PROJECT DOCUMENTATION

Customer Support Knowledge Base – FAST RAG App

Author: Srinivas
B.Tech CSE (AI & ML)

1. Introduction

The Customer Support Knowledge Base – FAST RAG App is an offline Retrieval-Augmented Generation (RAG) system designed to help users search and retrieve information from documents such as FAQs, Troubleshooting Guides, and Support Tickets. The system uses local embeddings and FAISS vector search to deliver fast and accurate answers without requiring internet access or API keys.

2. Problem Statement

Companies store large amounts of support information across multiple PDF documents. Searching manually takes time. This project solves that problem by converting documents into searchable vectors using modern NLP techniques.

3. Objectives

• Build an offline AI-based document search system.
• Perform end-to-end RAG using local models.
• Create a clean and simple Streamlit interface.
• Demonstrate embeddings, vector search, and PDF automation.

4. System Architecture

The architecture includes:

PDF Loader: Extracts text from documents.
Text Splitter: Breaks text into manageable chunks.
Embedding Model: Converts chunks into numerical vectors.
FAISS Index: Stores and retrieves similar vectors.
UI Layer: Streamlit application for user interaction.

5. Workflow

1. Load PDFs from the data folder.
2. Extract and chunk text.
3. Generate embeddings using MiniLM model.
4. Save vectors in FAISS database.
5. User enters a query.
6. Query is converted to vector format.
7. System retrieves closest matching answers.
8. UI displays results instantly.

6. Technologies Used

• Python
• Streamlit
• Sentence Transformers
• FAISS
• LangChain PDF Processing

7. Key Features

• Fully offline – no API keys required
• Fast retrieval with FAISS
• Lightweight and optimized for laptops
• Simple UI for easy usage

8. Applications

Practical use-cases include:
• Customer support automation
• Internal company search engines
• Document management systems
• Technical troubleshooting assistants

9. Conclusion

This project demonstrates a complete offline RAG system using open-source tools. It showcases key AI engineering concepts such as embeddings, vector indexing, NLP automation, and UI design. The system can be expanded into advanced intelligent search platforms for real-world use.

10. Developer

Srinivas
B.Tech CSE (AI & ML)