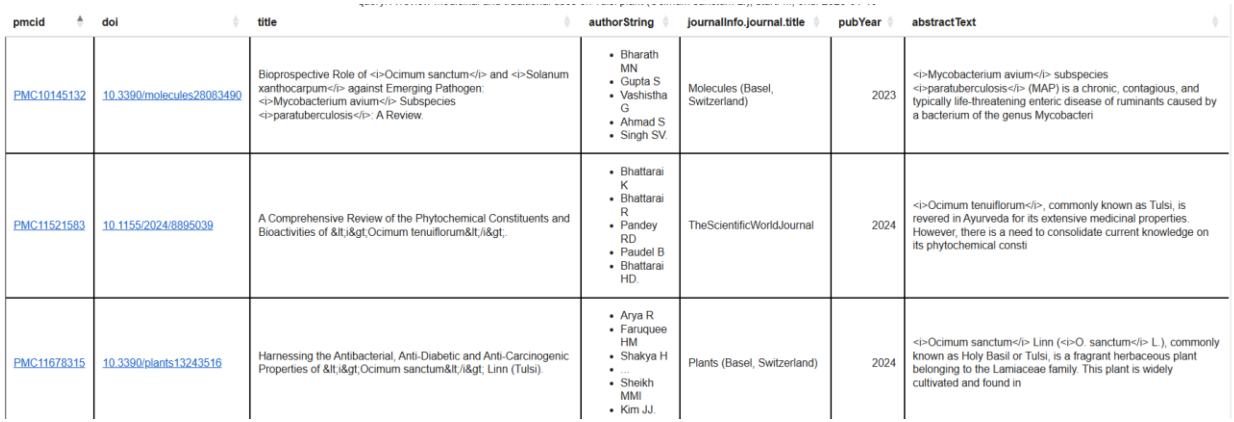


```
!amilib HTML --operation DATATABLES --indir downloaded_file
```



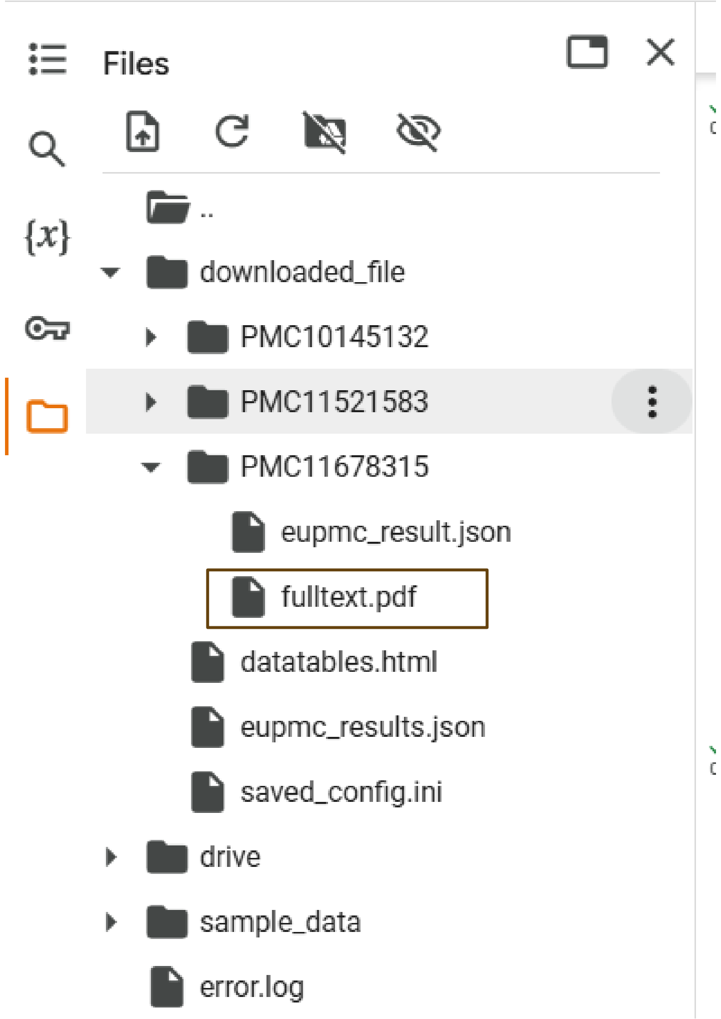


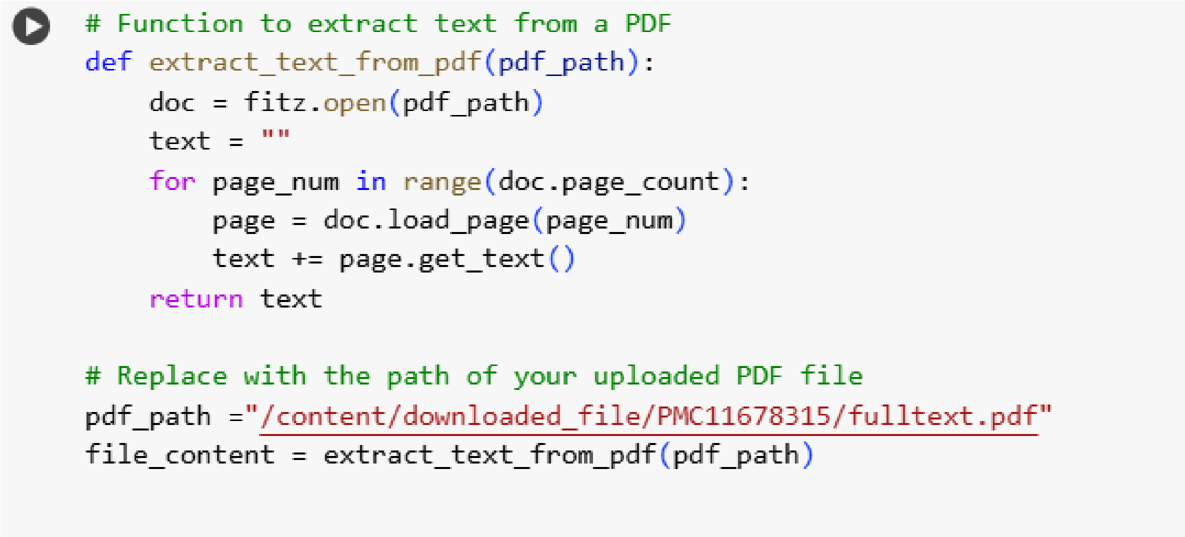
```python
from IPython.core.display import display, HTML

# Path to the HTML file
html_file_path = '/content/downloaded_file/datatables.html'

# Read the HTML file
with open(html_file_path, 'r', encoding='utf-8') as file:
    html_content = file.read()

# Display the HTML content
display(HTML(html_content))
```



pmcid	doi	title	authorString	journalInfo.journal.title	pubYear	abstractText
PMC10145132	10.3390/molecules28083490	Bioprospective Role of <i>Ocimum sanctum</i> and <i>Solanum xanthocarpum</i> against Emerging Pathogen: <i>Mycobacterium avium</i> Subspecies <i>paratuberculosis</i>: A Review.	• Bharath MN • Gupta S • Vashistha G • Ahmad S • Singh SV.	Molecules (Basel, Switzerland)	2023	<i>Mycobacterium avium</i> subspecies <i>paratuberculosis</i> (MAP) is a chronic, contagious, and typically life-threatening enteric disease of ruminants caused by a bacterium of the genus Mycobacteri
PMC11521583	10.1155/2024/8895039	A Comprehensive Review of the Phytochemical Constituents and Bioactivities of <i>Ocimum tenuiflorum</i>.	• Bhattarai K • Bhattarai R • Pandey RD • Paudel B • Bhattarai HD.	TheScientificWorldJournal	2024	<i>Ocimum tenuiflorum</i>, commonly known as Tulsi, is revered in Ayurveda for its extensive medicinal properties. However, there is a need to consolidate current knowledge on its phytochemical consti
PMC11678315	10.3390/plants13243516	Harnessing the Antibacterial, Anti-Diabetic and Anti-Carcinogenic Properties of <i>Ocimum sanctum</i> Linn (Tulsi).	• Arya R • Faruquee HM • Shakya H • ... • Sheikh MMI • Kim JJ.	Plants (Basel, Switzerland)	2024	<i>Ocimum sanctum</i> Linn (<i>O. sanctum</i> L.), commonly known as Holy Basil or Tulsi, is a fragrant herbaceous plant belonging to the Lamiaceae family. This plant is widely cultivated and found in

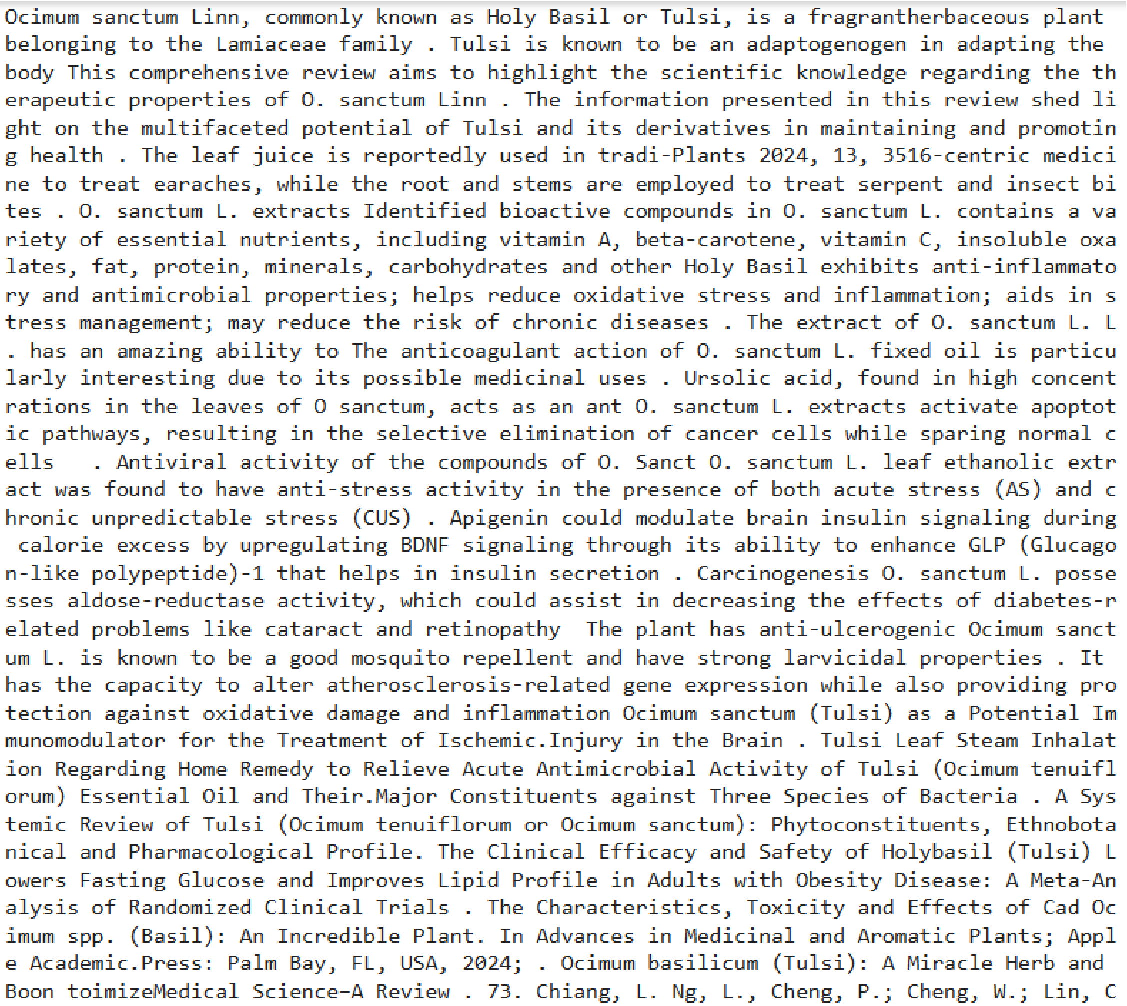
5.Loading and extracting text from pdf.





```python
# Function to extract text from a PDF
def extract_text_from_pdf(pdf_path):
    doc = fitz.open(pdf_path)
    text = ""
    for page_num in range(doc.page_count):
        page = doc.load_page(page_num)
        text += page.get_text()
    return text

# Replace with the path of your uploaded PDF file
pdf_path ="/content/downloaded_file/PMC11678315/fulltext.pdf"
file_content = extract_text_from_pdf(pdf_path)
```

OUTPUT:



```
Ocimum sanctum Linn, commonly known as Holy Basil or Tulsi, is a fragrantherbaceous plant
belonging to the Lamiaceae family . Tulsi is known to be an adaptogenogen in adapting the
body This comprehensive review aims to highlight the scientific knowledge regarding the th
erapeutic properties of O. sanctum Linn . The information presented in this review shed li
ght on the multifaceted potential of Tulsi and its derivatives in maintaining and promotin
g health . The leaf juice is reportedly used in tradi-Plants 2024, 13, 3516-centric medici
ne to treat earaches, while the root and stems are employed to treat serpent and insect bi
tes . O. sanctum L. extracts Identified bioactive compounds in O. sanctum L. contains a va
riety of essential nutrients, including vitamin A, beta-carotene, vitamin C, insoluble oxa
lates, fat, protein, minerals, carbohydrates and other Holy Basil exhibits anti-inflammato
ry and antimicrobial properties; helps reduce oxidative stress and inflammation; aids in s
tress management; may reduce the risk of chronic diseases . The extract of O. sanctum L. L
. has an amazing ability to The anticoagulant action of O. sanctum L. fixed oil is particu
larly interesting due to its possible medicinal uses . Ursolic acid, found in high concent
rations in the leaves of O sanctum, acts as an ant O. sanctum L. extracts activate apoptot
ic pathways, resulting in the selective elimination of cancer cells while sparing normal c
ells   . Antiviral activity of the compounds of O. Sanct O. sanctum L. leaf ethanolic extr
act was found to have anti-stress activity in the presence of both acute stress (AS) and c
hronic unpredictable stress (CUS) . Apigenin could modulate brain insulin signaling during
 calorie excess by upregulating BDNF signaling through its ability to enhance GLP (Glucago
n-like polypeptide)-1 that helps in insulin secretion . Carcinogenesis O. sanctum L. posse
sses aldose-reductase activity, which could assist in decreasing the effects of diabetes-r
elated problems like cataract and retinopathy  The plant has anti-ulcerogenic Ocimum sanct
um L. is known to be a good mosquito repellent and have strong larvicidal properties . It
has the capacity to alter atherosclerosis-related gene expression while also providing pro
tection against oxidative damage and inflammation Ocimum sanctum (Tulsi) as a Potential Im
munomodulator for the Treatment of Ischemic.Injury in the Brain . Tulsi Leaf Steam Inhalat
ion Regarding Home Remedy to Relieve Acute Antimicrobial Activity of Tulsi (Ocimum tenuifl
orum) Essential Oil and Their.Major Constituents against Three Species of Bacteria . A Sys
temic Review of Tulsi (Ocimum tenuiflorum or Ocimum sanctum): Phytoconstituents, Ethnobota
nical and Pharmacological Profile. The Clinical Efficacy and Safety of Holybasil (Tulsi) L
owers Fasting Glucose and Improves Lipid Profile in Adults with Obesity Disease: A Meta-An
alysis of Randomized Clinical Trials . The Characteristics, Toxicity and Effects of Cad Oc
imum spp. (Basil): An Incredible Plant. In Advances in Medicinal and Aromatic Plants; Appl
e Academic.Press: Palm Bay, FL, USA, 2024; . Ocimum basilicum (Tulsi): A Miracle Herb and
Boon toimizeMedical Science–A Review . 73. Chiang, L. Ng, L., Cheng, P.; Cheng, W.; Lin, C
```

5.Generating and evaluating the summary.

It's Your Turn!

- Time to Try It Out
- The code is ready for you in Google Colab.
- Follow along with the provided notebook.
- Feel free to ask questions as you go!

Let’s Dive In!

Thank you.

Shabnam Barbhuiya