CS-5821 MACHINE LEARNING

FINAL PROJECT

GENERATIVE AI (NLP)
Chatbot for Custom Data

TEAM : STRAW HATS

ADESH RAVICHANDRAN

HARIKRISHNA REDDY BENAKANA

NIHARIKA NAIK DHANAVATH

ROHAN VENKATESHA

SATHYANARAYANA RAMESH

LET'S TAKE A QUICK RECAP:

INTRODUCTION

A custom chatbot using RAG implementation to provide tailored responses and insights from textual content sourced from PDF documents and images.

What’s the need for Custom Chatbot

01.

Halucination faced by current models

A custom chatbot mitigates the issue of limited context understanding, ensuring accurate responses to complex user queries and improving overall user satisfaction.

02.

Enhanced Privacy

By tailoring the chatbot to specific documents or images, we maintain control over sensitive data, ensuring privacy and compliance with regulations.

NLP / MACHINE LEARING

NATURAL LANGUAGE PROCESSING

NLP techniques handle text processing and analysis tasks.

Tasks include extracting text from PDFs and understanding user inquiries.

NLP enables machines to comprehend and interact with human language in diverse contexts.

EMBEDDINGS

Embeddings represent words, phrases, or documents as numerical vectors.

These vectors are situated in a continuous vector space.

Embeddings empower machine learning models to better comprehend and process natural language.

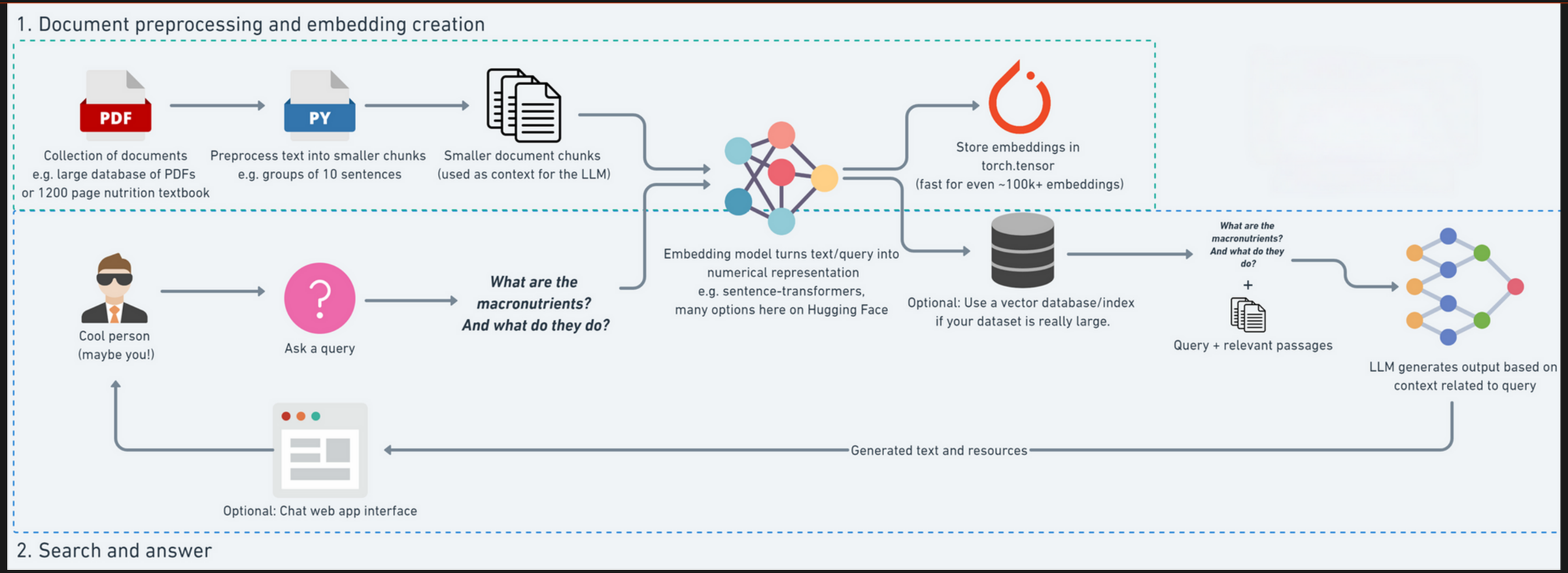
SIMILAR VECTOR SEARCH

Similar vector search algorithms efficiently find vectors with semantic similarity.

They measure similarity between vectors, often using methods like cosine similarity.

Similar vector search aids tasks such as information retrieval, particularly in large datasets.

Architecture



Leveraging Advanced Embedding Models and Efficient Similarity Search

Google Embeddings

Google embeddings are word embeddings or language models developed by Google.

Google embeddings are trained on large corpora of text data using deep learning techniques.

High dimensionality (768 dimensions) for capturing complex semantic relationships and strong performance metrics (MTEB score of 66.31)

Effectiveness in understanding textual entailment and semantic similarity.

Facebook AI Similarity Search

FAISS is specifically designed for handling high-dimensional data, such as the embeddings generated by Google embeddings

It employs advanced indexing techniques like hierarchical navigable small-world graphs.

Product quantization for fast and efficient organization and retrieval of vectors.

FAISS can handle large datasets with millions or billions of vectors.

Langchain Integration and Memory Management

Lang Chain

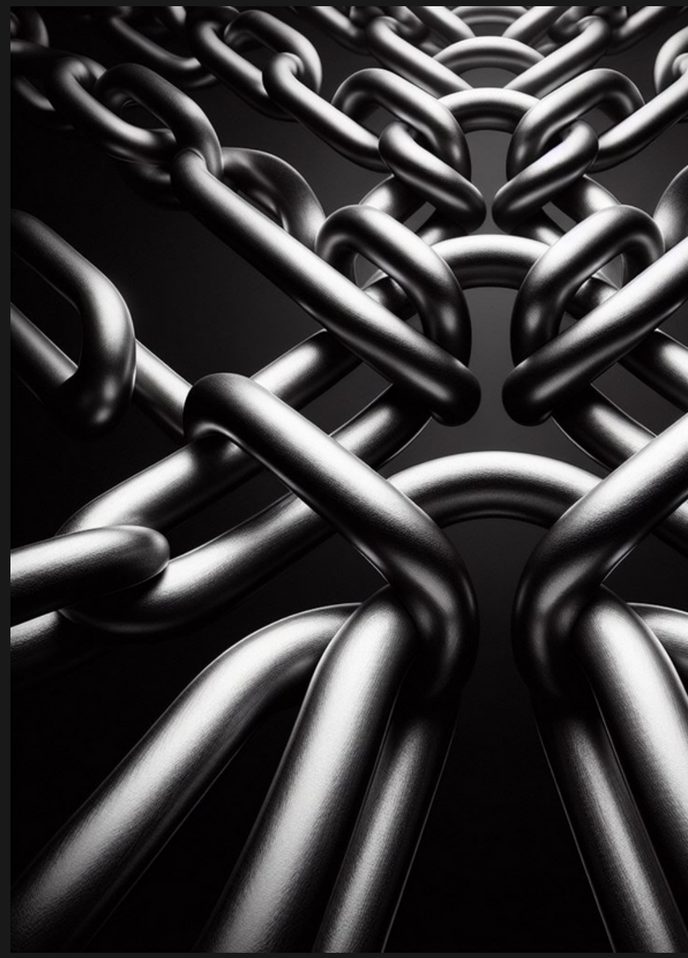
Langchain provides a modular framework that facilitates the integration of various text processing and analysis tasks seamlessly integration to multiple LLMs.

Level of customization enables the development of tailored solutions that address unique challenges and scenarios.

Memory Chain

Langchain includes a memory chain component that enables the chatbot system to store and retrieve conversational context.

This allows the chatbot to maintain continuity in conversations and provide more contextually relevant responses over time.



Quick Demo

Future Enhancement

01.

Implementing a variety of prompt styles enhances the efficiency and effectiveness of the chatbot's responses

02.

Developing a web extension enables seamless interaction with the chatbot directly to the browser.

Enhancing accessibility and user experience by facilitating quick and convenient access to document-related queries.

Thank You!

by Team Straw Hats