DATA533: Project Description

Zaed Khan, Yihang Wang, Aaron Sukare

DataSage is a Python package designed to deliver AI-powered insights from datasets by synthesizing data files (csv, xlsx, txt, pdf) with metadata documentation. It streamlines complex RAG (Retrieval Augmented Generation) workflows into three core components: an **Ingestion Engine** to structure raw data into chunks, an **Indexing Engine** to embed content into a vector index for context-aware retrieval, and a **Response Engine** to generate grounded, natural-language answers.

We plan to include several core functions: **rag_engine()** executes the full pipeline from ingestion to vector storage; **query()** performs top-k similarity searches to answer specific questions; to accelerate analysis, **summary()** provides instant, high-level insights; and **save()** exports responses. This system provides a streamlined implementation of RAG, enabling local analysis of confidential documents and integrating metadata, so users can query information without having to sift through lengthy or verbose documentation.