OmniSearch-Offline-Multimodal-RAG-Intelligence-Platform

Name : Arnab Kundu

REGN : 23BCE5013

Github Repository : https://github.com/arnab-kundu5013/OmniSearch-Offline-Multimodal-RAG-Intelligence-Platform

Problem Statement ID : 25231

Problem Statement Title

Design and build a multi-model Retrieval-Augmented Generation (RAG) system leveraging a Large Language Model (LLM) for OFFLINE mode that can ingest, index, and query diverse data formats such as DOC, PDF, Images and voice recordings within a unified semantic retrieval framework

Background

Centre routinely handles diverse data types: PDF, DOC, Images, screenshots, recorded calls, and free-form notes. Traditional search tools struggle with cross-format understanding, often isolating text, image, and audio searches. Retrieval-Augmented Generation (RAG) can address this by retrieving relevant data and grounding LLM outputs thus enhancing accuracy, reducing effort & time, and enabling content transparency. Multimodal RAG can extend this fusion to include images & audio, thus creating richer, context-aware intelligence.

Description

SIH Participants are to build a multimodal RAG-based system that:

Ingests multimodal inputs – Extract textual content from DOCX/PDF, generate embeddings for images, and perform speech-to-text conversion for audio.

Indexes all modalities in a shared vector space for seamless semantic retrieval.

Supports natural-language queries, retrieving relevant text snippets, images, and audio segments.

Generates grounded summaries or answers, integrating retrieved context via LLM.

Establishes cross-format links, e.g., connecting an audio transcript segment to a cited paragraph and screenshot.

Expected Solution

Unified Query Interface

● A simple chat or search box where users ‘type’ questions in plain language (e.g. “Show me the report that has a description about international development in 2024 OR show the report that references the screenshot taken at 14:32”).

● Optional - Support for different input modalities: other than text input - upload DOCX/PDF, drag-and-drop images, attach audio files, or speak their query.

Semantic & Cross-Modal Search

● Text-to-image search: Type a query like “email screenshot” to retrieve relevant images alongside matching text passages.

● Image-to-text search: Upload or select an image (e.g. screenshot) and surface relevant documents or audio transcript snippets that semantically match it.

Citation Transparency & Source Navigation

● Every answer includes numbered citations linking back to source files.

● Users can expand citations to open the original document, view full transcript segments, or inspect image metadata.

Short Description / Overview of Project :

In many organizations, important information is scattered across different formats PDF reports, Word files, screenshots, images, handwritten notes, and even recorded conversations. Searching through all of these manually is slow, frustrating, and often leads to missed details. Our project aims to solve this problem by creating an offline, multimodal Retrieval-Augmented Generation (RAG) system that can understand and connect all these formats in one place. The system extracts text from documents, recognizes content inside images, and converts audio into transcripts. All this information is then stored in a unified semantic index, allowing the user to simply ask a question in natural language just like chatting with an assistant. Whether the answer lies in a PDF paragraph, an email screenshot, or a specific moment in an audio recording, the system finds it, shows the relevant sources, and generates a clear, grounded response. By combining AI-powered retrieval with a local Large Language Model, this platform helps users save time, reduce manual searching, and build trust through transparent citations. Ultimately, it transforms scattered data into understandable, accessible knowledge instantly, and completely offline.

CATEGORY : SOFTWARE