News Tool Searcher

NAME – DEV SINGHMAR

ROLL NO - 21IT3017

MENTOR NAME – DR. SUSHAM BISWAS

TABLE OF CONTENTS

Problem Statement

- **Information Overload**: With the vast amount of information available online, users may struggle to find relevant information quickly and efficiently.

- **Time-Consuming Manual Search**: Manually searching through multiple news articles to find answers to specific questions can be time-consuming and tedious.

- So we need a smart tool that can quickly find important information from lots of news articles all at once, making it easy for people to get the details they need without wasting time.

Proposed Approach

It involves building a tool using Streamlit and various libraries to automate the extraction of information from multiple news articles. This includes:

Data Retrieval: Scraping content from provided URLs using BeautifulSoup and loading it into the tool.

Text Processing: Splitting the retrieved text into smaller chunks for efficient processing.

Text Embedding: Utilizing a pre-trained Sentence Transformer model to convert text into numerical vectors (embeddings).

Indexing: Creating an index using FAISS to efficiently search through the embeddings.

User Interaction: Allowing users to input a question and retrieving relevant information from the indexed text using a pre-trained question-answering model.

Why are we not using ChatGPT??

Copy pasting articles in ChatGPT is tedious task

Research analysts are busy folks they don't have time to go to website copy paste and then get the answer.

They also need an aggregate knowledge base because when they are asking questions they don't know where the answer might be.
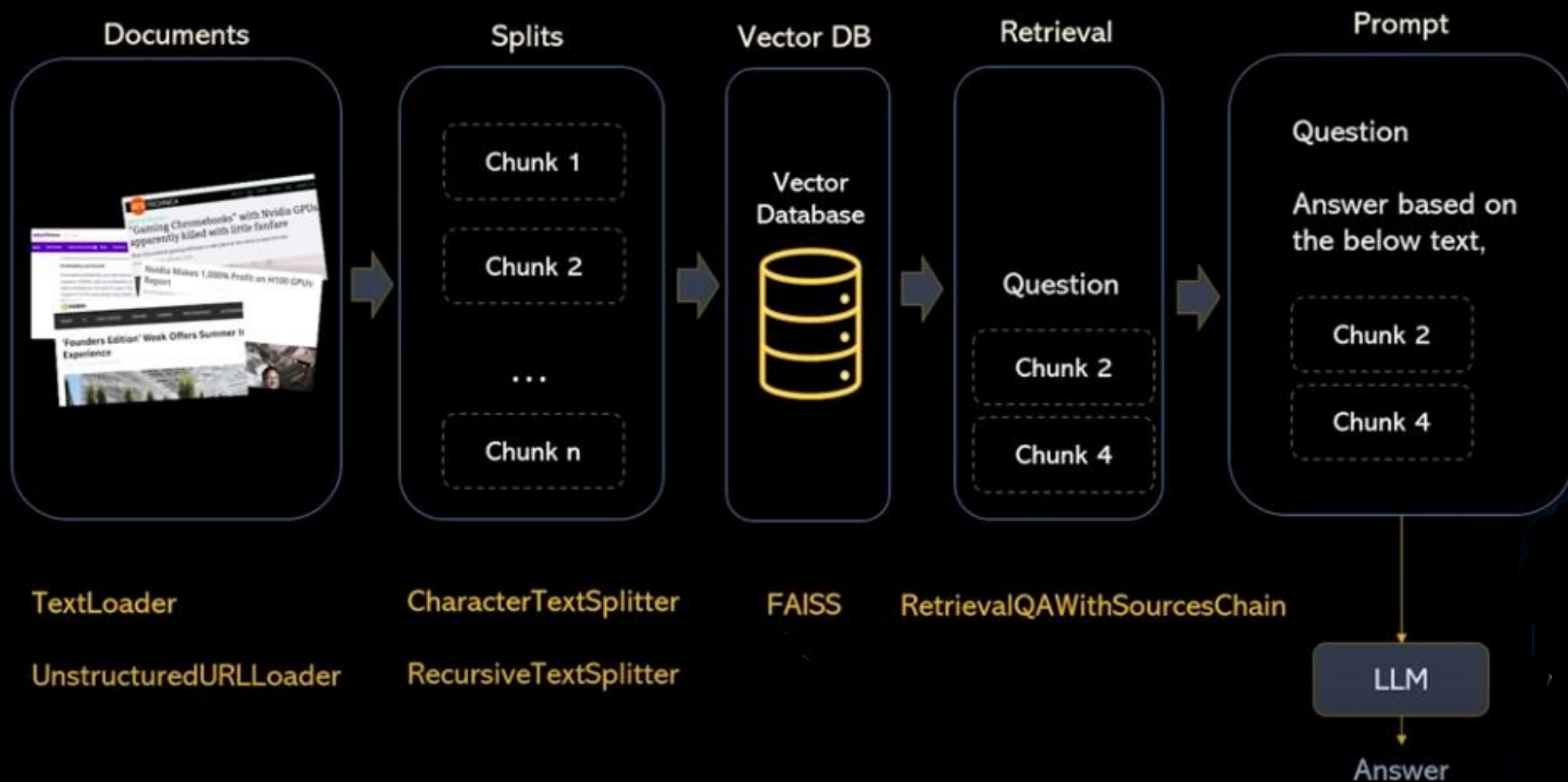
Chatgpt's word limit in chat GPT you can't copy paste a huge

Article has a limit on number of words you can supply

Workflow

Documents	Splits	Vector DB	Retrieval	Prompt

Documents → Splits (Chunk 1, Chunk 2, ... Chunk n) → Vector DB (Vector Database) → Retrieval (Question, Chunk 2, Chunk 4) → Prompt (Question, Answer based on the below text, Chunk 2, Chunk 4) → LLM → Answer

TextLoader   CharacterTextSplitter   FAISS   RetrievalQAWithSourcesChain

UnstructuredURLLoader   RecursiveTextSplitter

Results

Screenshot 1

News Article URLs

URL 1
`https://www.moneycontrol.com/news/bus`

URL 2
`https://www.moneycontrol.com/news/bus`

URL 3
`https://www.moneycontrol.com/news/bus`

Process URLs

News Research Tool 📈

Question:

What is the price of tata punch?

Answer

between Rs 7.1 lakh and Rs 9.68 lakh

Made with Streamlit

Screenshot 2

News Article URLs

URL 1
`https://www.moneycontrol.com/news/bus`

URL 2
`https://www.moneycontrol.com/news/bus`

URL 3
`https://www.moneycontrol.com/news/bus`

Process URLs

News Research Tool 📈

Question:

Incentive programme covers how many sectors

## Answer

14 sectors

Made with Streamlit

Conclusion
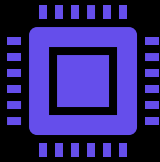
Efficiency and Automation: The tool streamlines the process of information extraction, saving users valuable time and effort by automating data retrieval, processing, and answering tasks.

Enhanced User Experience: With a user-friendly interface provided by Streamlit, users can easily input URLs, pose questions, and receive relevant answers, making the tool accessible and convenient to use.
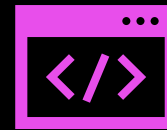
Technological Integration: The project demonstrates the integration of various technologies, including web scraping, natural language processing, and machine learning, to create an efficient and effective information retrieval system.

Impact and Utility: The developed tool has the potential to have a significant impact on various domains where quick access to relevant information from multiple sources is essential, such as journalism, research, and decision-making processes.

Future Scope

Integrating the news research tool with a stock prediction chart bot that harnesses real-time news updates and sentiment analysis to enhance predictive accuracy, enriching models with dynamic features and adaptive learning mechanisms.