

Title: Sanskrit Document Retrieval-Augmented Generation System

1. Introduction

This project implements a Retrieval-Augmented Generation (RAG) system for Sanskrit documents.

The system retrieves relevant document context and generates grounded answers using a CPU-based language model.

2. Objective

The objective is to build an end-to-end RAG system that works entirely on CPU and processes Sanskrit text.

3. System Architecture

The system consists of document loading, preprocessing, embedding generation, vector storage using FAISS, retrieval of relevant chunks, and answer generation using a language model.

4. Sanskrit Documents

The Sanskrit documents used in this project were provided as part of the assignment. They were processed in their original form without translation.

5. Preprocessing

Text preprocessing includes whitespace normalization and chunking with overlap to preserve semantic continuity.

6. Retrieval Mechanism

Multilingual embeddings are generated and stored in a FAISS vector index.

Top relevant chunks are retrieved based on semantic similarity.

7. Generation Mechanism

The retrieved context is passed to a CPU-based language model to generate grounded responses.

8. Performance

The system runs entirely on CPU with acceptable latency and low memory usage.

9. Limitations and Future Scope

Better Sanskrit-specific models and hybrid retrieval techniques can improve accuracy.

10. Conclusion

The project demonstrates a complete CPU-based RAG pipeline for Sanskrit documents with clear modular design.