Final Project Report

Sanskrit Document Retrieval System (RAG)

Intern Name: Harshal Borkar

Project Title: RAG-based Sanskrit Document Retrieval

1. Introduction

This project implements a Retrieval-Augmented Generation (RAG) based Sanskrit document retrieval system. The system allows users to query Sanskrit documents written in Devanagari script and retrieve the most relevant textual passages using semantic similarity search.

2. Objective

The objective of this project is to build an efficient, CPU-based, extractive question answering system for Sanskrit documents without hallucination, ensuring accurate and context-based responses.

3. System Architecture

- PDF Document Loader using UnstructuredPDFLoader
- Text cleaning and Devanagari validation
- Recursive text chunking
- Multilingual sentence embeddings (HuggingFace)
- FAISS vector database for similarity search
- Streamlit-based user interface

4. Technology Stack

- Python
- Streamlit
- LangChain
- FAISS
- HuggingFace Sentence Transformers

5. Features

- Sanskrit-only query validation
- Button-based document retrieval
- Query memory for previous searches
- Extractive answers from source documents

6. Conclusion

The Sanskrit Document Retrieval system successfully demonstrates the application of modern NLP and vector search techniques to classical language processing. The project is scalable, efficient, and suitable for educational and research use cases.